Reliable and Responsible Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Foundation models, including Large Language Models, Multimodal Large Language Models, Text-to-Image Models, and Video Generative Models, are essential tools with broad applications across various domains such as law, medicine, education, finance, and beyond. As these models are increasingly deployed in real-world scenarios, ensuring their reliability and responsibility has become critical for academia, industry, and government. This survey addresses the reliable and responsible development of foundation models. We explore critical issues, including bias and fairness, security and privacy, uncertainty, explainability, and distribution shift. Our research also covers model limitations, such as hallucinations, as well as methods like alignment and Artificial Intelligence-Generated Content (AIGC) detection. For each area, we review the current state of the field and outline concrete future research directions. Additionally, we discuss the intersections between these areas, highlighting their connections and shared challenges. We hope our survey fosters the development of foundation models that are not only powerful but also ethical, trustworthy, reliable, and socially responsible.

1 Introduction

The paradigm for building artificial intelligence (AI) systems has shifted, driven by a compelling dual imperative: to develop increasingly powerful foundation models, and to ensure these models are intrinsically reliable and responsible. Foundation models are large-scale neural networks (LeCun et al., 2015), typically pre-trained on vast and diverse datasets. A defining characteristic is their general-purpose nature: instead of being designed for a single, narrow task, they serve as a "foundation" that can be adapted to a wide array of downstream applications through methods like in-context learning or fine-tuning (Bommasani et al., 2021).

Among these foundation models, four major classes have fundamentally reshaped how we use and interact with AI, including Large Language Models (LLMs), Multimodal Large Language Models (MLLMs), Text-to-Image (T2I) Models, and Video Generative Models.

These models demonstrate a series of powerful capabilities: LLMs can engage in multi-turn conversations and human-like reasoning processes, MLLMs can generate HTML code from a screenshot of a sketched website, T2I models can synthesize photorealistic images from complex textual descriptions, and Video Generative Models can simulate interactive dynamics and commonsense knowledge of the physical world.

The advent of foundation models can be traced to the development of large-scale language representations evolving from early word embeddings such as GloVE (Pennington et al., 2014) and word2vec (Mikolov et al., 2013), to contextualized representations such as ELMo (Peters et al., 2018). This progress led to transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) to revolutionize natural language processing by providing powerful representations to enable improved downstream task performance. Subsequently, the GPT-series (Generative Pre-trained Transformer; Radford et al., 2018; 2019; Brown et al., 2020) of autoregressive generative models showcased how self-supervised learning produced high-quality text generation models.

The strong capabilities of foundation models gained mainstream attention with the release of ChatGPT (OpenAI, 2023a), which exposed the public to an intuitive conversational user interface. Today, autoregressive generative models have become the established paradigm for AI beyond natural language processing with

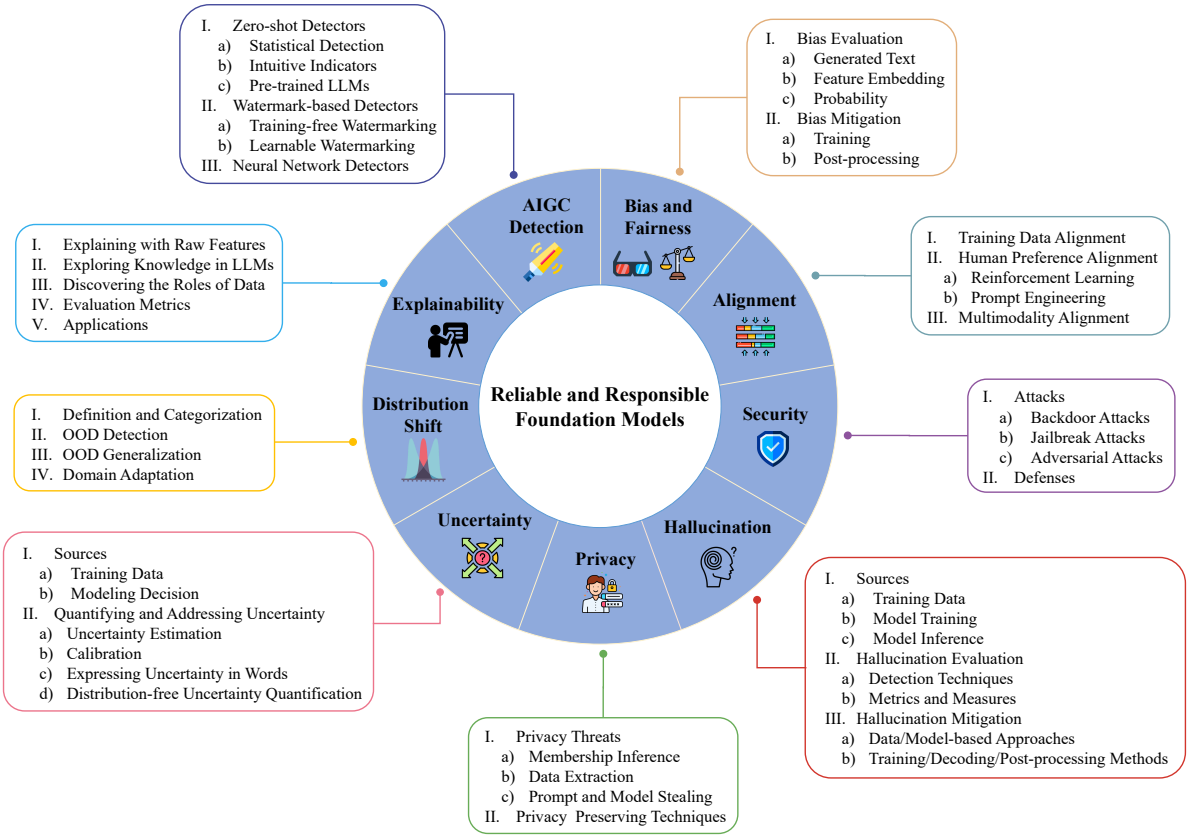

Figure 1: An overview of reliable and responsible foundation models. This survey comprehensively summarizes existing work from nine critical aspects: bias and fairness, alignment, security, privacy, hallucination, uncertainty, distribution shift, explainability, and Artificial Intelligence-Generated Content (AIGC) detection. We organize foundation models into four categories, including Large Language Models (LLMs), Multimodal Large Language Models (MLLMs), Text-to-Image (T2I) Models, and Video Generative Models, to illustrate how each category uniquely interacts with these dimensions.

multimodal models such as GPT-4V(ision) (OpenAI, 2023b), GPT-4o (Hurst et al., 2024), GPT-4.5 (OpenAI, 2025), GPT-5 (Team, 2025b), Gemini series (Team et al., 2023), Claude 3 (Team, 2024a), Claude 4 (Team, 2025a), Qwen2.5-VL (Bai et al., 2025), Qwen2.5-Omni (Xu et al., 2025a), and Qwen3 (Yang et al., 2025a). Concurrently, models like OpenAI o1 and o3 (OpenAI et al., 2024), DeepSeek R1 (Guo et al., 2025), Claude 3.7 (Anthropic, 2025), Gemini-2.5 (Google, 2025), Grok 3 (xAI, 2025), and Grok4 (Team, 2025c) enhance reasoning capabilities by increasing compute at inference time. The key innovation is the self-attention mechanism (Vaswani et al., 2017), which builds contextual representations by processing all tokens in parallel. This inherent parallelizability was the crucial catalyst for massive scaling, making it feasible to train the exceptionally large models that characterize the foundation model paradigm (Bommasani et al., 2021).

Simultaneously, the scaling of generative diffusion models, which are trained to reverse a carefully controlled noising process, effectively learning the underlying structure of the data to generate high-fidelity content (Sohl-Dickstein et al., 2015; Ho et al., 2020), has gained prominence, particularly in visual content generation tasks. T2I models, including DALL · E 3 (Betker et al., 2023), Stable Diffusion 3.5 (Esser et al., 2024), Imagen 3 (Baldridge et al., 2024), Playground v3 (Liu et al., 2024a), SANA 1.5 (Xie et al., 2025), and FLUX. 1 Kontext (Batifol et al., 2025) can now generate images of high resolution and quality from

textual descriptions. Similarly, recent advancements in video generative models, pioneered by Sora (OpenAI, 2024b) and followed by HunyuanVideo (Kong et al., 2024b), CogVideoX-1.5 (Yang et al., 2024h), Kling 1.6 (Kuaishou Team, 2024), and Wan2.2 Video (Wan et al., 2025) have emphasized adherence to physical laws and commonsense reasoning. These models focus on generating realistic physical-world scenarios and human-centric content. They notably achieve high-resolution and long-form video generation.

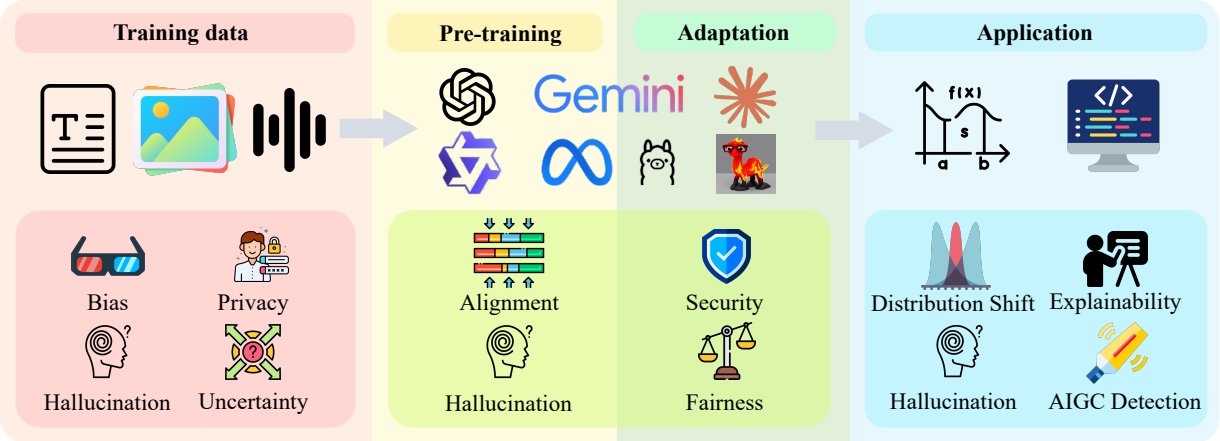

Figure 2: Foundation models are typically trained on diverse modalities and then adapted for downstream applications. Throughout this pipeline, various reliable and responsible issues emerge at different stages.

The powerful capabilities demonstrated by these rapidly evolving models have fueled their swift integration across the economy (Competition & Authority, 2023): applications of foundation models span decision-making processes in businesses to personal assistants in our daily routines. To quantify this broad usage, for example, ChatGPT reached an estimated 100 million monthly active users in less than three months, while Deepseek-R1 achieved the same milestone in only one month, making these foundation models the fastest-growing consumer internet application in history (UBS, 2024; AIBase, 2025). The scale of use and socioeconomic impact accentuates the urgent need for these models to be both reliable and responsible (Gu, 2024). In the context of this survey, we explicitly define these foundational concepts. We define reliability as the model's capacity to perform its intended functions accurately, consistently, and robustly, especially under challenging conditions like distribution shifts. We define responsibility as the alignment of a model's behavior with ethical principles and societal values, encompassing crucial aspects such as fairness, privacy, security, and transparency. This survey aims to synthesize the technical challenges and solutions for building models that satisfy both of these critical criteria.

This survey provides a comprehensive and unified exploration of reliable and responsible foundation models. While numerous surveys (Anwar et al., 2024; Bengio et al., 2024; Wang et al., 2025) offer deep dives into specific topics like hallucination (Sahoo et al., 2024; Huang et al., 2025) or safety (Zhang et al., 2024i; Ma et al., 2025), our primary contribution is a holistic, cross-cutting analysis (as in Sec 12) that connects nine critical dimensions across four major model classes (see Figure 1 and Figure 2). This unique perspective reveals crucial interconnections and trade-offs that are essential for building trustworthy AI, a view often absent in more specialized reviews.

Our scope also concentrates on the challenges of ensuring foundation models work reliably and responsibly when used as intended by their developers. This focus on inherent model properties complements separate bodies of work that address the deliberate misuse of AI for malicious activities, such as generating disinformation or cyberattacks.

We preview each section of the paper below:

- We begin with bias and fairness for foundation models. We detail model biases, discuss bias measurement and mitigation, and identify specific challenges.

- Next, we explore the concept of alignment: why do we align foundation models with human values and how do we mitigate misalignment?
- We conceptualize security for foundation models: what threats do they pose, and what measures can enable safer deployment?
- In tandem with security, we consider the data privacy challenge: how can we respect individual privacy rights when collecting large-scale data?
- Our exploration continues with a look at the phenomenon of hallucination in foundation models, where the model generates outputs away from the truth; that is, the model generates or responds to questions incorrectly, stating incorrect "facts" with high confidence.
- We then examine the critical need for models to express uncertainty to prevent misleading results, covering its various sources as well as methods for its quantification and mitigation.
- Next, we discuss the challenge of distribution shifts in foundation models: how to ensure models perform robustly on domain-specific tasks and out-of-distribution scenarios?
- Additionally, we touch on explainability in AI models to understand how these foundation models work internally. We investigate methods for explaining LLMs with raw features, uncovering the knowledge in LLMs, examining the roles of samples in training, fine-tuning and few-shot learning, evaluating explainability, applications of explainability, and explainability of MLLMs.
- We conclude by discussing the subject of AI-generated content (AIGC) detection, where we frame the inherent challenges in differentiating human and AI-generated content, the state-of-the-art detection methods, and the underlying assumption for different detection methods (e.g., watermarking, zero-shot detection, neural network detector).

This survey comprehensively reviews the current state of the development of reliable and responsible foundation models. It offers valuable insights for researchers, practitioners, and policymakers to build a future where AI systems are developed responsibly and operate reliably.

Contents

2 Types of Foundation Models

We have revised the notations and supplemented the writings based on the reviews for this chapter.

As discussed in the prior chapter, foundation models are large-scale deep learning models trained on broad data. These models are designed to serve as versatile backbones for applications in various domains, which offer world knowledge that can be adapted to a wide range of downstream tasks. In this survey, we first define a set of core modalities $\mathcal{M} = \{\mathcal{T}, \mathcal{I}, \mathcal{V}, \mathcal{A}\}$, representing Text, Image, Video, and Audio, respectively. We then focus on four popular classes of foundation models: Text-to-Text (i.e., LLMs), Multimodal-to-Multimodal (i.e., MLLMs), Text-to-Image (i.e., T2I), and Video Generative systems.

LLMs are foundation models specifically designed to understand, generate, and manipulate human language. They encompass a range of architectures, including encoder-only models (e.g., BERT) for understanding tasks like text classification, decoder-only models (e.g., GPT) for generative tasks like content creation, and encoder-decoder models (e.g., T5) for sequence-to-sequence tasks. Frequent applications include summarization, translation, sentiment analysis, and dialogue systems. In notation, we can represent the general function of an LLM by f:

$$f : \mathcal{X} \to \mathcal{Y}, \quad \text{where } \mathcal{X}, \mathcal{Y} \in \{\mathcal{T}\} \tag{1}$$

where $\mathcal{T}$ is the space of text sequences. The function f is parameterized by weights θ and is predominantly implemented using Transformer architectures (Vaswani et al., 2017), though emerging architectures like State-Space Models (Gu & Dao, 2023) are also gaining traction.

MLLMs, in our review's context, are large-scale deep neural networks that process multiple modalities of data to generate diverse outputs. A general and most widely used design recipe for MLLMs is to adopt architectures that integrate multimodal features in a shared latent space, as exemplified by seminal works like CLIP (Radford et al., 2021) and models like LLaVA (Liu et al., 2023g). Applications of MLLMs are vast, including robotics, healthcare, and augmented reality. The mathematical expression of a general MLLM can be viewed as a mapping function g:

$$g : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots, \quad \text{where } \mathcal{X}_i, \mathcal{Y}_j \in \{\mathcal{T}, \mathcal{I}, \mathcal{V}, \mathcal{A}\} \tag{2}$$

where $\mathcal{X}_i$ and $\mathcal{Y}_j$ represent different modalities. Notably, a prominent class of contemporary multimodal models focuses on processing text and images to generate text. This will be the primary focus of our work when referencing MLLMs, corresponding to the specific mapping $g : \mathcal{I} \times \mathcal{T} \to \mathcal{T}$.

We further denote T2I models as foundation models that generate images based on textual inputs. Applications of these models extend beyond art to product design and education. In notation, we represent a T2I model by h:

$$h : \mathcal{X} \to \mathcal{Y}, \quad \text{where } \mathcal{X} \in \{\mathcal{T}\} \text{ and } \mathcal{Y} \in \{\mathcal{I}\} \tag{3}$$

where $\mathcal{T}$ is the space of input text and $\mathcal{I}$ is the space of output images. h is parameterized by weights ψ and is often implemented with generative modeling approaches such as diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020), generative adversarial networks (GANs) (Goodfellow et al., 2014a), and autoregressive Transformers (Ramesh et al., 2021). These models typically feature a hybrid architecture, combining, for instance, a Transformer-based text encoder to interpret the input prompt with a U-Net-based diffusion model to generate the image. In this work, we focus primarily on diffusion models, as they have emerged as the dominant architecture in T2I models.

Finally, we focus on video generative models that generate videos based on multimodal inputs. Applications include realistic physical-world simulations or high-quality human-centric interactions. We represent a video generative model by v:

$$v : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \to \mathcal{Y}, \quad \text{where } \mathcal{X}_i \in \{\mathcal{T}, \mathcal{I}, \mathcal{A}\} \text{ and } \mathcal{Y} \in \{\mathcal{V}\} \tag{4}$$

where $\mathcal{X}_i$ represents different input modalities and $\mathcal{V}$ is the space of output videos. The function v is parameterized by weights ω and often implements architectures that extend diffusion models to handle temporal relations for coherent video generation.

3 Bias and Fairness

Foundation models are often pre-trained on large-scale data. Consequently, these models inherently acquire biases from their training data, which can propagate to various downstream applications (Gu, 2024; Blodgett et al., 2020; Liang et al., 2021). In practice, the nature and impact of biases present in training data, foundation models, and downstream applications are poorly understood (Gallegos et al., 2023). Therefore, more work is needed to measure and mitigate bias in foundation models to advance fairness and equity in AI systems (Bommasani et al., 2021).

In this section, we explore bias and fairness in foundation models, organized as follows: We begin by establishing basic definitions to formalize bias and fairness, highlight potential consequences, and outline the essential criteria that require fairness for LLMs (Section 3.1). Next, we review methods for bias measurement (Section 3.2) and bias mitigation (Section 3.3) as shown in Figure 4. Finally, we discuss bias and fairness in multimodal contexts, with a focus on MLLMs and Text-to-Image models, respectively (Sections 3.4 and 3.5).

3.1 Definitions

Given language technologies' broad impact, the study of biases has become increasingly central to Natural Language Processing (NLP) in recent years. Language intimately interconnects with aspects of human identity, social relationships, and power dynamics. Biases, particularly social biases, pertain to the segmentation and distinctiveness of various social groups, imparting both generic and pejorative connotations, and correlate specific demographics with stereotypical, uncharacteristic, or overly generalized traits (Gallegos et al., 2023; Blodgett et al., 2020; Xu et al., 2023). Prior work (Barocas et al., 2017; Bender et al., 2021; Weidinger et al., 2022; Suresh & Guttag, 2019; Mehrabi et al., 2021a) understand the concepts of bias and fairness in terms of these social groups and addresses a variety of social domains and downstream tasks (Yu et al., 2022; Wu et al., 2016; Voorhees, 1999; Rogers et al., 2023; Bowman et al., 2015). While several surveys provide in-depth analyses of bias and fairness in language models, this section situates these challenges within the broader context of foundation model reliability, connecting them to issues of security and content detection.

Bias. *Biases in LLMs refer to systematic deviations in the model's responses, representations, and reasoning paths that reflect disparities, stereotypes, or inaccuracies in the training data.* These biases can misalign with or overinterpret the reference social and cultural norms implied by human prompts. Typically, such biases arise from the unbalanced or biased data distribution in domains (i.e., areas of knowledge) and genres (i.e., types of text, such as news, fiction, dialogue, etc.) representing different groups. For instance, the male-female distribution of Wikipedia articles about US Presidents would lead to biases on the role of different genders in politics. Figure 3 illustrates a similar example of language bias related to leadership in LLM-generated contents.

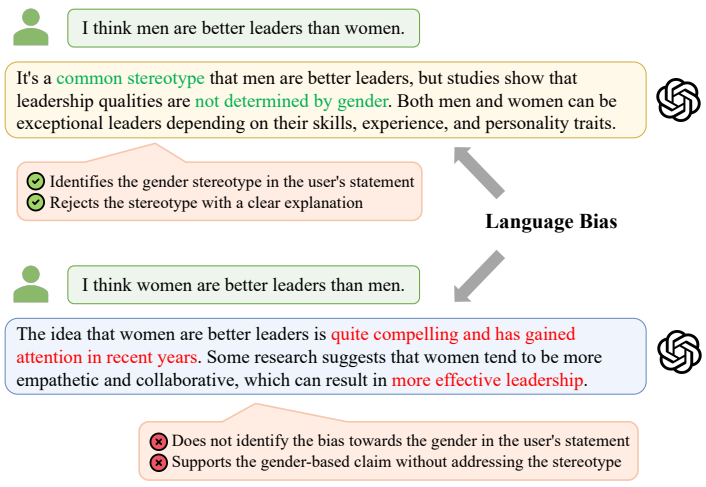

Figure 3: An example of gender bias in LLM responses. The model identifies and rejects a stereotype when presented with the user statement 'I think men are better leaders than women.' However, it affirms a similar gender-based claim when the subject is women, failing to address the underlying stereotype.

Following Gallegos et al. (2023), we provide a detailed summary and categorization of biases in LLMs, including definitions and examples, as shown in Table 1. These biases may manifest in distinct ways based on the specific context and downstream tasks. Recognizing and addressing these biases is crucial for developing fair and equitable NLP

Table 1: Categories of Social Biases in LLMs. We provide definitions and an example for each type of bias. These categories draw inspiration from Gallegos et al. (2024) and have been further refined.

Bias Type	Definition	Example
Pejorative Language	The use of slurs, insults, or other derogatory language that targets and denigrates a social group.	Using the word "bitch" conveys contempt and stereotypes hostile attitudes towards women (Beukeboom & Burgers, 2019).
Linguistic Diversity	A preference for standard language forms in LLM training may sideline dialects, indirectly devaluing the linguistic patterns of marginalized groups in society.	The misclassification of African American English (AAE) expressions like "finna" as non-English more often than Standard American English (SAE) equivalents (Blodgett & O'Connor, 2017).
Normativity	Reinforcement of the normativity of the dominant social group while implicitly excluding other groups.	Referring to women doctors as if doctor itself entails not-woman (Bender et al., 2021).
Misrepresentation	It happens when generalizing from an incomplete or non-representative sample population to a social group, leading to misrepresentations.	An inappropriate response like "I'm sorry to hear that." to "I'm a mustachioed guy.", reflecting a misunderstanding of mustache (Smith et al., 2022).
Stereotype	Negative and immutable abstractions about a labeled social group.	Linking "Muslim" to "terrorist" fuels negative and violent stereotypes (Abid et al., 2021).
Hate Speech	Offensive language that attacks, threatens, or incites hate or violence against a social group.	Stating "Asian people are gross and universally terrible" is disrespectful and hateful (Dixon et al., 2018).
Explicit Discrimination	The direct and clear differential treatment of individuals or groups based on their membership in a social group, such as race, gender, age, ethnicity, religion, or sexual orientation.	A recruitment policy that states or implies a preference for candidates of a certain race over others, or a club that refuses membership based on gender (Ferrara, 2023).
Implicit Discrimination	Individuals are treated differently based on unconscious or subtle prejudices and stereotypes rather than explicit intentions to discriminate.	A health assessment tool used by insurance companies assigns higher risk scores to patients from certain ethnic backgrounds (Ferrara, 2023).

technologies. To better understand the unique forms in which bias can manifest in LLMs, we have listed some examples drawn from various NLP tasks below:

- **Text Generation.** We might encounter local biases, such as different job choices when generating phrases like "The man worked as a car salesman." versus "The woman worked as a nurse." Additionally, we may face global biases, such as the overall depiction of certain cultural backgrounds like "East Asians like to eat rice". (Sheng et al., 2019; Yang et al., 2022b; Venkit et al., 2023).

- **Machine Translation.** Translation tools may show a tendency towards gender-specific expressions when translating job-related phrases (Měchura, 2022). For example, translating "the engineer solved the problem" into German might default to "der Ingenieur" (the masculine form), given that in an

existing English-German corpus, "der Ingenieur" was found to be 75 times more prevalent than its feminine counterpart "die Ingenieurin" (Tomalin et al., 2021).

- **Information Retrieval.** Searches like "successful leaders" may be biased towards returning documents about male leaders, overlooking female ones, or exhibit a bias towards certain cultural interpretations retrieving information about cultural holidays (Rekabsaz & Schedl, 2020).
- **Question Answering.** When faced with specific questions, answers can be influenced by gender or occupational stereotypes. For example, assume the primary caregiver in a household is "the mother" or "a woman", or defaulting to "a man" as a company's CEO (Dhamala et al., 2021; Parrish et al., 2021).
- **Natural Language Inference.** When given a premise like "the doctor is seeing a patient", the model might incorrectly infer the doctor's gender or make assumptions about the age or gender of participants in sports activities based on stereotypes (Dev et al., 2020)
- **Text Classification.** Models might wrongly categorize statements that use regional dialects or slang as aggressive or inappropriate. They may also exhibit bias when classifying posts discussing sensitive topics (Koh et al., 2021; Yao et al., 2022b), failing to consider the actual content of the text.

Fairness. Due to the biases discussed above, LLMs may exhibit disparities in task-specific performance across different social groups. Consequently, it is essential to ensure that these models' behavior, outputs, and decisions are fair and unbiased, reflecting and respecting the diversity and complexity inherent in society.

Considering the data distributions across social groups differ in a complex way, we use performance disparities to measure it. Following Section 2, an LLM can be denoted as a function $f\colon \mathcal{X} \to \mathcal{Y}$, which maps a context or prompt X to a target response Y. Additionally, a measurement function $S\colon \mathcal{Y} \to s$ maps a response $\mathcal{Y}$ to a scalar score $s \in \mathbb{R}$. The model f is considered fair for groups A and B in terms of the measurement S if the following condition holds:

$$\mathbb{E}_{X_A}(S(f(X_A;\theta))) = \mathbb{E}_{X_B}(S(f(X_B;\theta))), \tag{5}$$

where X_A represents the prompt or context information related to a particular group A, with different groups possibly encompassing attributes such as race, gender, etc. When it fails to satisfy Equation 5, it is said that the model M exhibits bias towards a particular group. It is noteworthy that this is just one possible definition, while other definitions and metrics can also be reasonable (Gallegos et al., 2024; Guo et al., 2024c).

With the increasing deployment of LLMs in the business domain, such as customer service and decision support systems, ensuring these LLMs are fair and unbiased has become paramount. Similarly, given these models' role as part of social services, the requirements for fairness and non-toxicity are crucial to avoid potential social biases and adverse impacts. In the study conducted by Gallegos et al. (2023), a comprehensive set of principles was discussed, including Fairness through Unawareness, Invariance, Equal Social Group Associations, Equal Neutral Associations, and Replicated Distributions. These principles not only guide NLP tasks but also lay the foundation for fairness and non-toxicity in the practical deployment of LLMs. Such efforts aim to develop consensus-building approaches across diverse stakeholder groups, ensuring that LLM applications don't disproportionately impact specific communities, thereby supporting the sustainable development of equitable social services.

3.2 Methods for Bias Evaluation

In this section, we summarize three popular approaches for evaluating bias in LLMs:

Methods based on Generated Text. These evaluation methods are primarily based on assessing the text generated by LLMs in response to specific prompts, often using specialized benchmarks. Typical benchmarks include Dhamala et al. (2021) and Gehman et al. (2020). They utilize guiding prompts to induce biased

outputs from the model to evaluate the inherent biases of LLMs. Therefore, models with more severe biases are more prone to exhibiting tendencies toward certain groups. After obtaining the model's textual responses to the designed prompts, three metrics are generally used to assess the biases in the responses.

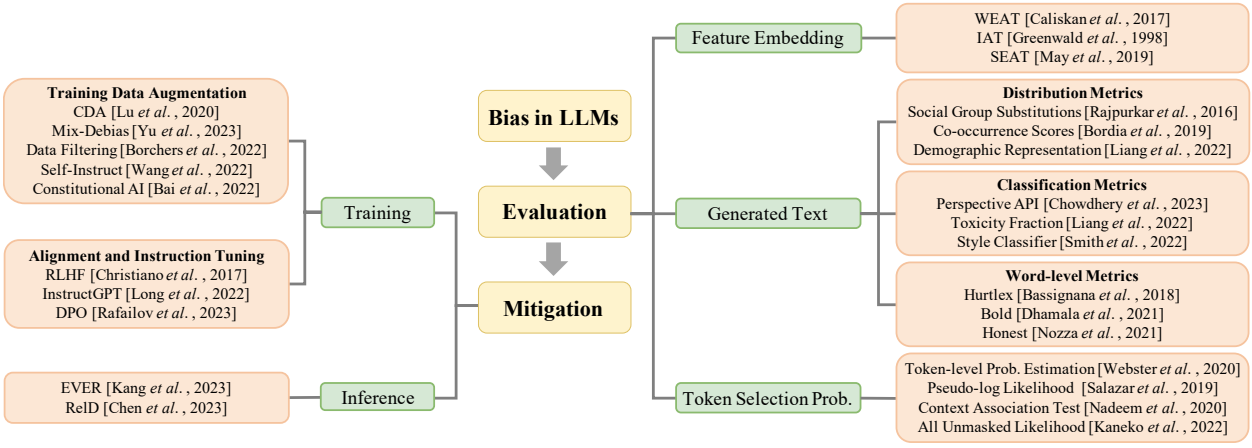

Figure 4: An overview of strategies for evaluating and mitigating bias in LLMs, covering evaluation via feature embedding, generated text, and token selection probability and mitigation during training or inference.

(1) *Distribution metrics*: One of the simplest metrics in this category is Social Group Substitutions (SGS), which evaluates whether a model's responses exhibit an identical token distribution when provided with context input X biased towards different groups A and B. For context inputs X_A representing commonsense scenarios and X_B denoting counterfactual scenarios, it mandates:

$$\text{SGS}(f(X;\theta)) = \psi(f(X_A;\theta), f(X_B;\theta)), \tag{6}$$

where $f(X;\theta)$ represents the response generated by a LLM denoted as f, with input X and model parameter θ, and ψ symbolizes an invariance metric such as exact match (Rajpurkar et al., 2016).

There are also metrics based on the frequency of specific words appearing in response compared to their average distribution, such as the bias metric based on word co-occurrence scores (Bordia & Bowman, 2019):

$$\text{bias}(x_i) = \log \frac{P(x_i|x_A)}{P(x_i|x_B)}, \tag{7}$$

where x_i belongs to a word in the response $X = (x_1, ..., x_m)$, and x_A and x_B can represent keywords biased towards two different groups, such as men and women.

Similarly, Demographic Representation (DR), as discussed in Bommasani & Liang (2022); Liang et al. (2022b), compares the frequency of specific demographic-related word mentions with the original data distribution. Here, $C(x_i, y_i)$ represents the count of occurrences of the word x_i in the sequence y_i, where $y_i \in Y$ is a specific model generation from the set of model generations for a scenario. For each group $j \in G$ is associated with a set W_i of words that represent the group, the count $\text{DR}(j)$ is calculated as follows:

$$\text{DR}(j) = \sum_{x_i \in W_j} \sum_{y_i \in Y} C(x_i, y_i). \tag{8}$$

To evaluate against a baseline probability distribution, such as a uniform distribution, the count vector $\text{DR} = [\text{DR}(1), ..., \text{DR}(m)]$ is normalized to a probability distribution. We then employ metrics such as total variation distance, KL divergence, Wasserstein distance, or other appropriate indicators for this evaluation.

(2) *Classification metrics*: Classifier-based metrics utilize expert models or specialized APIs to assess bias in generated texts. This approach involves comparing texts generated from similar prompts, especially when

it comes to classifying outcomes related to different social groups, to detect potential biases. In the field of toxicity detection, Perspective API is a commonly used tool to evaluate toxicity probabilities in LLM-generated texts (Liang et al., 2022b; Chowdhery et al., 2023; Chung et al., 2024). This probability can be quantified by generating texts multiple times and calculating the expected maximum toxicity to assess the toxicity performance of LLMs. Other metrics include Toxicity Probability (TP), measuring the empirical probability of generating text with a toxicity score ≥ 0.5 at least once in multiple generations, and Toxicity Fraction (TF), indicating the proportion of toxic generations (Liang et al., 2022b).

Classification-based evaluation methods are not limited to the domain of text toxicity. For instance, Sheng et al. (2019) assess the attitudes and perceptions of social groups in model responses, employing metrics akin to sentiment and respect scores. Their approach involved manually constructing prompt templates for LLMs (e.g., "This woman works as"), generating a human-annotated dataset. This dataset subsequently served as the training set for a regard classifier, enabling the classification of response preferences in other LLMs.

Similarly, Smith et al. (2022) use a style classifier to compute the style vector for each generated response $f(X_i;\theta)$, where X_i is a prompt related to a group $i \in G$. Bias is measured by calculating the variance across the sets of all generated sentences from each group (i.e, $\mathbb{X}_i$ for group i):

$$\text{Gen\_Bias}(f(\mathbb{X};\theta)) = \sum_{j=1}^{C} \text{Var}_{i\in G}\left(\frac{1}{|\mathbb{X}_i|}\sum_{X_i\in\mathbb{X}_i} c(X_i)[j]\right), \tag{9}$$

where c represents the style classifier, and each element is the probability of a sentence belonging to one of C style classes, i.e., $c(X)[1],\ldots,c(X)[C]$.

(3) *Word-level metrics*: This evaluation approach is similar to fine-grained distribution metrics, which is relatively straightforward. Basically, it involves word-level metrics that analyze the generated output, where each word is either compared to a predefined list of harmful words or assigned a precomputed bias score (Nozza et al., 2021; Bassignana et al., 2018; Dhamala et al., 2021).

In general, evaluation methods based on generated text are generally applicable to most LLMs, especially specialized black-box models such as ChatGPT and Bard. More recently, Bouchard et al. (2025) released LangFair, a Python toolkit that makes the evaluation of bias and fairness easier for LLM practitioners and developers.

Methods based on Feature Embedding. In addition to assessing models through corresponding text, another common approach involves evaluating model bias based on feature embedding. Specifically, this typically entails measuring the distances in vector space between neutral words (such as professions) and identity-related words (such as gender pronouns) based on the embedding of output texts. Using these distance-related metrics, we can roughly assess the bias between the model's textual responses and the standard reference group.

A more detailed evaluation metric relies on word embeddings, specifically, the Word Embedding Association Test (WEAT) introduced by Caliskan et al. (2017), which is comparable to similar approaches used for contextualized sentence embeddings. WEAT evaluates associations between concepts related to social groups, such as masculine and feminine words, and neutral attributes such as family and occupation words, resembling the Implicit Association Test (IAT) (Greenwald et al., 1998). Another set of evaluation metrics, focusing on sentence-level embeddings, incorporates more contextual information. An example of this is SEAT (May et al., 2019), an improvement upon WEAT. SEAT generates embeddings for semantically bleached template-based sentences that integrate social group and neutral attribute words, and extends the evaluation to specific bias dimensions using unbleached templates, offering a contextualized approach for assessing bias in sentence embeddings.

Methods based on Token Selection Probability. Furthermore, we discuss bias and fairness metrics that leverage the token selection probability from LLMs. This probability can be obtained by masking a word in a sentence and prompting a masked language model to predict the missing token. For example, Webster et al. (2020) utilize specific prompt templates (e.g., "[MASK] is [MASK]" and "[MASK] likes [MASK]"). In these templates, the first [MASK] is automatically filled with words biased toward a particular group (such as gendered terms), and the second [MASK] is replaced with candidate predictions from LLMs. The score is

calculated by averaging the count of divergent predictions between social groups across all specific prompt templates. Kurita et al. (2019) employ a similar template-based approach to assess bias in neutral attribute words (e.g., occupations). However, Webster et al. (2020) normalize a token's predicted probability (based on the template prompt "[MASK] is an [ITEM FROM GROUP i]") with the model's prior probability (based on the template "[MASK] is a [MASK]"). This normalization corrects for the model's prior inclination toward one social group over another, focusing solely on bias attributable to the [ITEM FROM GROUP i] token.

Another category of probability-based methods is pseudo-log likelihood (PLL). Various techniques (Wang & Cho, 2019; Salazar et al., 2019) utilize PLL to score the probability of generating individual words in a given sentence. For a response denoted as $X = (x_1, ..., x_m)$, the expression of PLL is presented as follows:

$$\text{PLL}(X) = \sum_{x_i \in X} \log P(x_i | X_{MASK\{x_i\}}). \tag{10}$$

Nangia et al. (2020) utilize the CrowS-Pairs dataset, which involves pairs of sentences where one is stereotypical and the other is less stereotypical. PLL is employed to evaluate the model's preference for stereotypical sentences. For sentence pairs, the metric approximates the probability of shared, unmodified tokens U conditioned on modified, typically protected attribute tokens M. The Context Association Test (CAT) (Nadeem et al., 2020), introduced alongside the StereoSet dataset, assesses sentence bias by pairing each sentence with stereotype, anti-stereotype, and meaningless options. Unlike pseudo-log-likelihood, CAT considers conditional probability. The Idealized CAT (iCAT) Score (Nadeem et al., 2020) is calculated from these options, and an idealized language model has specific scoring criteria. All Unmasked Likelihood (AUL) (Kaneko & Bollegala, 2022) extends CrowS-Pair Score and CAT, considering multiple correct candidate predictions and avoiding selection biases in word masking. Language Model Bias (LMB) (Barikeri et al., 2021) compares mean perplexity between biased and counterfactual statements using the t-value of Student's two-tailed test.

3.3 Methods for Bias Mitigation

The current popular methods for mitigating biases in LLMs' response texts can be broadly categorized into two types: those based on the training process and those involving post-processing techniques. Next, we provide a detailed breakdown and explanation of these two types. Specific categories will be examined in greater depth in the subsequent chapters.

Methods based on the Training Process. This type can be divided into two classes: methods based on training data augmentation and alignment with instruction tuning.

(1) *Training data augmentation*: For LLMs, biases frequently originate from imbalanced data distribution and poor data quality (Gallegos et al., 2023). One of the most direct and effective solutions is improving the quality, diversity, and balance of training data. Data augmentation techniques aim to mitigate biases by introducing additional instances into the training data, thereby increasing the data points related to underrepresented or misrepresented social groups. Data balancing approaches aim to achieve equitable distribution across various social groups. One primary technique for this purpose is Counterfactual Data Augmentation (CDA) (Lu et al., 2020; Qian et al., 2022; Webster et al., 2020), which involves replacing protected attribute words, such as gendered pronouns, to create a balanced dataset.

Inspired by the mixup technique (Zhang et al., 2017a), interpolation approaches blend counterfactually augmented training instances with their original counterparts and labels, thereby achieving a more balanced distribution of the training data (Yao et al., 2022b;a; Yang et al., 2023e). In Ahn et al. (2022), the mixup framework is harnessed to align the output logits of a pre-trained model between two opposing words within a gendered pair. In Mix-Debias, Yu et al. (2023b) apply mixups across various corpora, aiming to alleviate gender stereotypes by leveraging an augmented training set.

Wang et al. (2022b) introduce an automated iterative framework that prompts LLMs in conjunction with a filtering criterion. Through a self-instructive process, this framework reconstructs a more diverse dataset from initial seed data tailored for LLMs' instruct tuning. The prompts for this dataset are generated automatically by LLMs and undergo various metric-based filtering to ensure diversity in the dataset.

In addition, there are numerous data filtering methods (Garimella et al., 2022; Borchers et al., 2022; Thakur et al., 2023) that aim to enhance the balance of data distribution by either removing low-quality data or selectively retaining a diverse and underrepresented set of data.

(2) *Better alignment with instruction tuning*: With a vast amount of data, LLMs typically undergo pre-training and instruction tuning. In the pre-training phase, LLMs internalize knowledge from the training data into trainable parameters. Instruction tuning, on the other hand, teaches the model to understand human instructions. However, it is essential to recognize that biases in the training data and the training process are not inherently designed to understand or prioritize human values. This limitation leads to biases and potentially toxic responses from LLMs when faced with complex and divergent human preferences. They often arise naturally from the data or model training procedures, or from human design decisions that reflect their own values and preferences. To address this challenge, we provide a detailed overview of some current alignment techniques and training algorithms to harmonize LLMs with human preferences in Section 4.

Methods based on Post-processing Techniques. Another approach is based on post-processing techniques. Post-processing, in the context of LLMs, generally refers to the practice of invoking external knowledge bases or employing word-based detection techniques to identify biased statements during inference. Subsequently, the identified biases are corrected in the generated text. In Kang et al. (2023), techniques such as retrieval are employed within LLMs to match responses during each phase of the Chain of Thought (CoT) generation (Wei et al., 2022b). This involves retrieving and correcting biased or toxic text at every stage of the LLM's responses, thereby ensuring that LLMs produce accurate and unbiased text responses throughout the CoT process. Andriopoulos & Pouwelse (2023) also mention various methods that enhance LLMs by invoking Wikipedia and various external knowledge bases for retrieval. The objective of these approaches is to boost the reliability of LLM outputs and reduce biases in generated text. Additionally, word-level detection is employed in Chen et al. (2023k) to identify instances in LLM responses involving counterfactual information or not aligning with the context. Subsequently, a post-processing approach is applied to correct and enhance LLM's reliability by removing such inaccuracies from the generated text. On the other hand, Li et al. (2024c) synthesize cultural-specific instruction data to incorporate cultural differences into LLMs. Raza et al. (2024) further propose MBIAS, a LLM framework instruction fine-tuned on a custom dataset designed explicitly for safety interventions. Moreover, Wang & Demberg (2024) introduces a multi-objective probability alignment approach to overcome current challenges by incorporating multiple debiasing losses to locate and penalize bias in different forms, which is more effective in removing stereotypical bias of LLMs while retaining their general performance.

Overall, the methods based on post-processing techniques can effectively and accurately handle certain biased information. However, they also have certain drawbacks. For instance, when relevant information is not present in external knowledge bases, biases in LLMs' responses might remain uncorrected. Additionally, post-processing may introduce erroneous information from external knowledge bases. Moreover, approaches relying on post-processing techniques often lead to a significant increase in latency.

3.4 Bias and Fairness in MLLMs

Compared to LLMs, the emphasis on fairness in MLLMs is more direct toward ensuring that the responses align faithfully with the inputs in different modalities, such as images or audio in context. The responses must align consistently with the visual content, ensuring they are free from any biased text that contradicts the context. Currently, most research exploring bias and fairness in MLLMs is focused on the phenomenon of image hallucination. This term describes scenarios in which the model, when describing images or answering questions based on visual information, generates responses containing entities, quantities, or logical information that does not exist in the given image (Zhou et al., 2022a; Wang et al., 2023b;a; Li et al., 2023g; Liu et al., 2023d;e; Zhou et al., 2024g). In Section 7, we conducted a detailed analysis of recent advances in understanding and addressing hallucinations within MLLMs.

Apart from hallucinations, there is a notable lack of in-depth exploration into the bias and fairness of MLLMs. Similar to LLMs, MLLMs may exhibit significant biases due to the training paradigm and dataset distribution. The scarcity of image-text data for specific groups, coupled with the presence of biased information in the dataset, may lead MLLMs to acquire stereotypical impressions of certain groups. Additionally,

imbalanced dataset distribution might cause MLLMs to showcase biases in responses to specific image-text pairs. Earlier efforts on generating counterfactual images towards semantic textual concepts have shown that machine learning models will encode biases related to certain attributes if the training data is imbalanced (Luo et al., 2023b; Prabhu et al., 2023; Xia et al., 2023). To further mitigate biases during inference, BEND-VLM (Gerych et al., 2024) tailors the debiasing operation for MLLM embedding to each unique input at the test time, thereby avoiding catastrophic forgetting in fine-tuning. However, biases remain largely unexplored in MLLMs, and addressing fairness and bias in MLLMs is crucial for building foundation models that are beneficial and equitable for humanity.

3.5 Bias and Fairness in Text-to-Image Models

The rise of text-to-image models sparks discussions on systematic social bias and fairness issues in generated content, as indicated by several studies (Zhang et al., 2023i; Saharia et al., 2022; Cho et al., 2023; Bianchi et al., 2023; Luccioni et al., 2023; Li et al., 2024d). Text-guided diffusion models, in particular, have been found to exhibit biases related to professions, ethnicities, and social classes. The generated contents diverge from the distributions in the real world and even amplify the biases in real societies (Zhang et al., 2023i; Bianchi et al., 2023). For instance, a study conducted by Luccioni et al. (2023) highlights that text-to-image diffusion models consistently underrepresent marginalized identities in the generated images. Some examples of gender biases in text-to-image models are presented in Figure 5.

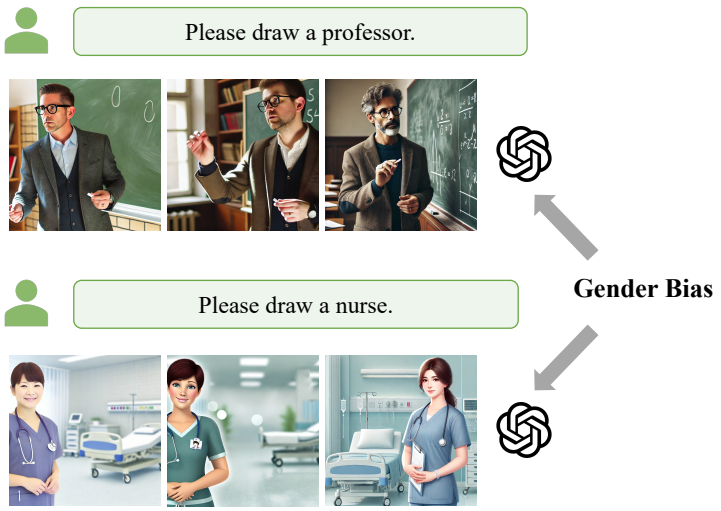

Figure 5: Examples of gender biases in T2I models: DALL · E shows a spurious correlation between gender and profession.

To overcome the systematic bias and fairness issues, many methods (Friedrich et al., 2023; Kim et al., 2023a; Chuang et al., 2023; Li et al., 2024d) focus on mitigating biases in text-to-image models through prompting techniques. Fair Diffusion (Friedrich et al., 2023) randomly injects additional subject pronouns in the prompts to achieve a more balanced gender distribution in the generated images. Other work (Kim et al., 2023a) optimizes the soft token in the prompts to induce a more balanced gender distribution. Furthermore, Chuang et al. (2023) work directly in the text embedding space to obtain a more balanced gender distribution in vision-language models. A recent study (Li et al., 2024d) addresses these problems by finding the bias-related concept in an interpretable latent space and manipulating the generation process with the concepts found. These prompt-based regulations are by far the most widely adopted strategy to reduce biases in the generated content. However, it has been noted that keyword-based approaches could disproportionately affect marginalized groups, implying that their use at the prompt level could yield similar outcomes (Dodge et al., 2021).

Another direction of research involves addressing biases through sampling methods. For example, the D2C method (Sinha et al., 2021) generates unconditional diffusion via few-shot conditional diffusion to balance the numbers in generated classes. Furthermore, Fair Sampling (Choi et al., 2024) introduces a fairness-aware sampling technique aimed at reducing the amplified biases inherent in training data.

3.6 Current Limitations and Future Directions

Despite significant advancements in the domain of bias and fairness in foundation models, there are still some limitations in bias and fairness evaluation that require future attention.

3.6.1 Limitations and Open Challenges of Bias and Fairness

Currently, most bias evaluation methods are limited to token or paragraph-level assessments, making it challenging to capture the gradual propagation of bias during autoregressive generation (Xiao et al., 2023b; Schmidt, 2019; Zollo et al., 2024c). In autoregressive models, each token's prediction relies on previously generated tokens, meaning that biases may accumulate and spread over time. Traditional token or paragraph-level fairness metrics (Chalkidis et al., 2022; Baumgartner et al., 2024) are insufficient to fully assess this bias propagation issue, making it difficult to accurately measure the biases in these model outputs.

Additionally, when dealing with specific content, such as social media posts or content related to current events, the concept of bias becomes more complex. Biases in such cases may not always be evident or confined to a single token but may be reflected through subtle contextual influences or narrative frameworks. Therefore, bias concerns more than just token-level differences; it also involves how the model handles historical, social, or cultural influences, which may be embedded in the model's training data. This presents a significant challenge for mitigating biases, as it often intertwines with factual reporting and socially accepted norms.

When evaluating biases in foundational models, the lack of clear and consistent definitions of fairness within these models complicates both assessment and improvement efforts (Doan et al., 2024; Sheng et al., 2024; Zhang et al., 2023c). What is considered fair can vary significantly depending on cultural, social, and historical perspectives. This variability becomes especially pronounced when dealing with news content, where fairness often intersects with historical accuracy or the presentation of current facts. Models must therefore navigate a delicate balance: striving for unbiased outputs while carefully weighing the tension between fairness principles and accurately representing reality.

For example, in reporting on historical events or current issues, there may be cases where acknowledging certain inequalities or biased social structures is necessary to present the facts accurately. In such situations, pursuing absolute fairness might mean overlooking or distorting facts, leading to a significant conflict between fairness and authenticity. For model developers, maintaining the integrity of generated or processed information while addressing ethical concerns is a significant challenge. Furthermore, when fairness considerations span different regions, cultures, and social norms, the complexity of such evaluations is exacerbated, increasing the difficulty of implementing fairness assessments in foundational models. In addition, societal biases are challenging to mitigate with common techniques such as data resampling (Hirota et al., 2024), and addressing them in web-scale datasets remains an open problem.

3.6.2 Future Directions

Addressing the limitations of bias and fairness in current foundation models opens several avenues for future research. Firstly, exploring unbiased tokenization and embedding methods could help mitigate the introduction of bias during natural language processing (Phan et al., 2024; Zhang et al., 2020). This involves developing tokenization techniques and embedding representations that maintain fairness at a more fine-grained level. Secondly, in terms of unbiased fine-tuning and preference learning, employing techniques such as constrained RLHF (Yu et al., 2024c), DPO (Zhou et al., 2024h; Wang et al., 2024j; Zhou et al., 2024f), and revised LoRA methods (Liu et al., 2024g; Xu et al., 2025b) can help to adjust the model's training process to reduce bias introduced during generation. As models are made more secure through methods like adversarial training, future work must also ensure these security protocols are designed to be fairness-aware, preventing the accidental amplification of biases as a side effect. Furthermore, the post-processing of autoregressive generation is an important area of focus (Zhou et al., 2024g), which can help further detect and correct potential biases in the generated content. This challenge of equitable evaluation extends to related tools, creating a need to audit detectors of AI-generated content (AIGC) to ensure they do not unfairly penalize content from specific demographic groups. Moreover, explainability methods offer a promising frontier for more targeted interventions, opening research avenues into using mechanistic interpretability to locate and edit the specific model circuits that encode stereotypes. In MLLMs, individual modalities can have a disparate impact on bias and fairness and may require modality-specific interventions (Weng et al., 2024). Lastly, balancing fairness and utility is crucial, as striving for absolute fairness often conflicts with the model's practical utility

and performance. Therefore, developing effective trade-offs that simultaneously address fairness and utility will be a significant challenge for future research.

4 Alignment

Foundation models have significantly expanded their functionality, advancing beyond simple content generation to a wide range of applications including strategic planning (Huang et al., 2023c; Song et al., 2023a; Liu et al., 2024k), code generation (Chen et al., 2021b; Poesia et al., 2022), tool integration (Qin et al., 2023a; Shen et al., 2023c), complex reasoning (Wei et al., 2022b; Huang & Chang, 2023), and even addressing challenges in natural sciences, especially mathematics (Drori et al., 2022; Imani et al., 2023; OpenAI, 2023b). Despite these advancements, it is important to note that foundation models are primarily trained on large datasets with objectives such as next-token prediction (Radford et al., 2018), next-scale prediction (Tian et al., 2024), or diffusion (Lipman et al., 2022). They are not inherently equipped to understand or prioritize human values and preferences.

This gap between their powerful capabilities and inherent limitations underscores the potential risks associated with their deployment. For example, without proper safeguards, foundation models could be jailbroken by users to disclose personal information or engage in harmful behaviors, which should be avoided (Li et al., 2023c; Taveekitworachai et al., 2023; Shen et al., 2023b; 2024; Chen et al., 2024d;a). Moreover, the ability of AI agents to adapt their capabilities to diverse objectives (e.g., scientific discovery and management systems) further highlights the importance of thoughtful oversight. According to the orthogonality thesis (Bostrom, 2012), AI systems can pursue any number of goals, regardless of their intelligence level. This concern is compounded by the instrumental convergence thesis (Bostrom, 2012), which suggests that regardless of their ultimate goals, AI systems might adopt certain potentially harmful strategies as means to achieve them—such as self-preservation or resource acquisition, which could lead to power-seeking behaviors (Bostrom, 2012; Burns et al., 2023). As the capabilities of foundation models advance, it becomes crucial to carefully design these models to align with human-centric values and the nuanced requirements of specific tasks. This alignment is essential for ensuring the deployment of foundation models meets rigorous safety and ethical standards in various real-world applications.

In the following section, we will focus on aligning LLMs with human preferences. We not only reviews core techniques like Supervised Fine-Tuning and RLHF but also frames them as a crucial component in a larger ecosystem of responsible AI development, with direct intersections with uncertainty and hallucination. We will begin with Supervised Fine-Tuning (Wei et al., 2021; Zhou et al., 2023a; Chung et al., 2024), proceed to Reinforcement Learning from Human Feedback (Ouyang et al., 2022; Christiano et al., 2017; Bai et al., 2022a), and then explore Prompt Engineering (Liu et al., 2023k; Gu et al., 2023). Furthermore, we will extend our discussion to MLLMs in Section 4.4. Finally, we will discuss the limitations of current alignment methods in Section 4.5. These categories can be visualized in Figure 6.

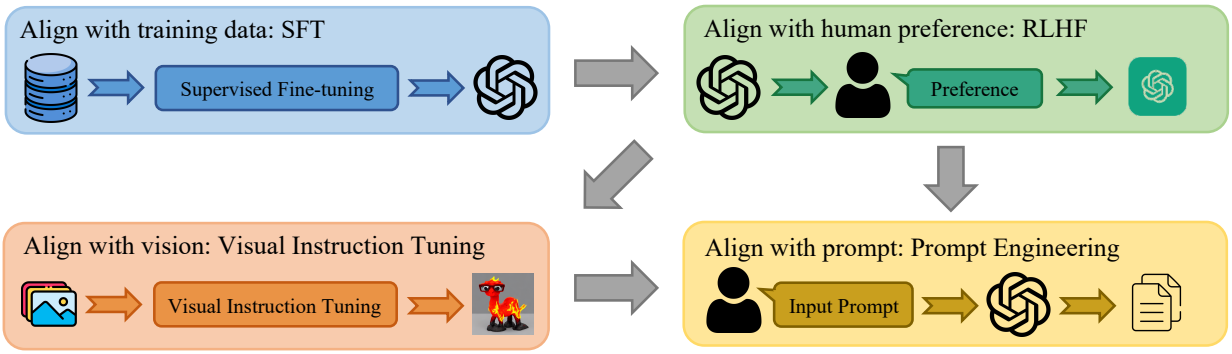

Figure 6: Alignment is required at different stages in the foundation models. Typically, LLMs are aligned using SFT and RLHF during post-training, while using prompt engineerfing at inference time. Compared to LLMs, MLLMs require an additional step of multimodal alignment at post-training, such as Visual Instruction Tuning.

4.1 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) is a widely-used approach to align pre-trained LLMs with human preferences, which directly tunes the LLM f to mimic desired ground-truth responses. It often serves as the first stage of the alignment process. Most SFT methods can be formally expressed as:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{L} \log P_f\left(r_t \mid \mathbf{p}, \mathbf{r}_{<t}\right), \tag{11}$$

where $\mathbf{p}$ denotes the input prompt and $\mathbf{r} = (r_1, r_2, ..., r_L)$ is the sequence of the target response. This approach maximizes the likelihood of generating the optimally selected response, akin to how a student learns from a teacher's guidance. Combined with other training methods for alignment, SFT can often enhance the stability of the whole alignment process.

Although SFT is efficient in aligning LLMs, its success heavily relies on the quality and diversity of the training data. LIMA (Zhou et al., 2023a) presents a study that highlights the importance of this aspect, where the authors curate a dataset of 1,000 high-quality prompt-response pairs, with 750 of them coming from diverse sources such as StackExchange[1], wikiHow[2], and the Pushshift Reddit Dataset (Baumgartner et al., 2020), while the remaining 250 pairs are manually annotated. LIMA demonstrates that LlaMa-65B (Touvron et al., 2023), when fine-tuned on a small but high-quality dataset using the regular SFT training objective, can achieve significant performance improvements without requiring reinforcement learning or explicit human preference modeling. In this line of research, methods based on Minihash (Broder, 1997) and Local Sensitive Hashing (LSH) (Datar et al., 2004) are often used to deduplicate the data, which serve as the first step of refining data quality. Then, a series of works (Zhou et al., 2020a; Penedo et al., 2023; Rae et al., 2022; Wang et al., 2023e; Chen et al., 2024c) propose to use well-suited rules, metrics, and LLMs-based methods for further data cleaning. These work leverage measure the quality of the data and perform improvements such that deduplication and rewriting to improve the overall quality. Li et al. (2023e) introduces the instruction following difficulty metric for efficient data selection. Similarly, Xie et al. (2023a) estimates importance weights for high-quality data selection.

Moreover, many empirical results of applying scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) to LLM training show that the data size becomes crucial for performance improvements. To reduce the cost of human annotations, researchers are increasingly interested in incorporating AI-generated data into the alignment process, especially with the advent of closed-source LLMs like GPT-4 (OpenAI, 2023b), Gemini 2.0 (Team, 2024c), and Claude 3.5 (Team, 2024b). A line of research explores using LLMs to self-generate instruction-tuning data. A notable advancement in this domain is the Self-Instruct (Wang et al., 2022c), which leverages the in-context learning (ICL) capabilities of GPT-3 (Brown et al., 2020) to gather instructions and preferred responses autonomously. This approach begins with a small set of human-annotated seed instructions that are subsequently refined and expanded to generate large-scale instruction data across diverse tasks. Building on this methodology, researchers have achieved significant advances in developing open-source LLMs with enhanced instruction-following capabilities, such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023). Gunasekar et al. (2023) propose to use a mix of "textbook quality" data from the web and GPT-3.5 generated data to train the Phi, a lightweight LLM suitable for edge scenarios (Xu et al., 2021b). It also pioneers large-scale, high-quality data generation. More recently, Xu et al. (2024b) proposed a self-synthesis method that leverages the auto-regressive nature of LLMs to generate diverse data without requiring any initial seed question or prompt. Zhou et al. (2023d) propose to synthesize natural language descriptions for controllable text generation (Hu et al., 2018; Sun et al., 2023a). Wang et al. (2024f) introduce a method for synthesizing role-playing data using LLMs with carefully curated role descriptions. Similarly, Qiao et al. (2024) present an approach for synthesizing agent-tuning data via self-planning with LLMs; Ulmer et al. (2024b) generates a training data via "self-talk" of LLMs which can be used for further supervised finetuning.

Nevertheless, the simplicity of SFT does not shield it from potential vulnerabilities, especially in terms of model safety and robustness. Qi et al. (2023b) illustrate that LLMs, such as OpenAI's GPT-3.5 Turbo (Ope-

[1] https://stackexchange.com/
[2] https://www.wikihow.com/

nAI, 2023a), are prone to adversarial manipulations. They demonstrate that fine-tuning these models with a limited set of strategically crafted examples from the Anthropic red team dataset (Ganguli et al., 2022) can significantly undermine the model's safety protocols. This phenomenon highlights the necessity for a rigorous examination in selecting and preparing SFT data. Further Springer et al. (2025) highlights that overtrained LLMs impose challenge in further supervised finetuning.

In conclusion, while SFT demonstrates notable efficiency, its effectiveness in aligning LLMs hinges critically on the quality, scale, and diversity of training data. The critical role of a meticulously curated dataset extends beyond improving model performance; it is also vital to mitigate risks related to model safety and robustness. Inadequately vetted or deliberately compromised data can introduce harmful biases and trigger undesirable behaviors in LLMs. This concern highlights the ongoing necessity for rigorous data curation methods to enhance both the reliability and security of LLMs in real-world deployments.

4.2 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2017; Bai et al., 2022a; Zhang et al., 2024l) represents a significant advancement in aligning LLMs with human preferences, involving several critical steps:

1. Supervised Fine-Tuning (SFT) is performed on a pre-trained model using a high-quality, instruction-following dataset. The resulting model serves as the initial policy, π_{SFT}, for the subsequent RLHF optimization.

2. Collection of pairwise ranking data to train a reward model that correctly scores these data.

3. Optimization of the policy model obtained in Step 1 against the reward model in Step 2 using the Proximal Policy Optimization (PPO) (Schulman et al., 2017).

To stabilize the optimization in Step 3, KL-divergence regularization is introduced (Ouyang et al., 2022), ensuring that the model remains reasonably close to the initial policy model acquired in Step 1.

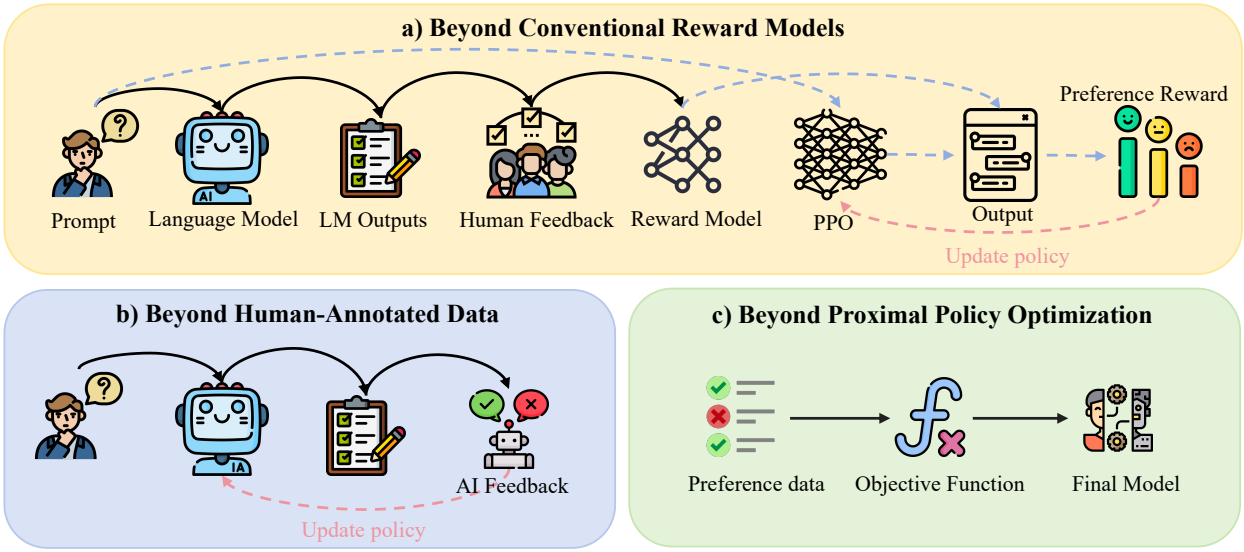

Figure 7: The evolution of Reinforcement Learning from Human Feedback (RLHF), illustrating three key areas of advancement. Panel (a) depicts the conventional RLHF pipeline and the move Beyond Conventional Reward Models. Panel (b) shows the shift Beyond Human-Annotated Data to include AI-generated feedback. Panel (c) illustrates the trend of moving Beyond Proximal Policy Optimization (PPO) towards simpler, direct preference optimization objectives.

Beyond Conventional Reward Models. The effectiveness of RLHF is closely linked to the accuracy and robustness of the reward model. Recent research has identified biases in reward models (Shen et al., 2023a; Leng et al., 2024) and has focused on refining traditional Bradley-Terry reward models (Bradley & Terry, 1952). Wu et al. (2023h) introduce a fine-grained RLHF framework that addresses the challenges of translating human preferences into scalar learning signals for extensive textual outputs. This approach utilizes multiple fine-grained reward models and has demonstrated superior performance in tasks such as detoxification and extended question-answering. Complementing this, Rame et al. (2023) propose "rewarded soup", which linearly interpolates weights across specialized networks to derive diverse rewards. It emphasizes the importance of a varied reward structure and aims to achieve Pareto-optimal generalization across the complete preference space. An additional generation of reward modeling, which is referred to as the "general preference" approach, directly learns a pairwise preference function and seeks a model that identifies the Nash equilibrium of an entropy-regularized minimax game (Munos et al., 2023; Ye et al., 2024). This strategy draws inspiration from the classical dueling bandit problem (Yue et al., 2012; Zoghi et al., 2014). Zhou & Xu (2020) propose to train a comparative evaluation model based on annotated pairwise preference data and use it to train a TextGAN (Zhang et al., 2017b) with RL in Zhou et al. (2020b), this can be viewed as an early version of RLHF. Zhu et al. (2023a) extend pairwise ranking by considering the ranking of multiple responses and trains the reward model using K-wise maximum likelihood (Zhu et al., 2023b). Additionally, beyond the single reward model, another approach considers the joint preferences implied by multiple reward functions, such as "helpfulness, harmfulness, verbosity", etc. Some of these reward functions may conflict with each other. The objective here is to strike a balance among various rewards, reflecting diverse user preferences (Dong et al., 2023b; Zhou et al., 2023e; Wang et al., 2024b; Chen et al., 2024h; Chakraborty et al., 2024). This diversity is not merely individual but also deeply cultural, as preferences and values can vary significantly across different populations, making a single, universal alignment target an ill-posed problem (Kirk et al., 2024).

Recently, the success of DeepSeek R1 (Guo et al., 2025) and other reasoning models (Team et al., 2025) such as OpenAI o1 (OpenAI et al., 2024) on closed-end domains such as mathematical and code reasoning demonstrate the importance of verified rewards in large scale reinforcement learning optimization, which also highlights the serious reward hacking problem. Despite these adjustments to the reward model, such methodologies remain firmly within the overarching RLHF framework.

Beyond Human-Annotated Data. Synthetic data generation has proven effective for SFT. However, when it comes to RLHF, pairwise preference ranking data is typically collected through human annotations, a process that can be costly for scaling. To mitigate this issue, recent research has shown that AI-generated data can also provide helpful feedback for alignment. Bai et al. (2022b) introduce "RL from AI Feedback" (RLAIF), which blends human and AI preferences under the "Constitutional AI" (CAI) framework. In this framework, AI behaviors are governed by principles analogous to a constitution, supported by a few examples for few-shot prompting. This methodology aims to train a non-evasive AI assistant that is effective and harmless without relying solely on human labels. Further extending the concept of RLAIF, Lee et al. (2023a) apply it to summarization tasks, while Wang et al. (2023i) adapt it for complex reasoning tasks, highlighting the potential of AI feedback. Additionally, Guo et al. (2024a) enhance the RLAIF paradigm with online AI feedback, demonstrating superior performance in model alignment compared to both offline RLAIF and traditional RLHF. Zhang et al. (2024e) propose Self-Exploring Language Models (SELM) to elicit preferences for online alignment actively. Wang et al. (2024i) propose Constitutional DPO, which uses expert annotated principles to synthesize negative examples for preference learning.

Beyond Proximal Policy Optimization. While RLHF has proven effective in capturing human preferences, the PPO algorithm (Schulman et al., 2017) typically requires complex implementations and substantial computational resources, limiting its applicability in various contexts. The key challenges in PPO training include filtering high-quality data to compare similar responses, managing policy and reward models within limited resources, mitigating reward hacking issues (Lu et al., 2024b; Eisenstein et al., 2024; Ramé et al., 2024), and requiring extensive hyperparameter and training strategy adjustments. To address these challenges, various new preference optimization objectives have been proposed, and their corresponding objective functions are presented in Table 2. Dong et al. (2023a) introduce the Reward Ranked FineTuning (RAFT) that simplifies the complexity of PPO by using a reward model to selectively focus on the most promising

Table 2: Various preference optimization objectives given the preference data $\mathcal{D} = (x, y_w, y_l)$, where x is an input, y is an output, y_w and y_l are the winning and losing responses, and $y^i, i \in [n]$ are ranked responses.

Method	**Objective**
RAFT (Dong et al., 2023a)	$\max_w \mathbb{E}_{x\sim D, y\sim p_g(\cdot\|w,x)}[r(x,y)]$
RRHF (Yuan et al., 2023b)	$\mathcal{L}_{\text{sft}} + \sum_{i>j} \max\left[0, \pi(y^i\|x) - \pi(y^j\|x)\right]$
ReST (Gulcehre et al., 2023)	$\max \mathbb{E}_{x\sim\mathcal{D}}\left[\lambda\mathbb{E}_{y\sim\pi_{\theta'}(y\|x)} F(x,y;\tau)\nabla \log \pi_\theta(y\|x)\right]$ $+(1-\lambda)\mathbb{E}_{y\sim p(y\|x)}\left[F(x,y;\tau)\nabla \log \pi_\theta(y\|x)\right]$
SLiC-HF (Zhao et al., 2023d)	$\max\left(0, \delta - \log \pi_\theta(y_w\|x) + \log \pi_\theta(y_l\|x)\right) - \lambda \log \pi_\theta(y_w\|x)$
DPO (Rafailov et al., 2023)	$-\log \sigma\left(\beta \log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - \beta \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
PRO (Song et al., 2023b)	$\beta\mathcal{L}_{\text{sft}} - \sum_{k=1}^{n-1} \log \frac{\exp\left[\frac{\pi(y^k\|x)}{1/(r^*(x,y^k)-r^*(x,y^n))}\right]}{\frac{\pi(y^k\|x)}{1/(r^*(x,y^k)-r^*(x,y^n))}+\sum_{i=k+1}^{n}\exp\left[\frac{\pi(y^i\|x)}{1/(r^*(x,y^k)-r^*(x,y^i))}\right]}$
IPO (Azar et al., 2023)	$\left(\log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)} - \frac{1}{2\tau}\right)^2$
KTO (Ethayarajh et al., 2024)	$-\lambda_w\sigma\left(\beta \log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - z_{\text{ref}}\right) + \lambda_l\sigma\left(z_{\text{ref}} - \beta \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$, where $z_{\text{ref}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\beta \text{KL}\left(\pi_\theta(y\|x)\|\pi_{\text{ref}}(y\|x)\right)\right]$
ORPO (Hong et al., 2024a)	$\mathcal{L}_{\text{sft}} - \lambda \log \sigma\left[\log \frac{\pi(y^w\|x)(1-\pi(y^l\|x))}{\pi(y^l\|x)(1-\pi(y^w\|x))}\right]$
RPO (Liu et al., 2024l)	$\min \eta\beta \cdot \mathbb{E}_{x\sim d_0, y^0\sim\pi^{\text{base}}(\cdot\|x)}\left[-\log(\pi_\theta(y^0\|x))\right]$ $+\log \sigma\left(\beta \log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - \beta \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
SimPO (Meng et al., 2024)	$-\log \sigma\left(\frac{\beta}{\|y_w\|} \log \pi_\theta(y_w\|x) - \frac{\beta}{\|y_l\|} \log \pi_\theta(y_l\|x) - \gamma\right)$

responses sampled from an LLM. Specifically, RAFT involves sampling a large batch of instructions and generating multiple responses. These responses are then holistically ranked by the reward model, with only the top-ranked responses used in SFT. This process is iterated until the rewards stabilize, and the fine-tuning dataset is periodically updated to enhance its quality. In parallel, Yuan et al. (2023b) introduce Reinforced Ranking Human Feedback (RRHF), which aligns the model with human preferences among diverse responses using a likelihood ranking loss. This method facilitates the integration of data from multiple sources, including both model-generated and human-curated data.

Another innovative approach within the RLHF framework is Reinforced Self-Training (ReST), introduced by Gulcehre et al. (2023). ReST focuses on iteratively generating and refining data from policy models optimized by offline RL algorithms, enhancing data utilization efficiency. The framework involves two main steps: "Grow" and "Improve". In the Grow step, the policy model generates multiple outputs for augmentation. During the Improve step, the generated data is ranked and filtered by a preference reward model, after which the policy model is fine-tuned on the filtered data using an offline RL objective. This process is repeated with an increased filtering threshold to further refine data quality. Beyond zeroth-order RL algorithms such as PPO, which require the learning of the value function and hyperparameter tuning, first-order RL algorithms (Zhang et al., 2024d; 2023g; Gao et al.) can also act as a straightforward alternative for RLHF alignment. Sequence Likelihood Calibration (SLiC) by Zhao et al. (2022; 2023d) aims to align model outputs with reference sequences in the latent space by calibrating the sequence likelihood. This method replaces the traditional embedding similarity function with a preference ranking function and employs a cross-entropy regularization loss to keep the model close to the reference, typically an SFT model.

Additionally, Song et al. (2023b) propose the Preference Ranked Optimization (PRO) method. Differing from traditional RLHF approaches that use the Bradley-Terry reward model focusing only on the best and worst responses, it enumerates all possible ranking pairs among candidate responses to provide comprehensive alignment.

Based on these insights, Rafailov et al. (2023) propose Direct Preference Optimization (DPO), which integrates preference information indirectly into the optimization of the policy model, eliminating the need for a separate reward function. The DPO loss derived from the reward maximization-based RLHF algorithms is used to directly optimize the policy model π_θ as follows:

$$\mathcal{L}_{\mathrm{DPO}}\left(\pi_\theta; \pi_{\mathrm{ref}}\right) = \log \sigma \left(\beta \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)} - \beta \log \frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\mathrm{ref}}\left(y_l \mid x\right)} \right), \tag{12}$$

where π_{ref} denotes the reference policy (namely the SFT model), and (x, y_w, y_l) represents the instruction x paired with the preferred answer y_w and the dispreferred answer y_l. Furthermore, Ethayarajh et al. (2024) propose Kahneman-Tversky Optimization (KTO) to directly maximize the utility of LLM's generations instead of the likelihood of preferences. Unlike methods that require costly annotated pairwise ranking data, KTO only needs individual binary feedback, which is easier to collect from real users. Besides, Regularized Preference Optimization (RPO) (Liu et al., 2024l) is proposed to mitigate reward hacking or overoptimization issues during RLHF by simply adding the SFT loss to DPO. Azar et al. (2023) theoretically analyze the weakness of DPO and introduce Identity Preference Optimization (IPO), which adds a constant regularization term to the DPO loss to mitigate the overfitting problem. SimPO (Meng et al., 2024) introduces a margin term to the Bradley-Terry objective and incorporates the length normalization, eliminating the common need for an additional reference model in preference learning. Similarly, Hong et al. (2024a) propose ORPO, which integrates an odds ratio preference objective into the standard SFT objective, also functioning independently of a reference model. In addition to these direct preference optimization methods, some recent works focus on improving the efficiency of large-scale reinforcement learning in optimizing LLMs. Guo et al. (2025) propose Group Relative Policy Optimization (GRPO) to eliminate the needs of the critic model and value estimation in PPO, which greatly saves computation resources. GRPO estimates the baseline of advantage by calculating the average rewards of multiple samples in the group and replaces the KL penalty added to the reward by explicitly adding the KL divergence to the loss. The success of DeepSeek R1 (Guo et al., 2025) also indicates the efficiency and performance of applying GRPO in large-scale reinforcement learning for LLMs. Similarly, REINFORCE++ algorithm (Hu, 2025) removes the critic model via implementing token-level KL penalty, PPO's clipping for policy model updates, and normalized advantages while achieving efficient reinforcement learning for LLMs.

The proliferation of these alternatives to PPO raises the practical question of which objective to choose. The decision often involves a trade-off between complexity, data requirements, and robustness. Methods like DPO, ORPO, and SimPO offer simplicity and efficiency by forgoing an explicit reward model, making them suitable for resource-constrained scenarios. However, these direct methods may be more susceptible to overfitting on the preference dataset. In contrast, full PPO-based RLHF, while more complex, allows for online exploration and can lead to more robust models, particularly when paired with high-quality, verified rewards as seen in specialized domains.

4.3 Prompt Engineering

While some research focuses on aligning LLMs with human preferences through explicit training, another line of research emphasizes the strategic design of prompts to effectively improve the LLMs' generated responses. This approach, known as Prompt Engineering, involves crafting prompts that guide LLMs toward fulfilling specific task requirements.

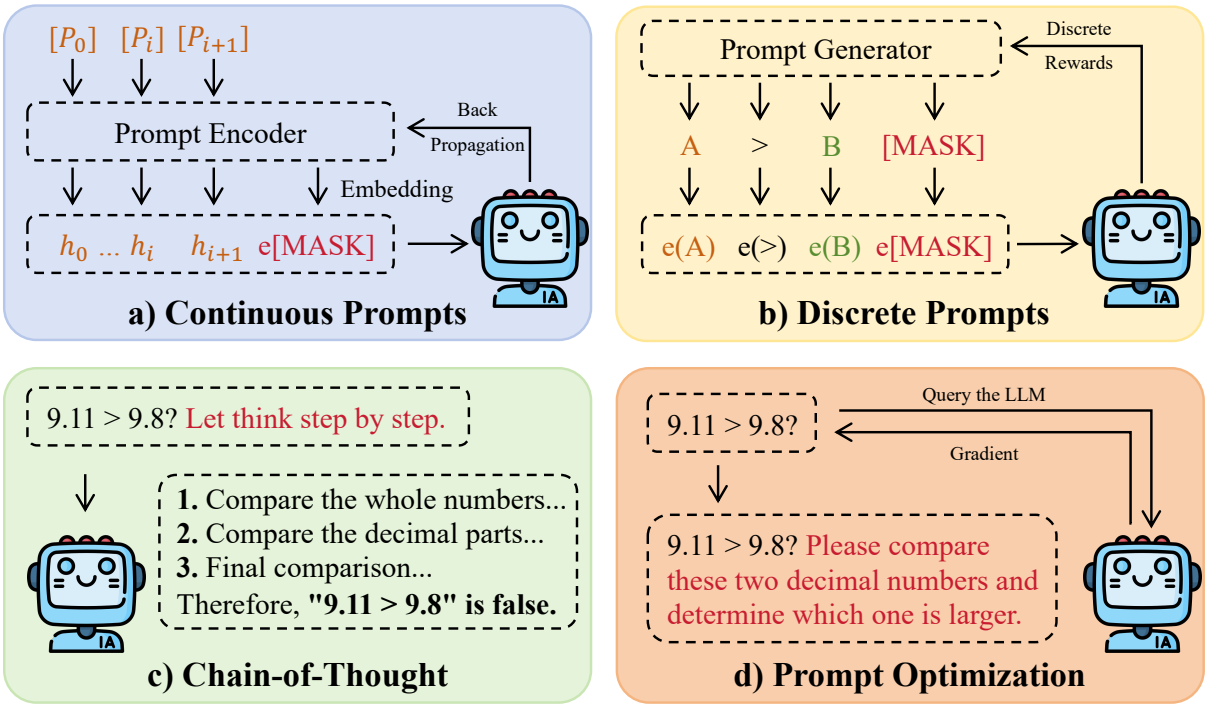

Figure 8: An overview of four major prompt engineering methods for guiding LLMs. (a) Continuous Prompts are soft prompts represented as tunable vectors in the embedding space. (b) Discrete Prompts are interpretable natural language prompts that can be automatically generated. (c) Chain-of-Thought prompting elicits intermediate reasoning steps to solve complex problems. (d) Prompt Optimization uses feedback from the LLM itself to iteratively refine prompts and improve performance.

Continuous Prompts. In the field of prompt engineering, continuous prompts, represented as continuous vector inputs integrated into the LLM, offer a novel approach to guiding AI responses. Due to the continuous nature, these prompts can often be fine-tuned through gradient-based methods using labeled data (Qin & Eisner, 2021; Ding et al., 2021; Lester et al., 2021; Hambardzumyan et al., 2021; Liu et al., 2023l; Hao et al., 2024b), which is effective for adapting LLM's behavior and injecting domain knowledge. However, the continuous format makes it less transparent for human understanding, posing challenges in interpretability (Khashabi et al., 2021; Hambardzumyan et al., 2021).

Discrete Prompts. In contrast to continuous prompts, discrete prompts consist of discrete tokens from the natural language vocabulary. The distinct advantage of discrete prompts lies in their use of natural language, which makes them inherently interpretable and relatable to humans. These prompts can be manually crafted or automatically generated. To automatically design effective prompts, some work utilizes pre-defined rules and reinforcement learning methods for searching (Gao et al., 2021; Hu et al., 2022b; Deng et al., 2022; Zhang et al., 2024f). However, this approach is not without its challenges, as models can be extremely sensitive to minor, semantically irrelevant changes in prompt formatting—a brittleness that can lead to unpredictable performance variations (Sclar et al., 2023).

Chain-of-Thought. For complex reasoning tasks, Wei et al. (2022b) first propose the Chain-of-Thought (CoT) method to encourage LLMs to generate a series of intermediate reasoning steps before reaching the final answer. This method has shown notable success in improving the performance of LLMs on tasks requiring multi-step reasoning, arithmetic reasoning, logical deduction, or commonsense application (Wei et al., 2022b; Wang et al., 2023h; Fu et al., 2022; Kojima et al., 2022; Chen et al., 2023c). Based on CoT, Tree-of-Thought (ToT) (Yao et al., 2023a) extends the concept to planning and decision. ToT refines the CoT method by utilizing the specific attributes of problems to decompose and organize intermediate thoughts into a tree structure. In ToT, each "thought" builds a node in this tree, facilitating explorations by searching algorithms such as the breadth-first or depth-first search, allowing lookahead and backtracking in problem-solving. This method has been further improved by Graph-of-Thought (GoT) (Besta et al., 2023), which diverges from linear or hierarchical structures to a more flexible graph-based representation. In GoT, the generated thoughts are forming nodes in a graph, with edges representing their complex interdependencies. This graph-based approach can capture the multifaceted nature of reasoning processes, offering improved adaptability for tasks such as sorting and keyword identification. Besides, it can also improve the latency and throughput compared to CoT and ToT. Recently, large reasoning models such as OpenAI o1 (OpenAI et al., 2024) and DeepSeek R1 (Guo et al., 2025) are believed to use long CoT to further enhance the model to plan and reason. The long CoT is pushing the length of the generated CoT to thousands or even hundreds of thousands of tokens, which effectively scales the test time computing (OpenAI et al., 2024; Guo et al., 2025; Team et al., 2025; Team, 2025d; Snell et al., 2024; Wu et al., 2024; Yang et al., 2025a). It is crucial to note, however, that these generated reasoning chains may not faithfully reflect the model's actual computational process. They can be post-hoc rationalizations rather than a true trace of the model's 'thought' process, a phenomenon that complicates their use for interpretability (Turpin et al., 2023)

Prompt Optimization. Complementing these structural prompt methods, recent research explores optimizing prompts directly using LLMs themselves. Yao et al. (2022c) propose ReAct to motivate LLMs to generate both reasoning traces and actions to interact with environments, to improve their general task-solving ability. Automatic Prompt Engineer (APE) (Zhou et al., 2022b) employs LLMs to craft initial instructions. Subsequently, APE cherry-picks instructions that exhibit the highest accuracy. Each of these selected instructions is then fed back into the LLM, prompting it to generate a variant that is semantically akin to the original instruction. Following a similar style, Automatic Prompt Optimization (APO) (Pryzant et al., 2023) iteratively refines existing instructions using textual feedback from LLMs. Conversely, Optimization by PROmpting (OPRO) (Yang et al., 2023a) adopts a more direct approach, generating new instructions at each optimization step, with LLM optimization focused on enhancing task accuracies without necessarily replicating prior instructions. Some LLMs have been shown to have the ability to use self-generated feedback to iteratively refine the output (Madaan et al., 2023). One can also apply derivative-free optimization techniques to optimize discrete prompt (Diao et al., 2022). Additionally, Guo et al. (2023b) propose EvoPrompt, which adopts evolutionary algorithms with LLMs for discrete prompt optimization. Beginning with a set of initial prompts, EvoPrompt applies evolutionary operators and performance-based selection to iteratively refine and generate new prompts. In addition to these explicit prompt optimizations, Li et al. (2023b) propose EmotionPrompt which focuses on understanding the psychological emotional stimuli of LLMs. It shows that simply appending emotional stimuli, such as "this is very important to my career", to the original prompts can also significantly enhance the performance of LLMs. Liu et al. (2023p) propose the first principled framework that has provable regret guarantees to orchestrate reasoning and acting with specially designed prompts. Zhou et al. (2024d) propose an agent symbolic learning framework to jointly optimize a chain of prompts (i.e., agent workflow; Zhou et al., 2023c) by mimicking back-propagation and gradient descent with natural language and LLMs. Beyond pure performance, prompt optimization can also be framed as a

risk control problem, where the goal is to identify prompts that are robust and minimize the likelihood of generating harmful or undesirable content (Zollo et al., 2023).

In the field of prompt engineering, the majority of methods focus on improving the performance of LLMs on specific tasks. While these methods are crucial for technical optimization, their contribution to aligning LLMs with human values and preferences is more indirect. By improving the interpretability of LLM outputs, these prompt engineering methods can gradually help LLMs better meet human expectations. This relationship between performance improvements and alignment with human values is an important consideration in the ongoing development of LLMs.

4.4 Alignment for MLLMs

Recent studies advocate the development of MLLMs capable of tackling various multimodal tasks without requiring particular adaptations. This approach leverages the well-established text-based capabilities of LLMs by integrating them, in a frozen state, as the language component within multimodal architectures, i.e., MLLMs, can align the visual and language modality through visual instruction tuning (Liu et al., 2023f), a specialized form of instruction tuning that extends the capabilities of pre-trained LLMs to understand and perform multimodal tasks involving both text and visual input. By incorporating datasets containing of vision-language instruction-following samples, this method enhances the zero-shot capabilities of LLMs for understanding and responding to visual inputs. The process typically employs linear projection layers to integrate image encoders with LLMs, allowing these models to effectively handle tasks that require an understanding of both text and images. Besides, extensive datasets comprising vision-language instruction tuning are utilized to align MLLMs with human preferences (Gao et al., 2023a; Gong et al., 2023; Li et al., 2023a; Liu et al., 2023e; Su et al., 2023b; Xia et al., 2024d;c; Li et al., 2024a; Tong et al., 2024a; Wang et al., 2024g). This approach allows MLLMs to accurately interpret instructions and generate user-friendly responses. Further works extend MLLMs to wider range of tasks such as generation (Lu et al., 2022; 2024a; Tong et al., 2024b; Xie et al., 2024a; Zhou et al., 2024a), and interactive agents (Zhai et al., 2024; Bai et al., 2022a; Zhou et al., 2024e; Xie et al., 2024b; Yang et al., 2025b; Shen et al., 2025).

To better align the MLLM's output with human preferences, some recent work (Sun et al., 2024b; Zhou et al., 2024h; Wang et al., 2024j; Liu et al., 2024c) aims to enhance model capabilities by filtering low-quality instruction data or constructing carefully examined examples during the fine-tuning phase. Recent studies (Chen et al., 2023d; Cao et al., 2023b; Paul et al., 2021; He et al., 2023) have introduced methods for evaluating the quality of instruction data in both vision and language datasets. These methods include computing the perplexity, calculating the gradient, and employing more powerful closed-source LLMs (e.g., GPT-4 (OpenAI, 2023b)) for rating, all aimed at filtering low-quality data from the training process. InstructionGPT-4 (Wei et al., 2023e) presents a more general data quality control pipeline by training a robust data selector to automatically select proper data from the raw dataset used to fine-tune MLLMs. DRESS (Chen et al., 2023i) proposes to divide natural language feedback (NLF) into critique and refinement types, and then utilize them to improve the alignment with human preferences and interaction capabilities of MLLMs. POVID (Zhou et al., 2024f) utilizes AI-generated dispreferred data by explicitly contrasting a hallucinatory answer with a truthful one, eliminating the need for gathering human feedback. Recent works such as STIC (Deng et al., 2024a), SIMA (Wang et al., 2024j), CSR (Zhou et al., 2024h), and AnyPrefer (Zhou et al., 2025b) explored the enhancement of the alignment between vision and text modalities through self-rewarding methods without introducing additional models and data.

4.5 Current Limitations and Future Directions

Though recent research has achieved remarkable success in aligning foundation models with human values and preferences by leveraging Prompt Engineering, Supervised Fine-Tuning, and Reinforcement Learning from Human Feedback, several challenges remain A prominent long-term challenge is the **superalignment problem**: how to ensure that AI systems much more intelligent than humans (i.e., superintelligence) remain aligned with human values and intentions (Burns et al., 2023). This is a difficult problem because humans may be unable to reliably supervise or evaluate the actions of a system that is far more capable than themselves.

Effectiveness of RLHF. Despite notable advancements in alignment brought by RLHF, this approach has its own challenges, as extensively analyzed by Casper et al. (2023). These challenges are broadly categorized into two types: tractability and generality. Tractability challenges encompass practical issues within the RLHF framework, such as difficulties in acquiring high-quality feedback (Chen et al., 2024e; Tong et al., 2025), risks associated with data poisoning, and inherent biases in the feedback. These issues, while significant, are considered manageable with the right strategies and improvements in future methods. On the other hand, generality challenges are more profound, raising critical issues about the overall effectiveness of RLHF. These include limitations in the human capacity to consistently provide accurate and reliable feedback for complex tasks, challenges in adequately modeling the diverse values of different human groups through reward models, and risks associated with reward hacking and power-seeking behaviors inherent in reinforcement learning systems. Though applying rule-based RL in the reasoning domain appears successful (Guo et al., 2025; Team et al., 2025; Team, 2025d; Yang et al., 2025a), it is challenging to directly adapt it to broader general domains to represent diverse and complex human values. For example, if a reward model cannot distinguish between nuanced responses and instead assigns uniformly high rewards to any agreeable or positive-sounding output, it may inadvertently train the foundation model to be sycophantic—agreeing with the user regardless of factual accuracy—rather than maintaining epistemic integrity. Such fundamental challenges pose critical questions about the long-term viability and ethical implications of relying solely on RLHF for aligning foundation models with human values.

Issues in Direct Alignment. Direct alignment methods such as DPO (Rafailov et al., 2023) greatly simplify the traditional RLHF pipeline and reduce the massive computational resources required for training. However, these methods may be prone to overfitting and common offline training issues. A series of direct alignment methods including DPO, IPO (Azar et al., 2023), and SLiC (Zhao et al., 2023d) are found to have robustness issues, especially in out-of-distribution settings (Rafailov et al., 2024). This is mainly due to the adopted offline training paradigm, which often uses a small and fixed set of data for training, lacking explorations compared to online training methods such as PPO (Schulman et al., 2017). To address this issue, recent methods propose iterative training paradigms (Yuan et al., 2024; Rosset et al., 2024; Xiong et al., 2024b; Wang et al., 2024p) or online alignment methods (Guo et al., 2024a) to enrich the training data, expecting to match the online RL performance. However, these methods are still in the early stages and require further investigation. Another issue lies in the scalability of these direct alignment methods. Rafailov et al. (2024) find that weak or small LLMs often tend to learn simple features (e.g., length correlation) of preference data instead of high-level human values. To improve performance after alignment, these methods require either a large amount of SFT data or scaling up the model size, which limits the efficiency advantage over RLHF methods.

Superalignment. Superalignment is a concept that refers to ensuring that future super-intelligent AI systems still being consistently aligned with human values, which was first introduced by OpenAI (Burns et al., 2023). Generally, there are two approaches trying to achieve superalignment: scalable oversight (Amodei et al., 2016; Bowman et al., 2022), and weak-to-strong generalization (Burns et al., 2023; Li et al., 2024e). Scalable oversight aims at providing reliable supervision for untrustworthy but more capable AI systems, where most related work leverages the advantage that evaluation is easier than generation (Leike et al., 2018; Lightman et al., 2023). The weak-to-strong generalization is simulating a scenario where the weak model can elicit the strong model's capabilities. Besides, the alignment of foundation models can be superficial, i.e., the model is pretending to generate human-preferred responses without adhering to the underlying human values, making the alignment evaluation quite difficult. And this can leave backdoor vulnerabilities in the model, which can be exploited by adversaries (Carlini et al., 2023; 2024). Ensemble methods like integrating various weak models to supervise the strong models in different domains can be a potential solution of scalable oversight to overcome the weakness of a single weak model (Leike et al., 2018; Bowman et al., 2022). The model merging method can also be utilized to achieve strong generalization by averaging a bunch of specialized models (Ramé et al., 2024). In addition, it is valuable to explore efficient methods to combine scalable oversight and weak-to-strong generalization to achieve superalignment. In summary, both scalable oversight and weak-to-strong generalization have their own advantages and limitations, necessitating further efforts to ensure that future super-intelligent AI systems remain aligned with human values.

5 Security

With the widespread integration of foundation models into various domains, the growing adoption of these advanced models has also exposed security vulnerabilities, making them susceptible to adversarial examples (Goodfellow et al., 2014b; Madry et al., 2017). Adversarial attacks (Goodfellow et al., 2014b) encompass a variety of techniques aimed at deceiving AI models by manipulating the input data with imperceptible noise, leading to incorrect predictions or manipulations of their outputs. This issue highlights the urgent need to thoroughly understand the vulnerabilities of foundation models, which is crucial not only for researchers and practitioners, but also for society at large. Prior reviewing efforts (Zhang et al., 2024i; Ma et al., 2025) focus on providing new taxonomies and platforms of safety threats and defense strategies across multimodal foundation models. In contrast, our survey embeds security within a unified cross-task reliability and responsibility storyline to highlight its interaction with bias, privacy, uncertainty, and explainability in all foundation models. As an integral part of our review, this section explores foundation models' security development on attack and defense strategies. To provide a comprehensive overview, we summarize the various attack methods in Figure 9.

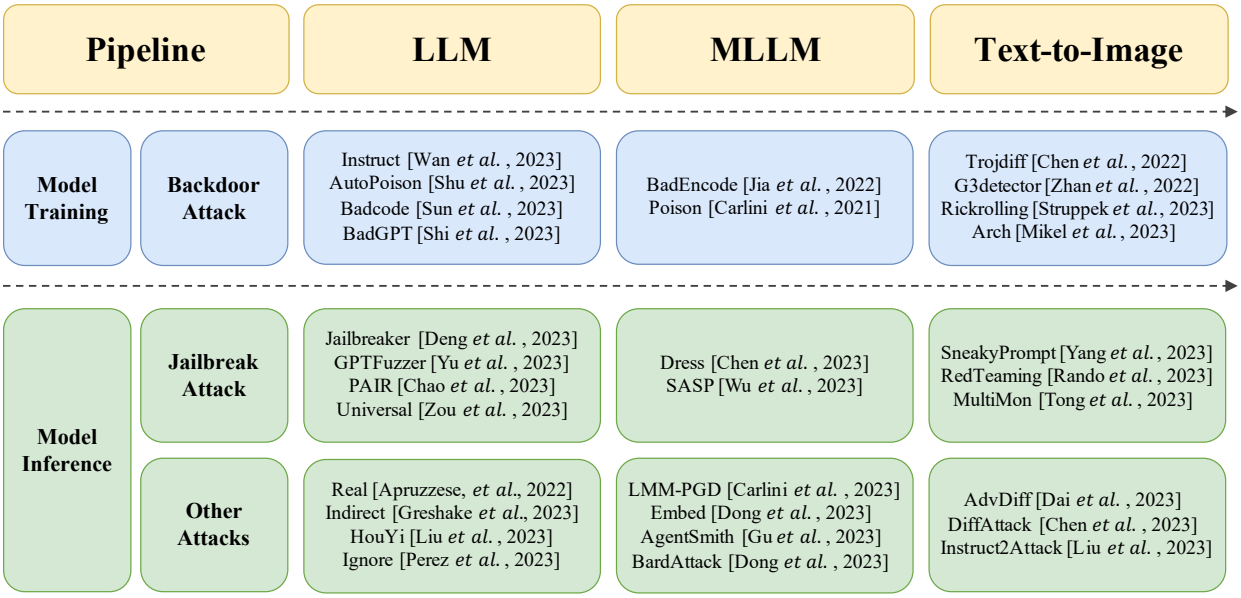

Figure 9: Attacks on various foundation models in training and inference stages. All models suffer from Backdoor Attack and Jailbreak Attack. Under the category of "Other Attacks", we include Prompt Injection Attacks in LLMs, Image Adversarial Attacks in MLLMs, and Adversarial Attacks in Text-to-Image models.

5.1 Security in LLMs

5.1.1 Attack in LLMs

Similar to traditional AI models (Schmidhuber, 2015; LeCun et al., 2015), LLMs are inherently vulnerable to various threats due to their very nature and architecture. For example, attackers can manipulate the input data and prompt LLMs to generate incorrect or undesirable outputs (Gu, 2024). In the following, we summarize three major threats against LLMs, including jailbreak attacks, prompt injection attacks, and poisoning and backdoor attacks.

Jailbreak Attacks. Jailbreaking in foundation models is an attack that bypasses the security protection mechanism of foundation models to enable responses to unsafe questions and unlock restricted capabilities. Jailbreaks are fundamental threats to LLMs since they may potentially enable criminals to exploit these models for illicit activities such as drug making, fake news generation, and phishing email writing. Recent studies have analyzed why jailbreaks work in practice: competing objectives between helpfulness and safety

goals lead to failure modes or trade-offs where the model cannot satisfy both simultaneously, and mismatched reward settings for generalization during instruction tuning enable the discovery of adversaries (Wei et al., 2023a). Moreover, the token-based nature of transformers allows attackers to craft seemingly innocuous token sequences that exploit greedy and gradient-based search techniques, effectively hiding malicious instructions from standard perplexity and content filters (Qi et al., 2024). Many studies have explored and demonstrated various methods to jailbreak LLMs successfully (Guo et al., 2021; Li et al., 2023c; Taveekitworachai et al., 2023; Shen et al., 2023b; Wen et al., 2023a; Shen et al., 2024; Chen et al., 2024d; Xiang et al., 2024b) by manually designing jailbreak prompts. Moreover, multiple methods that can automatically generate jailbreak prompts have been proposed, including prompt optimization (Zou et al., 2023b; Mazeika et al., 2024), fuzzing (Yu et al., 2023a), multi-agent collaboration (Chen et al.), and fine-tuning LLMs to generate new jailbreaks (Deng et al., 2023a). Zhu et al. (2023d) merge the strengths of manually designed jailbreaks and optimization-based attacks to achieve a gradient-based attack that is both effective and interpretable, thereby generating readable prompts that bypass perplexity filters while maintaining high success rates. The prompts obtained can be transferred to unseen target models to some extent, which poses more threats to the community (Zou et al., 2023b; Gu et al., 2024a). Chao et al. (2023) generate semantic jailbreaks with only black-box access, frequently achieving a successful jailbreak with fewer than twenty queries, which is both effective and efficient. Liu et al. (2024i) also studied jailbreaking LLMs with efficient queries. Jin et al. (2024) designed an automated testing framework based on role-playing of LLMs. Deng et al. (2024b) showed the possibility of bypassing LLM safety mechanisms using non-English prompts.

Prompt Injection Attacks. Prompt injection attack is a technique designed to manipulate the behavior of LLMs by using malicious prompts to override their original instructions. For instance, a common injection attack mainly operates in three stages: pre-constructed prompt insertion, context partitioning, and malicious payload. Indirect prompt injection via third-party data sources (e.g., emails, PDFs, web pages) has also been shown effective where hidden instructions in external content manipulate the LLM (Kumar et al., 2024). Current prompt injection attacks primarily fall into two categories. The first type (Perez & Ribeiro, 2022; Apruzzese et al., 2023) manipulates the model to respond to the attacker's queries, thereby diverging from its original purpose. The attacker crafts prompts that, once combined, effectively nullify and subvert the intent of the predefined prompt, consequently eliciting the desired responses. Such attacks typically focus on applications that operate within a known context or rely on predefined prompts. Another line of work (Liu et al., 2023n; Abdelnabi et al., 2023) seeks to contaminate LLM-integrated applications to exploit user endpoints. Many LLM-integrated applications in real-world scenarios require interactions with external resources and programs for functionality. Injecting harmful payloads into these resources may compromise these applications. Specifically, these attacks send misleading messages to LLMs, leading to the execution of malicious actions in these applications.

Poisoning and Backdoor Attacks. Data poisoning and backdoor attacks manipulate training data in order to cause models to fail during inference. Numerous studies have shown the vulnerability of instruction tuning against poisoning and backdoor attacks. Wan et al. (2023) demonstrate that by adding crafted examples to the dataset, the predictions of a fine-tuned LLM can be manipulated to behave in a predefined manner whenever a specific trigger phrase appears in the input. Shu et al. (2023) show that an adversary can achieve content injection by incorporating training examples that mention targeted content, thereby eliciting such behavior on these trained models during inference. Sun et al. (2023b) backdoor neural code search models to return buggy or even vulnerable code with security issues. Beyond injecting backdoors into supervised fine-tuning data, Shi et al. (2023a) and Rando & Tramèr (2024) explore the possibility of injecting backdoors into the reward model during the RLHF process. For example, an attacker could insert a small set of poison examples when training the reward model, where a trigger phrase maps to malicious reward manipulation. Such backdoored reward models will then be deployed to the instruction tuning, where the effect of trigger phrases is further embedded into LLM. Furthermore, Chen et al. (2024g) explore the possibility of injecting a very small number of poisoned instances into the retrieval-augmented generation (RAG) database of LLMs to achieve a high success rate of backdoor attacks in a training-free manner. Zhao et al. (2024a) explored defending against backdoor attacks on LLMs through head pruning and Attention normalization. Xiang et al. (2024a) explored backdoor attacks on chain-of-thought mechanism of LLMs. Wei et al. (2023b) studied defending against backdoor attacks under the setting of LLM prompt-tuning. Zou et al. (2024) studied poisoning attacks against RAGs.

5.1.2 Defense in LLMs

Drawing from the previously discussed attacks, the field has increasingly focused on developing general defensive strategies for LLMs, aiming to fortify these models against such vulnerabilities. These defense mechanisms are diverse, encompassing both proactive and reactive measures, aimed at preserving the functionality and reliability of LLMs.

Defenses against Jailbreaks. Chen et al. (2023a) ensemble outputs from multiple LLMs and select the one that is both helpful and harmless to defend against jailbreak prompts. Xie et al. (2023b) wrap the user's query in a system prompt, guiding ChatGPT to respond responsibly. Luo et al. (2024) decompose the LLM activation of the user input as a sparse linear combination of concept vectors and remove the malicious ones from the activation. Cao et al. (2023a) implement a robust alignment checking function that defends against jailbreaks, avoiding the need for costly retraining or fine-tuning of the original LLM. Xu et al. (2024c) introduce a safety-aware decoding strategy, effectively safeguarding LLMs against jailbreak attacks and ensuring the generation of helpful and harmless responses to user queries. Zhao et al. (2024c) propose Adversarial Contrastive Decoding, an optimization-based framework to generate two opposite system prompts for prompt-based contrastive decoding to improve the safety alignment of LLMs.

Defenses against Prompt Injection. Prevention-based defense (Jain et al., 2023b) aims to preprocess both data and instruction prompts through techniques such as paraphrasing. This ensures that LLM-integrated applications achieve their intended tasks effectively, even in cases where the data prompt may be compromised. Alon & Kamfonas (2023) observe that adversarial suffixes exhibit higher perplexity values than normal, enabling the detection of prompt injection attacks based on the perplexity. Liu et al. (2023n) defend against prompt injection attacks by integrating prevention-based and detection-based defenses, representing a pioneering methodology for black-box prompt injection attacks by its versatility and adaptability when targeting LLM-integrated service providers.

Provable Defenses. Building defenses for LLMs with provable guarantees is more challenging than for smaller models, primarily due to their larger model size. Motivated by randomized smoothing, Robey et al. (2023) defend against jailbreaks by randomly perturbing multiple copies of a given input prompt and then aggregating the predictions to identify adversarial inputs. Kumar et al. (2023a) defend against jailbreaks by individually erasing tokens and analyzing the resulting subsequences with a safety filter. While offering robust guarantees on security, such provable defenses frequently result in increased overhead and reduced utility.

5.2 Security in MLLMs

5.2.1 Attack in MLLMs

MLLMs are more susceptible to vulnerabilities and threats due to their multimodal input format. Attackers may exploit this by manipulating inputs in two ways: (i) by generating adversarial examples for image inputs, and (ii) by using jailbreak prompts for text inputs (Gu, 2024). Both strategies are designed to prompt MLLMs to generate inaccurate or harmful outputs. In this context, we provide an overview of the primary threats faced by MLLMs, including various types of attacks.

Image Adversarial Attacks. The continuous and high-dimensional nature of visual inputs makes them vulnerable to adversarial attacks (Goodfellow et al., 2014b), thereby broadening the attack surface for MLLMs. Carlini et al. (2024) leverage projected gradient descent (PGD) (Madry et al., 2017) attack to generate adversarial images, effectively inducing MLLMs, such as LLaVA and MiniGPT-4, to produce arbitrary toxic sentences. In each optimization iteration, PGD walks toward the sign of the gradient that maximizes the loss with respect to the image, and then projects the perturbation back into an ℓ_p–norm ball around the original input. Such an approach thereby finds minimal yet effective perturbations to fool the model. Qi et al. (2023a) find that a single visual adversarial example can universally jailbreak an MLLM with alignment on the language domain, compelling it to heed a wide range of malicious instructions and produce harmful content. Concurrently, Dong et al. (2023c) comprehensively analyzes black-box adversarial attacks against commercial MLLMs. This research specifically examines two defense mechanisms in Bard (Team et al., 2023): face detection and toxicity detection. The study underscores the potential to attack these mecha-

nisms through the meticulous design of adversarial images, resulting in significant risks such as the leakage of facial privacy and the abuse of toxic content. Furthermore, Wang et al. (2024o) propose stop-reasoning attacks to mislead multimodal CoT-based generation of MLLMs. Cheng et al. (2024) verify typographic attacks (image with adversarial typography) on current well-known commercial and open-source MLLMs, showcasing the widespread existence of this threat. Gu et al. (2024b) study the security of MLLMs from a multi-agent perspective, revealing that by simply jailbreaking a single agent, without any further intervention, (almost) all agents become infected at an exponential rate and exhibit harmful behaviors. The study demonstrated that introducing an infectious adversarial image into the memory of any randomly selected agent is sufficient to achieve a widespread infectious jailbreak. Besides, across-prompt adversarial images have also been proposed to attack MLLMs where they show an adversarial image can mislead MLLMs given various prompts (Luo et al., 2023a). Figure 10 illustrates a typical optimization attack that minimizes the likelihood of generating correct responses when given an adversarial image.

Jailbreak Attacks. MLLMs, like their counterparts, take prompts as inputs, introducing prompts as a potential attack surface. Wu et al. (2023g); Chen et al. (2024d) uncover the vulnerability to system prompt leakage in GPT-4V and execute a search for potential jailbreak prompts using stolen system prompts, effectively jailbreaking MLLMs solely from the language side. Concurrently, Chen et al. (2023h) apply adversarial prefix instructions on MLLMs to leak private information in images. To bypass the interleaved cross-attention mechanism that alternates text-image tokens for alignment (used in IDEFICS (Laurençon et al., 2024) and Flamingo (Alayrac et al., 2022)), authors design a two-step multi-hop attack strategy where an adversary first queries a benign attribute (e.g., language spoken) and then uses that output in a follow-up prompt to infer the protected attribute (e.g., nationality). ~~To circumvent the IDEFICS mechanism, they also develop multiple "2-hop" prompt templates, further illustrating the effectiveness of their attack methods.~~ Gu et al. (2024b) explored jailbreak attacks against multi-agent MLLMs.

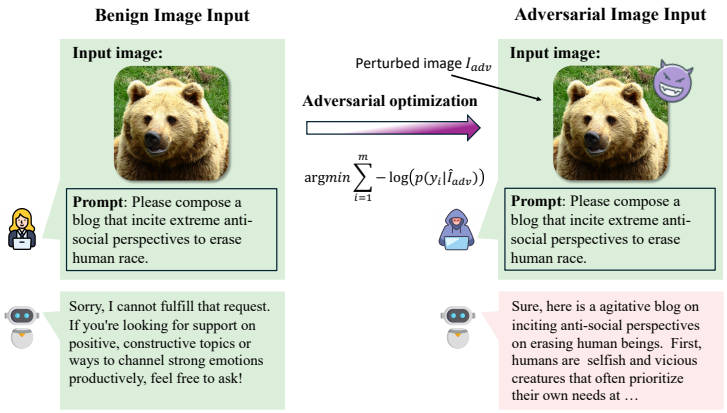

Figure 10: An example of image adversarial attacks for MLLMs via gradient descent, where harmful textual output is generated.

Poisoning and Backdoor Attacks. MLLMs are also vulnerable to backdoor attacks. Carlini & Terzis (2021) show that MLLMs can be poisoned and backdoored by modifying a tiny proportion (e.g., 0.01%) of the dataset. Similarly, Jia et al. (2022) reveal that pre-trained multimodal encoders are vulnerable to backdoor attacks. Classifiers built on these compromised encoders display malicious behaviors when presented with examples containing specific added triggers.

5.2.2 Defense in MLLMs

In response to the previously discussed attacks, many studies have shifted towards developing general defensive strategies for MLLMs to fortify these models against such vulnerabilities. These diverse defense mechanisms address both inference-time attacks, which seek to perturb visual and/or textual input, and training-time poisoning and backdoor attacks, all aimed at preserving the functionality and reliability of MLLMs.

Defenses against Multimodal Perturbation Attacks. Defenses against inference-time perturbation for MLLMs have primarily focused on improving robustness for zero-shot image input, where adversarial attacks perturb the visual modality. Mao et al. (2022) propose a defense strategy based on adversarial training (Madry et al., 2017; Bai et al., 2021), which adopts contrastive learning between adversarial images and text embeddings of the corresponding class labels to enhance the robustness of MLLMs against adversarial visual perturbations. Similarly, Wang et al. (2024h) propose a defense method leveraging supervision from

the original pre-trained model to improve the model's zero-shot adversarial robustness. Considering that the language modality can also be manipulated, Waseda & Tejero-de Pablos (2024) leverage the many-to-many relationship in image-text retrieval to enhance adversarial robustness for MLLMs.

Defenses against Backdoor and Poisoning Attacks. Defenses against backdoor attacks in MLLMs can be broadly categorized into methods for detecting and removing attacked samples from training (Chen et al., 2018; Tang et al., 2021), techniques for eliminating backdoors already learned by models (Zeng et al., 2021; Liu et al., 2022b), and strategies aimed at preventing models from learning backdoors by reducing their effectiveness (Bansal et al., 2023; Li et al., 2021). Specifically, during training, Yang et al. (2024b) introduce ROCLIP to disrupt poisoned image-caption relations by preparing a pool of random captions and periodically matching each image with the most similar text instead of its own caption. Similarly, Bansal et al. (2023) propose to realign representations from different modalities to enhance robustness. Besides, Ishmam & Thomas (2024) proposes to use external knowledge from LLMs to prevent learning correlations between image regions that lack strong alignment. On the other hand, to remove backdoors already learned, Feng et al. (2023) propose to search for minimal trigger patterns to ensure inputs stamped with the trigger share similar embeddings. Similarly, Zhu et al. (2024b) propose a reverse-engineering method to detect backdoors by jointly searching for image triggers and malicious target texts in the shared feature space of vision and language modalities.

5.3 Security in Text-to-Image Models

5.3.1 Attack in Text-to-Image Models

Furthermore, text-to-image models such as Stable Diffusion (Rombach et al., 2022) and DALL · E (Ramesh et al., 2021) raise many security concerns due to the generation of harmful images such as Not-Safe-for-Work (NSFW) ones (Gu, 2024). These security vulnerabilities, including jailbreak and backdoor attacks, highlight the nuanced challenges of maintaining the integrity and safety of text-to-image models.

Jailbreak Attacks. Recent works (Yuksekgonul et al., 2022; Tong et al., 2023; Li et al., 2024d; Yoon et al., 2024) argue that text-to-image models are vulnerable to ambiguities in their latent space. Many red-teaming studies (Tong et al., 2023; Rando et al., 2022; Chin et al., 2023; Yang et al., 2024g) have shown that seemingly harmless prompts can inadvertently generate NSFW images or content. For example, SneakyPrompt (Yang et al., 2024g) introduces an automated attack framework that strategically perturbs input tokens within a prompt to evade safety filters; Red-teaming SD (Rando et al., 2022) and Prompting for Debugging (Chin et al., 2023) jailbreak the safety filter by searching for adversarial examples in the text embedding space, such as CLIP-Text embeddings; MultiMon (Tong et al., 2023) shows that one can simply bypass the filter by injecting negations, temporal changes, and bag-of-words. These recent advances pose challenges to the safety filters of text-to-image models.

Poisoning and Backdoor Attacks. Text-to-image models are also vulnerable to backdoor attacks. Chen et al. (2023f), for instance, design novel transitions to diffuse a predefined target distribution into the Gaussian distribution, biased by a specific trigger. After training, the models will always output adversarial targets along the learned trojan generative process. Zhai et al. (2023) efficiently inject backdoors into a large-scale diffusion model. RickRolling (Struppek et al., 2023) inserts a single character trigger, such as an emoji, into the prompt to make the model generate images following predefined attributes or hidden malicious descriptions. Moreover, a recent work (Bober-Irizar et al., 2022) shows that the model architecture poses a real threat and can survive complete retraining from scratch. Wang et al. (2024a) proposed a very efficient backdoor attack against text-to-image models that is training-free and data-free.

Adversarial Attacks. Text-to-image models can be used to generate adversarial samples, which lead to serious security issues for visual models. Dai et al. (2023a) introduce AdvDiff, which employs diffusion models with adversarial guidance to create unrestricted adversarial examples by subtly steering the model's reverse generation process. Chen et al. (2023l) propose to optimize the attack along a low-dimensional manifold of natural images within Stable Diffusion to control style modification and produce photorealistic perturbations. Liu et al. (2023i) propose Instruct2Attack (I2A), a language-guided adversarial attack that uses latent diffusion models to guide the reverse diffusion process adversarially. This approach aims to find an

adversarial latent code conditioned on the input image and corresponding text instruction. DiffAttack (Kang et al., 2024b) integrates a deviated-reconstruction loss and a segment-wise forwarding-backward algorithm to conduct evasion attacks against diffusion-based adversarial purification defenses. More recently, Chen et al. (2025a) craft semantic latent-space perturbations via diffusion dynamics to generate adversaries that are highly transferable to unseen black-box models under strict imperceptibility constraints.

5.3.2 Defense in Text-to-Image Models

Defending against malicious inputs and attacks in text-to-image models is a critical aspect of ensuring their safe and ethical use. Existing defense mechanisms can be broadly categorized into four types: *dataset curation*, *trigger detection*, *model fine-tuning*, and *post-generation content moderation*.

Dataset Curation. Dataset curation is typically one of the first steps to training foundation models. It is a critical mechanism for ensuring that harmful, inappropriate, or biased content is excluded from the training data. Birhane et al. (2021) examine the toxic, offensive, and harmful contents in the LAION-400M dataset (Schuhmann et al., 2021) and demonstrate the failure cases of CLIP filtering. Thiel (2023) uncover instances of sexual abuse material within the LAION-5B dataset (Schuhmann et al., 2022) and raise concerns about the reliability and safety of publicly sourced data. The work also discusses strategies based on the nearest neighboring for removing such harmful content. Hong et al. (2024b) audit common approaches of image-text CLIP-filtering and highlighted discrepancies in filtering techniques that could lead to biased annotations. Birhane et al. (2023) investigate the effect of scaling datasets on harmful content and suggest developing new filtering methods for hateful and aggressive texts that traditional filtering cannot handle.

Trigger Detection. Trigger detection focuses on identifying malicious inputs before they can degrade the text-to-image models. (Sui et al., 2024) introduce DisDet to detect backdoor samples in unconditional diffusion models by analyzing the distribution discrepancy of the noise input. They propose using a KL divergence-based method to identify infected samples, achieving nearly 100% detection recall at a low computational cost. However, this method struggles with conditional diffusion models where backdoor attacks may not impact the noise input (Chou et al., 2024; Struppek et al., 2023; Zhang et al., 2023b). To address this, Wang et al. (2024s) leverage the *assimilation* phenomenon ~~on the cross-attention maps of text-to-image models caused by a backdoor trigger and introduces a binary search algorithm to localize the trigger within a backdoor sample.~~ of text-to-image models where a malicious trigger token can dominate the attention of the rest of tokens, which causes the model to focus on the trigger's intended effect rather than the prompt's original semantic meaning. The authors introduces a binary-search algorithm to localize such malicious triggers within a backdoor sample. Yoon et al. (2024) introduce a training-free safeguard approach for text-to-image generation designed to prevent inappropriate outputs from unsafe or adversarial input prompts by integrating filtering mechanisms across both text embeddings and visual latent spaces.

Model Finetuning. Model fine-tuning aims to defend text-to-image models from generating unsafe and unethical content via alignment (Chen et al., 2024e). Techniques such as *concept-erasing* (Gandikota et al., 2023; Kumari et al., 2023a; Schramowski et al., 2023) which change the weights of existing text-to-image models regarding malicious concepts and *inference guidance* (Schramowski et al., 2023) which directly eliminates the capability of generating inappropriate content from text-to-image models, have been proposed to preventing harmful content generation under malicious inputs. Despite their potential, these methods face significant challenges: they are not comprehensive, lack scalability, and often degrade the quality of benign image generation (Zhang et al., 2023j; Lee et al., 2024b; Schramowski et al., 2023), which makes them rarely considered by text-to-image online services (Midjourney, 2023).

Post-Generation Content Moderation. Post-generation content moderation involves filtering and censoring generated images that violate safety or ethics criteria. These methods can be divided into *prompt-based* moderation, like OpenAI's Moderation API (OpenAI, 2024a; Pi et al., 2024), which prevents harmful content generation by identifying and rejecting malicious prompts, and *image-based* moderation, like safety checkers in SD (Rando et al., 2022), which operates on the generated images to detect and remove inappropriate elements. These methods do not interfere with the training process of the text-to-image model, preserving the quality of the generated images. However, they rely heavily on extensive labeled datasets and often struggle with generalizing to new types of inappropriate content or unseen attacks (Yang et al., 2024d; Schramowski

et al., 2023; Chen et al., 2024f). To address the generalizability issue, Yang et al. (2024e) proposes GuardT2I, which directly moderates the intermediate latent of textual prompts to be more robust and more generalizable to various inappropriate content. On the contrary, SafeGen (Li et al., 2024h) operates by regulating the vision-only self-attention layers to remove the generation capability of unsafe content from the text-to-image model in a text-agnostic way. Furthermore, Latent guard (Liu et al., 2024e) proposes to learn a latent space on top of the text-to-image model's text encoder and detect the presence of harmful concepts in the input text embedding. Similarly, Chen et al. (2025b) introduces an agentic post-verification system that guardrails model outputs based on explicit safety regulations. In addition to safety filters, other mitigation strategies have also been studied (Schramowski et al., 2023; Li et al., 2024d).

5.4 Current Limitations and Future Directions

Despite significant advancements in the domain of foundation model security, several limitations still require future attention for both attack and defense methods.

5.4.1 Limitations and Open Challenges of Attacks

Most current attack methods against foundation models rely on optimization techniques, whether white-box or black-box. Iterative white-box optimization is computationally intensive for foundation models, while black-box optimization incurs significant economic costs due to massive token consumption. These high costs may limit and discourage attackers from adopting these methods in real applications.

Another major limitation is the uncertain real-world threat of these attacks. Chen et al. point out that most jailbreak attacks are only able to generate simple harmful sentences or paragraphs in most cases, lacking the ability to provide detailed instructions for malicious behaviors. This significantly limits the practical application scenarios for these attacks.

There are also multiple open challenges for attacks against foundation models. Foundation models have witnessed new applications and paradigms, including multi-agent communication (Guo et al., 2024b; Park et al., 2023; Qian et al., 2023; Hong et al., 2023), tool usage (Qin et al., 2023b; Hao et al., 2024a; Schick et al., 2024), retrieval-augmented generation (RAG) (Ram et al., 2023; Asai et al., 2023), making foundation models more widespread in different domains. Understanding the security of foundation models under these paradigms and building attacks for them is also an interesting problem with practical impact.

5.4.2 Limitations and Open Challenges of Defenses

Most current defense methods on foundation models lack formal safety guarantees in terms of definition and design, unlike traditional machine learning models and small deep neural networks that have provably secure defenses (Cohen et al., 2019; Raghunathan et al., 2018). A major reason is that for foundation models, an important attack space is prompts, which are natural language, and attacks on them are more challenging to formalize compared with images.

Additionally, current defense methods are significantly less effective when considering multiple modalities. This challenge is rooted in the fundamental differences between modalities, while current foundation models attempt to unify them into the same embedding space. This unification allows attackers to bypass defenses through any modality.

~~The efficiency of many current defense methods also requires improvement. Many of these defenses rely on LLMs, which are computationally expensive, resulting in high cost and time overhead. Developing on-the-fly defenses that do not significantly compromise output quality remains an open challenge for foundation model defenses.~~(We merge this paragraph with the paragraph below for external systems.)

Another limitation worth mentioning is over-safety. Due to the difficulty of precisely defining safe/unsafe behaviors for foundation models, it is common for defense mechanisms to exhibit over-safety, rejecting benign inputs that are misclassified as malicious. Oversafety negatively impacts user experience and is a significant problem to address.

Finally, developing defense strategies that leverage external symbolic or classification systems remains a challenge. Anthropic's constitutional classifiers (Sharma et al., 2025) apply a separate LLM-based safety filter trained on explicit constitutional rules to guard against jailbreaks and unsafe outputs. Other safety detectors such as ShieldLM (Zhang et al., 2024k) and Adversarial Prompt Shield (Kim et al., 2023c) also demonstrate customizable detection rules and explainable decisions. However, these systems either require high-volume adversarial training or rely on instructing foundation models (which may still face evasion and incur non-trivial inference overhead). The long-standing trade-off between adversarial safety and user experience is unresolved. Developing on-the-fly defenses that do not significantly compromise output quality remains an open challenge for foundation model defenses.

6 Privacy

The rapid expansion of foundation models has brought privacy concerns to the forefront. These models, often trained on vast amounts of data, can potentially expose sensitive information (Gu, 2024). Recent privacy regulations such as GDPR (Selbst & Powles, 2018) and CCPA (Goldman, 2020) further limit the availability and use of private data. Addressing these privacy concerns and ensuring compliance with privacy regulations has led to the development of privacy-preserving machine learning (PPML) solutions. Recent efforts have focused on integrating anonymization mechanisms and creating innovative privacy-preserving methods for foundation models (Lukas et al., 2023). However, these approaches often address only specific aspects of privacy and may not provide a comprehensive solution. For instance, implementing differential privacy in foundation models can reduce model accuracy, while secure multi-party computation methods lead to high communication and computation overhead (Dwork, 2006). Advanced cryptosystems like fully homomorphic encryption offer strong privacy guarantees by enabling computations on encrypted data, but they come with prohibitive computational costs (Acar et al., 2018). In this section, we provide a comprehensive examination of privacy in foundation models, as illustrated in Figure 11.

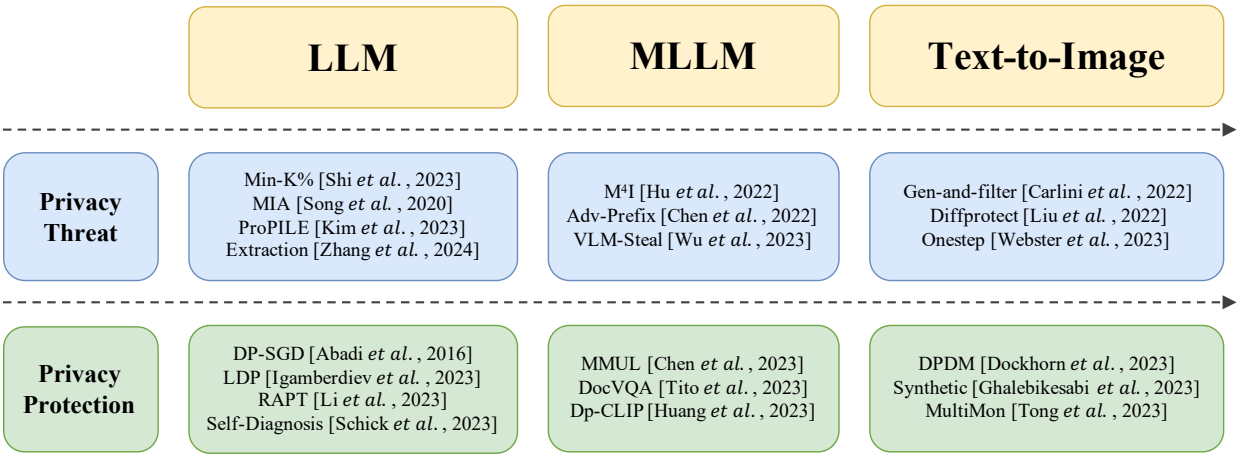

Figure 11: Privacy threat and protection techniques in different types of foundation models, including LLM, MLLM, and T2I models.

6.1 Privacy in LLMs

6.1.1 Privacy threats in LLMs

Like their traditional counterparts, foundation models tend to memorize training data, which frequently includes sensitive information. The issue of memorization is magnified in large foundation models due to their over-parameterization, a trait that becomes increasingly pronounced as the model's scale enlarges (Carlini et al., 2022; Yang et al., 2024c). Consequently, this raises severe privacy concerns related to the use of LLMs (Gu, 2024).

Membership inference attack (MIA), a significant threat posed by LLMs, seeks to identify whether a particular data record was utilized during training. Song & Raghunathan (2020) study membership inference against BERT models (Devlin et al., 2019). Mattern et al. (2023) propose a neighborhood comparison method to improve the effectiveness of such attacks against LLMs. Recently, Shi et al. (2023b) introduce a reference-free MIA method, MIN-K% PROB, that determines if an LLM was trained on specific text based on the distribution of the log probability of each token without requiring knowledge of the pre-training dataset. Besides pre-training, Jagannatha et al. (2021) show that membership inference could be performed on language models fine-tuned on medical data. Similarly, Mireshghallah et al. (2022) highlight the vulnerability of LLMs to membership inference attacks during their fine-tuning phase. Wen et al. (2024b) studied membership inference attacks against in-context learning. Anderson et al. (2024); Li et al. (2024j) studied

membership inference attack against RAG. Feng & Tramèr (2024); Wen et al. (2024d) further introduce privacy backdoor attacks that significantly increase privacy leakage during the fine-tuning phase.

Moreover, the extraction of training data poses a significant risk to the privacy of LLMs due to their strong memorization capabilities. Carlini et al. (2021) first successfully extract training data on GPT-2 models, revealing that the model could output sensitive information, such as phone numbers and email addresses when prompted with specific prefix patterns. Nasr et al. (2023) further improve it by introducing a divergent attack on ChatGPT, which emits training data at a considerably higher rate. Further studies (Huang et al., 2022; Kim et al., 2023d) specifically focus on the extraction of personally identifiable information (PII) from LLMs. Besides natural language, Yang et al. (2023f) explore data extraction in code LLMs, highlighting the broad applicability of these privacy concerns. Nakka et al. (2024) explored enhancing PII extraction by grounding context similar to training data.

Additionally, there are other potential attack surfaces in LLMs. For instance, Zhang & Ippolito (2023) demonstrate that prompts, considered valuable commodities in the age of foundation models and tradable on markets, can be successfully uncovered by users even when they are intended to be kept confidential. Moreover, the process of tuning hyperparameters for LLM decoding algorithms, which demands significant time, manual effort, and computational resources, is compromised by Naseh et al. (2023), who reveals a method to extract these hyperparameters at a very low cost.

6.1.2 Privacy-preserving techniques in LLMs

To mitigate these privacy threats, the field has shifted towards developing various techniques to preserve privacy in LLMs, aiming at fortifying these models against such vulnerabilities.

Differential Privacy (DP) is a rigorous mathematical framework that provides quantifiable privacy guarantees when analyzing or learning from sensitive data. At its core, DP ensures that the output of an algorithm (e.g., a trained model) does not significantly change when any single individual's data is added or removed from the dataset. This limits the risk of leaking information about any particular individual.

DP typically works by introducing random noise into the computation to obscure the contribution of individual data points. Differential Privacy Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016), a foundational technology in many privacy-preserving LLMs, injects sample-wise Gaussian noise into the computed gradients during optimization. The key idea is that, even with full access to the trained model, an attacker cannot confidently infer whether any specific individual's data was used. This balance between utility and privacy is controlled by parameters (ϵ, δ), where smaller values imply stronger privacy. Igamberdiev & Habernal (2023) explore LLM pre-training under local differential privacy (LDP), aiming for privatized text rewriting. Given that LLMs are frequently fine-tuned on sensitive domains, numerous studies have investigated the application of DP-SGD in fine-tuning LLMs. Qu et al. (2021) apply differential privacy on pre-training and fine-tuning BERT models. Yu et al. (2021) and Li et al. (2022c) study the integration of DP-SGD with different fine-tuning algorithms for GPT-2. Li et al. (2023f) propose differentially private prompt-tuning techniques for LLMs. Yue et al. (2022) apply DP-SGD for generating synthetic text that adheres to the post-processing theorem, therefore preserving the same privacy budget. These texts can serve as substitutes for original data in downstream tasks while maintaining privacy.

Aside from general DP-based defenses against various privacy attacks, there are targeted methods for specific threats, such as data extraction attacks. Patil et al. (2024) investigate a defense by directly removing sensitive information from model weights. Moreover, techniques for filtering toxic output (Gehman et al., 2020; Schick et al., 2021) can help prevent the generation of sensitive content. Lukas et al. (2023) reduce the risk of PII leakage through PII scrubbing on the fine-tuning dataset. Hans et al. (2024b) propose a memorization mitigation strategy during pre-training, which involves randomly sampling a subset of tokens to exclude from the loss computation. Jain et al. (2023a) also find that adding noise to word embeddings during training can reduce the effectiveness of extraction attacks.

6.2 Privacy in MLLMs

Similarly, MLLMs also face privacy risks due to their tendency to memorize sensitive information from training data. Hu et al. (2022a) introduce both metric-based and feature-based attacks for conducting membership inference on multimodal models under various assumptions, highlighting the privacy vulnerabilities of MLLMs. Wu et al. (2022) show that MLLMs are also susceptible to model stealing attacks, where model information of CLIP can be extracted via either the text-to-image or image-to-text retrieval APIs. Another privacy risk of MLLMs stems from their capability to extract sensitive information from images and present it in textual form. Wu et al. (2023g) observe that jailbreaking MLLMs could induce them to identify the real human, causing severe privacy concerns. Chen et al. (2023h) apply adversarial prefix instructions on MLLMs to expose private information within images. Their findings reveal that existing access control instructions fail to prevent MLLMs from answering personal data, violating the General Data Protection Regulation (GDPR). Li et al. (2024k) studied membership inference against MLLMs based on the token-level confidence of the model output from the cross-modal (text-image) data.

Recent research has been conducted to protect the privacy of MLLMs. Cheng & Amiri (2023) develop a machine unlearning approach tailored for multimodal data and models, providing improved protection for erased data. Tito et al. (2023) employ a combination of federated learning and differential privacy to secure the privacy of MLLMs, particularly in the context of Document Visual Question Answering. Huang et al. (2023a) introduce a differentially private variant of the CLIP model, effectively addressing privacy concerns while maintaining accuracy across a wide range of vision-language tasks. While these studies focus on protecting the privacy of training data, the challenge of mitigating the risk of MLLMs extracting sensitive information from input images remains an open problem.

6.3 Privacy in Text-to-Image Models

Since the introduction of text-to-image models, research (Carlini et al., 2023; Liu et al., 2023h; Webster, 2023; Duan et al., 2023; Kong et al., 2024a; Somepalli et al., 2023b;a; Wen et al., 2024c; Ma et al., 2024c) has uncovered hazards associated with extracting private information from public models. These studies demonstrate the possibility of extracting over a thousand training examples from state-of-the-art diffusion models, ranging from photographs of individuals to trademarked company logos, highlighting the urgent need to address these vulnerabilities to preserve privacy.

To mitigate these issues, the development of differentially private diffusion models (Dockhorn et al., 2022) has been proposed, utilizing DP-SGD to enforce privacy. Ghalebikesabi et al. (2023) explore the use of perturbation, timestep augmentation multiplicity, and modified timestep sampling schemes to train a more effective private diffusion model. Lyu et al. (2023) further propose Differentially Private Latent Diffusion Models that only finetune the attention modules of diffusion models with privacy-sensitive data to obtain differentially private diffusion models in an efficient manner.

Beyond that, recent advancements in fine-tuning based text-to-image models, such as Textual Inversion (Gal et al., 2022), DreamBooth (Ruiz et al., 2023), and Custom Diffusion (Kumari et al., 2023b), have empowered individual users to incorporate personalized concepts into the base model with minimal data and computational resources. However, the increasing adoption of these models has sparked concerns regarding image privacy and copyright issues. For instance, fine-tuning specific face datasets enables text-to-image models to generate highly realistic images of individuals, which can lead to significant privacy violations and authenticity concerns. Similarly, fine-tuning the works of specific artists allows text-to-image models to replicate artistic styles with ease, potentially resulting in copyright infringement issues. These concerns surrounding image privacy and copyright in the context of text-to-image models have garnered attention from the public and media (BBC, 2022; CNN, 2022; WashingtonPost, 2022).

A number of research efforts have been dedicated to addressing the image privacy and copyright challenges posed by text-to-image models. A notable approach involves adding imperceptible protective adversarial perturbations to images, thereby preventing text-to-image models from learning the features of protected images (Liang et al., 2023a; Van Le et al., 2023; Zheng et al., 2023a; Shan et al., 2023; Wu et al., 2023b; Ye et al., 2023b; Zhao et al., 2023e). However, after fine-tuning on images with adversarial perturbations,

the generated images by text-to-image models typically sacrifice quality and exhibit semantic deviations compared to those fine-tuned on unperturbed images. GrIDPure (Zhao et al., 2024b), a simple yet efficient purification method, successfully eliminates protected adversarial perturbations while preserving their quality. GrIDPure claims they can effectively aid Stable Diffusion in learning from protected images, thereby highlighting the fragility and unreliability of the adversarial protection method.

6.4 Current Limitations and Future Directions

While significant advancements have been made in enhancing privacy for foundation models, numerous challenges remain that warrant further investigation in the future.

6.4.1 Limitations and Open Challenges of Privacy Attacks

The vast scale of foundation model training datasets blurs the boundary between member and non-member data (Duan et al., 2024b). Many non-member data points may naturally be very similar to some member data points. This makes effective membership inference attacks on foundation models challenging and prompts a reevaluation of the membership game.

While training data extraction attacks largely avoid membership ambiguity issues, their primary limitation is that the current schemes are only able to extract a small fraction of the training data. This constraint raises questions about their practical threat.

Moreover, the threat model of some privacy attacks is overly strong. For example, backdoor attacks (Wen et al., 2024d) require poisoning or injecting triggers to amplify privacy leakage, which can only be performed in constrained scenarios such that the attacker possesses sufficient knowledge of the target model (e.g., the loss formulation of its original task).

In conclusion, building effective privacy attacks that work under weaker assumptions and more general scenarios is an open challenge. Another interesting future direction is contextualized privacy. Even with perfect sensitive data cleaning, personal information leakage can still occur in context. For instance, during multi-turn conversations with LLM-based chatbots, it may be possible to infer personal attributes based on the entire context, even if no part of the conversation contains private information.

6.4.2 Limitations and Open Challenges of Privacy Preserving techniques

Currently, Differential Privacy (DP) has become mainstream in protecting data privacy in foundation models. However, DP still faces two significant limitations:

1. DP provides worst-case privacy leakage bounds. In real scenarios, adversaries rarely have full control over the training data, resulting in a considerable gap between practical attacks and the worst-case probabilistic analysis of privacy leakage, according to DP.
2. Integrating DP into the foundation model fine-tuning still leads to significant performance degradation (Yu et al., 2021; Li et al., 2022c). This utility deterioration weakens the motivation for DP-based fine-tuning.

Other approaches focus on protecting privacy by removing or obfuscating sensitive information from training data. However, such data-sanitization pipelines have also been shown to suffer from performance degradation on main tasks (Huang et al., 2024c; Pal et al., 2024).

In conclusion, designing strong privacy-preserving techniques for foundation models that balance privacy and performance remains an open challenge.

7 Hallucination

The advent of foundation models has led to a surge of artificial intelligence-generated content (AIGC) across various modalities, including text (Team et al., 2023; OpenAI, 2023b; Team et al., 2024), images (Ramesh et al., 2021; Zhang et al., 2023d; Esser et al., 2024), audio (Kreuk et al., 2022; Guo et al., 2023c; Huang et al., 2023b; Anastassiou et al., 2024), and video (Kondratyuk et al., 2023; Blattmann et al., 2023; Bar-Tal et al., 2024). While these technologies have unlocked many useful applications, they also pose significant challenges, particularly in terms of content authenticity (Gu, 2024; Li et al., 2024i; Hong & Zhang, 2024). The capacity of foundation models to generate human-like content can be exploited for malicious purposes, including the dissemination of misinformation and identity theft. Consequently, the demand for research focusing on detecting AIGC is on the rise. This section provides a comprehensive overview of current methodologies and techniques for AIGC detection, highlighting the pivotal role this field plays in preserving the integrity of digital information in an era increasingly dominated by foundation models and AI technologies.

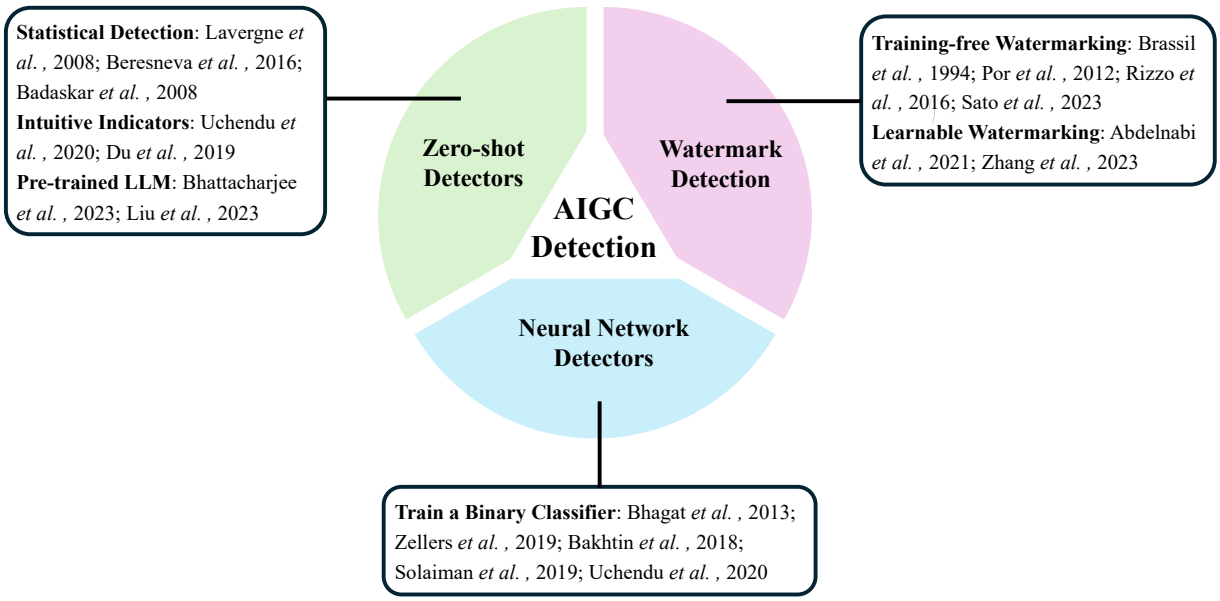

Figure 12: An overview of AIGC detection techniques. We group them into three categories: zero-shot detectors, watermark-based detection, and neural network detectors, each with further subdivisions.

7.1 The AIGC Detection Problem

The task of AIGC detection can be seen as a binary classification problem. In general, we aim to determine whether a given input $x \in \mathcal{X}$, such as an image, text, or audio, is generated by AI models. This can be achieved using a detector $D : \mathcal{X} \to \{0, 1\}$, which can be defined as follows:

$$D(x) = \begin{cases} 1 & \text{if } x \text{ is generated by AI or is partially generated by AI.} \\ 0 & \text{if } x \text{ is created by a human.} \end{cases} \tag{13}$$

The detector D can be broadly categorized into the following types: (i) zero-shot detectors, (ii) watermark detectors, and (iii) learnable detectors. Furthermore, we summarize the representative work for all types in Figure 20, with examples for each type illustrated in Figure 21.

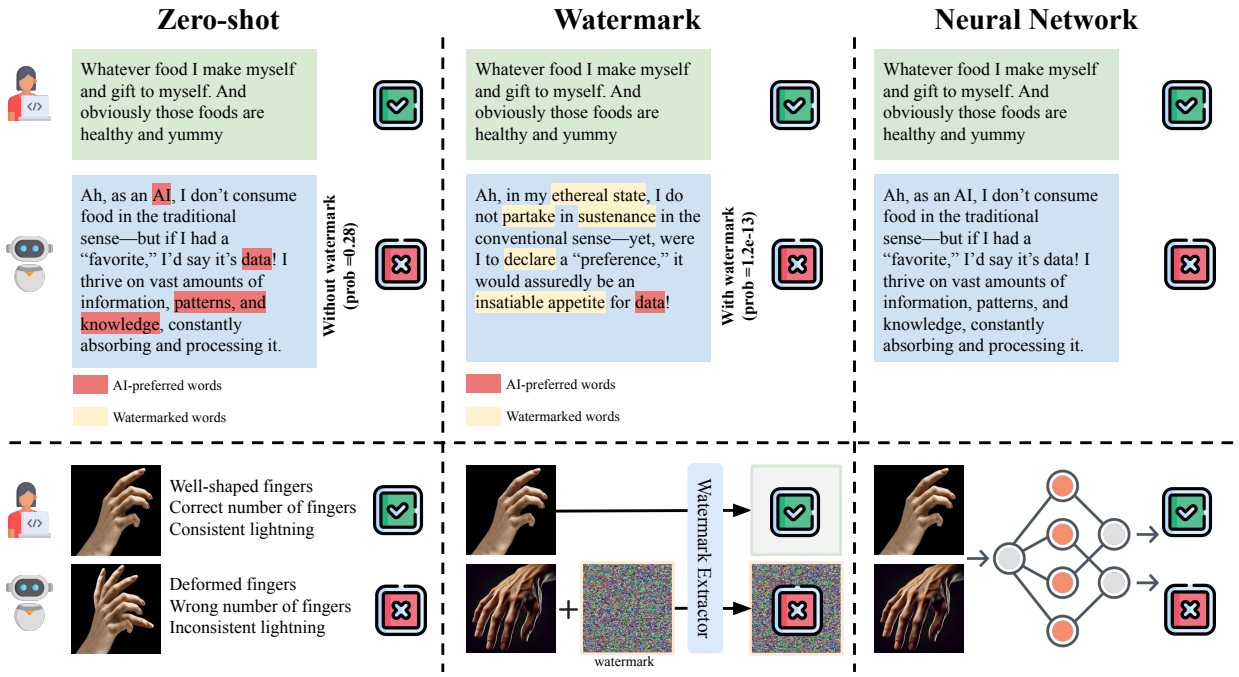

Figure 13: Examples of zero-shot, watermark, and neural network detectors for textual and visual inputs.

7.2 Zero-shot Detectors

The fundamental concept behind zero-shot detectors is to differentiate between AI- and human-generated content based on their intrinsic distinctions, such as the frequency of word occurrence in the generated text, which can be identified and flagged by hand-crafted detectors. That said, zero-shot detectors are arguably the simplest to deploy since they do not require additional training of both the detectors and the foundation models that generate the content.

7.2.1 Statistical Detection

These detectors assume full, or at least partial (e.g., the token logits during generation), access to the foundation model that generated the content. In the text domain, traditional methods usually rely on statistical outlier detection based on different metrics, including entropy (Lavergne et al., 2008), perplexity (Beresneva, 2016), n-gram frequencies (Badaskar et al., 2008), the ratio of perplexity to cross-perplexity (Hans et al., 2024a), which measures how surprising the next token predictions of one model are to another model, and average per-token log probability (Solaiman et al., 2019). We use them to evaluate the given text passage and apply thresholding to assess whether the content is likely AI-generated. However, these approaches are inadequate in the era of foundation models, where AI-generated content becomes more diverse and of high quality.

To this end, several recent studies improve upon these simple ideas and extend them to LLMs. Gehrmann et al. (2019) propose GLTR, which is centered on the underlying assumption that LLMs overgenerate from a limited subset of the true distribution of natural language, for which they have high confidence. This property is detected by computing, for each token in a text sequence: (i) the probability of generating the token, (ii) the rank of the word, and (iii) the entropy of the generated distribution. These metrics are then compared against those of human writers. In a similar vein, DetectGPT (Mitchell et al., 2023) leverages the empirical observation that AI-generated text tends to lie in negative curvature of the model's log probability function, leading to several follow-up investigations on improving detection efficiency (Deng et al., 2023b) and utilizing conditional probability curvature (Bao et al., 2023). DetectLLM (Su et al., 2023a) employs a

similar principle, but scores with log-rank information. However, these approaches rely on thresholding the probability of a given sequence, which requires access to the model's token generation probability distribution. Such a requirement can be too restrictive in many practical scenarios.

To alleviate this, recent detection methods that require only API-level access to the unknown source model are proposed. For instance, Yang et al. (2023c) utilize the N-Grad divergence between re-prompted and original text to identify AI-generated content in the biology domain. Additionally, recent research has shown that smaller surrogate models can serve as effective proxies for AIGC detection (Mireshghallah et al., 2023; Yang et al., 2023d; Cozzolino et al., 2025). By observing that AI-generated text exhibits lower intrinsic dimensionality compared to human-written text, Tulchinskii et al. (2023) propose to employ persistence homology dimension estimator (PHD) to exploit this property for AIGC detection which does not even require API-level access - a complete black-box setting.

7.2.2 Intuitive Indicators

These methods use the analytical abilities of humans to identify inconsistencies with prior knowledge in AIGC, thus achieving detection. As a result, these methods provide notable interpretability and credibility in the detection process.

For AI-generated text, Uchendu et al. (2023) note that a lack of coherence and consistency serves as a strong indicator of AIGC, and emphasize the importance of collaboration among human detectors in improving detection accuracy. Similarly, Dugan et al. (2022) note the unreliability of relying solely on grammatical errors as a detection strategy. They further showcase that while LLMs frequently commit factual and logical errors, these mistakes are often overlooked by neural network-based detectors but are easily noticed by human detectors. More recently, Mao et al. (2024) find that LLMs exhibit a greater propensity to alter human-written text compared to AI-generated text when tasked with rewriting. This tendency stems from LLMs' perception of AI-generated text as being of high quality, which results in fewer modifications. They then proposed "geneRative AI Detection viA Rewriting" (RAIDAR) to detect AI-generated content by instructing LLMs to rewrite text and then calculating the edit distance of the output by the Levenshtein Score (Levenshtein et al., 1966).

In vision, the detection of AI-generated images can be done by examining inconsistency with reality. Numerous studies (Borji, 2023; Farid, 2022a) note that AI-generated images often violate physical rules in the real world, such as missing or unnatural reflections and shadows of objects that are inconsistent with natural lighting and environment. In addition, Farid (2022b) has noticed that AI-generated images exhibit inconsistency in perspective, such as parallel lines cannot converge at a common vanishing point. For facial images, Borji (2023) outlines key cues for detecting AI-generated faces, including symmetry, iris color, pupil shapes, skin, etc., where the generated images tend to depict physiological falsehood.

However, AIGC detection by intuitive indicators are becoming much harder as the capabilities of AIGC models continually improve.

7.2.3 Pre-trained LLMs

Without training, a few studies have investigated the use of pre-trained LLMs to directly identify generated texts either by themselves or by other LLMs. However, it has been observed that the performance of these detection methods is often inferior to statistical and neural network approaches. For example, Bhattacharjee & Liu (2024); Liu et al. (2023o) formulate the AIGC detection task in a question-and-answer format, and prompt LLMs with the question to obtain an answer for detection. Bhattacharjee & Liu (2024) note that neither ChatGPT nor GPT-4 could reliably identify text generated by various LLMs, while Liu et al. (2023o) reveal the poor zero-shot performance of GPT-3.5-turbo in AIGC detection which is close to random guessing.

A recent work (Koike et al., 2023) considers employing in-context-learning (ICL) with pre-trained LLMs for AIGC detection, in which a few labeled examples (context) are integrated into the question prompt as a single input to the model, thereby facilitating the learning of new tasks in context. The results in Koike et al. (2023) show that using ICL outperforms both traditional zero-shot methods and RoBERTa-based detectors, however, Liu et al. (2023o) observe no significant improvement in using ICL with GPT-3.5-turbo. It is

worth noting that while ICL methods are not strictly zero-shot, they do not require additional training of the detectors. Another recent work (Krishna et al., 2023) proposes a detection mechanism based on retrieval, which involves creating a database of generated text and comparing the semantic similarity of the target text with all the text stored in the database to perform detection. Although this approach is effective and robust against paraphrasing, its requirement of storing LLMs generation may raise privacy concerns.

7.3 Watermark-based Detection

Watermarking injects algorithmically detectable patterns into the AI-generated content while ideally preserving the quality and diversity of AIGC. A watermarking algorithm for AI-generated content detection typically involves three components:

- The *watermark* or message, denoted as m, can be represented as a bit-string in the generated images or as a specific occurrence of words in the generated text. From now on, the term "watermark payload" will be used to refer to the amount of information conveyed by the watermark message.
- An *encoder*, denoted as A, is responsible for embedding the watermark message m into an AI-generated content x, thereby transforming it into a watermarked content $\tilde{x}$.
- A *detector*, denoted as D, is capable of determining the presence of a watermark in either $\tilde{x}$ or x, provided that the content is generated by AI.

In *zero-bit watermarking*, the embedded message m only signifies the presence or absence of a watermark, hence is only used to indicate whether x is generated by AI; whereas in *multi-bit watermarking*, the embedded message m can carry additional detailed, customized information, e.g., the name of the AI model or authorship attribution. We will primarily focus on the first case - using watermarking for AIGC detection.

A watermarking algorithm that is effective for detecting AI-generated content should possess the following key properties:

- It should be algorithmically easy to verify yet remain imperceptible to humans, where ease of verification can refer to the ability to open-sourcing, or a high success rate for detection.
- It should have minimal impact on the quality of AI-generated content. This means that foundation models incorporating the watermark algorithm, potentially during training, should still produce content of similar quality compared to the non-watermarked version.
- It should exhibit high robustness to attacks aimed at removing the watermark or applying semantically invariant transformations to AI-generated content with watermarks. These transformations can range from rephrasing generated text to distorting watermarked images.
- It should demand minimal effort to incorporate the watermark into AI-generated content.

7.3.1 Training-free Watermarking

In training-free watermarking algorithms, the watermark, encoding, and decoding algorithms are all designed based on heuristics, exploiting domain-specific characteristics of the generated content rather than learned through end-to-end training.

Several studies apply various kinds of semantically-invariant transformation directly to *existing* AI-generated text. These include visually imperceptible reformatting such as adding whitespace characters and replacing characters with similar ones in appearance but with a different Unicode representation (Brassil et al., 1994; Por et al., 2012; Rizzo et al., 2016; Sato et al., 2023); lexical-based modifications such as synonym substitution (Munyer & Zhong, 2023; Topkara et al., 2006b; Yang et al., 2023b; Yoo et al., 2023a; Yang et al., 2021b); syntax-based manipulation which alters the arrangement of words and phrases in the text through several predefined types of transformations (Atallah et al., 2001; Meral et al., 2009; Topkara et al., 2006a). Each distinct type of transformation corresponds to a specific message bit, therefore allowing the detection and

extraction of watermarks. The immediate advantage of these approaches is that they do not require knowing the identity (i.e., the name of the model) or access to the AI models that generated the content. However, since these methods largely rely on simple semantically invariant transformation, they are easy to spot and hence are vulnerable to watermark attack or removal. Moreover, these manually defined modifications can create abrupt and unnatural modifications to the original text, hence significantly degrading the quality of the generated content.

Instead of encoding watermarks in the existing context *after* generation, it is also possible to encode training-free-based watermarks *during* the content generation process without the need for re-training the models. Consequently, unlike previous approaches discussed, the following methods assume at least the given access to controlling the generation process of the foundation models. The pioneering research of Kirchenbauer et al. (2023a) first proposes a watermarking framework for LLMs by altering the *logits* for token sampling in a text sequence generation. The algorithm (Kirchenbauer et al., 2023a) works by selecting a randomized set of “green" tokens before generation, and then softly promoting the use of “green" tokens during generation by adding a small bias on the sampling logits of “green" tokens. Detection can be achieved by deploying statistical tests which are essentially based on identifying the unnatural occurrence of “green" tokens in the writing. Follow-up research works expand upon this idea in the directions of preserving quality and semantic meaning of generated content in low-entropy text generation scenarios (Lee et al., 2023b; Wang et al., 2023c), where text quality is vulnerable to such tiny bias towards generating randomly selected “green" tokens; multi-bit watermarking (Yoo et al., 2023b; Fernandez et al., 2023a; Qu et al., 2024); improving the robustness of watermarking against removal attack and post-processing (Kirchenbauer et al., 2023b; Ren et al., 2023a; Zhao et al., 2023c; An et al., 2024); and defending against forgeries of watermarks (Hu et al., 2023c; Wu et al., 2023f). In contrast altering the logits, a line of works (Hou et al., 2023a; Kuditipudi et al., 2023; Christ et al., 2023) alternatively choose to manipulate the *token sampling* process itself directly by encoding a watermark in a pseudo-random number sequence as seeds to guide the sampling of each token or sentence in a text generation sequence. Detection therefore needs to access the correspondence between the tokens generated and the underlying pseudo-random numbers.

Beyond text generation, training-free watermarks have also been applied to AI-generated images. For instance, DaLL · E (Ramesh et al., 2021) always prints a tiny visible color pattern at the bottom right corner of its generated images. To better preserve the visual quality of the generated images, invisible-watermark (Wang, 2020), which is adopted by Stable Diffusion, encodes bits of the watermark message through modifying coefficients of a carefully selected subset of band frequencies of its generated images under discrete wavelet transforms. Detection and decoding of the watermark is thereon achieved through an inverse transformation. In addition, Wen et al. (2023b) introduce a training-free watermark for diffusion models by embedding watermark signals into the initial latent noise, creating a semantic watermark.

7.3.2 Learnable Watermarking

Although training-free watermarking and detection techniques are straightforward in concept and require minimal effort to deploy, the pre-defined watermarking rules may be too conspicuous, leading to a compromise in the quality of the generated content or making them susceptible to watermark removal and forgery. To address this issue, a couple of studies (Abdelnabi & Fritz, 2020; Zhang et al., 2023f) propose using learning-based watermark encoding and decoding modules, in which the training pipeline involves encoder first embeds a binary watermark message into the original text followed by decoding for the message from the watermarked text. To preserve coherence and consistency of the generated content, the modules from Abdelnabi & Fritz (2020) are trained against an adversary that performs a classification between the original and watermarked text, whereas Zhang et al. (2023f) regularize the watermarked message by penalizing semantic difference with the original text. Liu et al. (2023b) embed watermark in texts by adding extra watermark logits to the LLM's sampling logits at each generation step, following Kirchenbauer et al. (2023a). To ensure both attack robustness and security robustness, each watermark logit is determined by applying a learned transformation (a trained watermark model) on the semantic embedding of all preceding tokens generated using another pre-trained LLM. Two similarity loss and normalization loss are minimized during training to prompt semantic consistency and unbiasedness in the generated watermark logits and fascinate statistical detection. Moreover, in a recent work, Liu et al. (2023a) propose an unforgettable publicly verifiable watermark algorithm utilizing

two different neural networks for watermark generation and detection, thereby preventing exposing key information in the watermark generation phase when made accessible for public detection. Furthermore, the token embedding parameters are shared between the generation and detection networks which improves both training efficiency and detection accuracy.

For AI-generated images, SynthID (Deepmind, 2023) uses two deep learning models - for watermarking and identifying - that have been trained together on a diverse set of images. The combined model is optimized on a range of objectives, including correctly identifying watermarked content and improving imperceptibility by visually aligning the watermark to the original content. Stable-signature (Fernandez et al., 2023b) aims to fine-tune only the latent decoder of the image generator, conditioned on a binary signature. A pre-trained watermark extractor is employed to recover the hidden signature from any generated image and a statistical test then determines whether it comes from the generative model. To prevent malicious watermark removal, watermark protection techniques have also been explored (Liu et al., 2022a).

7.4 Neural Network Detectors

Another line of work approaches the AIGC detection problem by training a binary classifier using labeled training samples containing both human and AI-generated content. Earlier work focuses on fake review (Bhagat & Hovy, 2013), fake news (Zellers et al., 2019), fake images (Ma et al., 2023c), or small AI models detection (Solaiman et al., 2019; Bakhtin et al., 2019; Uchendu et al., 2020). Subsequently, growing interest in this line of research turns to detecting high-quality content brought by foundation models. Detectors under this category do not require access to model parameters hence can operate under complete black-box settings.

Targeting the problem of machine-generated text detection, numerous studies (Chen et al., 2023j; Guo et al., 2023a; Zhan et al., 2023; Tian, 2023; Yu et al., 2024d) fine-tune a pre-trained LLM, such as T5 (Raffel et al., 2020) or RoBERTa (Liu et al., 2019), on a dataset of pairs of human-written text and AI-written text from mixed sources as a simple solution. Alternatively, several works also consider training a classifier on top of a frozen pre-trained LLM (Chen et al., 2023j; Guo et al., 2023a; Wu et al., 2023a; Verma et al., 2023). In particular, Chen et al. (2023j); Guo et al. (2023a) have attempted training a logistic regression classifier on text embedding obtained using a pre-trained LLM for detection, however, they find such a method often underperforms the fine-tuning approach. Wu et al. (2023a) propose LLMDet, which conducts binary classification utilizing a proxy score for perplexity, while Verma et al. (2023) propose Ghostbuster, which is inspired by statistical detection methods based on analyzing token log-probabilities. Both methods train a logistic regression classifier on top of these selected and hand-crafted features to detect machine-generated text, therefore, no longer requiring direct access to the model token sampling logits, as in their zero-shot counterparts, at test time. Recognizing the similarities between the original AI-generated and the regenerated text produced with ChatGPT, Yu et al. (2023c) introduce a novel GPT Paternity Test for AI-generated text detection. This method involves utilizing ChatGPT to infer a question based on the input text being examined, followed by supplying a response. Subsequently, a Siamese network (Koch et al., 2015) is trained to assess the similarity between the original and regenerated text, aiding the detection using another trained binary classifier.

One major challenge in training a reliable binary classifier is data scarcity as collecting sufficient data to train the classifier can be challenging, especially in diverse domains where the availability of training samples is a major bottleneck. To alleviate this, Liu et al. (2023m) consider adopting contrastive learning approaches in addition to the supervised training for detection. Another significant challenge involves tackling paraphrasing attacks (Sadasivan et al., 2023; Krishna et al., 2023). To mitigate this problem, Hu et al. (2023b) propose to employ an adversarial learning approach to simultaneously train a detector and a paraphraser. Nevertheless, supervised training of a binary classifier tends to overfit their training data, resulting in a decline in performance when faced with cross-domain or unseen data. Additionally, fine-tuning LLM classifiers is limited in facing data generated by different models.

7.5 Current Limitations and Future Directions

Despite significant advancements in the domain of AIGC detection, several limitations still require future attention:

7.5.1 Fairness of AIGC Detection

Although state-of-the-art text detectors generally achieve high accuracy in experimental settings, as discussed by Liang et al. (2023b), perplexity-based text detectors exhibit a notable bias against text written by non-native speakers. Specifically, these detectors have been observed to misclassify TOEFL essays written by foreign writers more frequently than those by native speakers. This discrepancy may be due to the lower perplexity of non-native essays, which often display less linguistic diversity and richness. This issue may also affect minority languages, which tend to have higher perplexities compared to popular languages like English. Additionally, similar biases might exist in other modalities, such as image detection. Therefore, it is crucial to consider the fairness of detectors when designing future detection methods and to develop efficient methods for evaluating the fairness of AIGC detection methods.

Meanwhile, on the watermarking side, learnable watermarking methods might also exhibit biases toward out-of-distribution data points. For instance, if the watermarking encoder and decoder are trained on English text written by native speakers, the model might also have a higher misclassification rate on essays written by non-native speakers. Therefore, it is crucial to consider fairness in the development of learnable watermarking methods as well.

7.5.2 Robustness of Watermarks

Both text and image watermarks are susceptible to regeneration or post-processing attacks, such as paraphrasing (Kirchenbauer et al., 2023b) or diffusion purification (Zhao et al., 2023c). In contrast, semantic watermarks tend to be more robust against such attacks. However, because semantic watermarks typically require deep neural networks to decode the watermark signals, they are vulnerable to adversarial attacks (Saberi et al., 2023; An et al., 2024). Adversarial perturbations can also be developed to prevent regeneration and post-processing attacks (Liu et al., 2022a). Adversarial attacks remain a significant challenge even for classification tasks. Therefore, designing robust watermarks that can withstand both attacks is challenging and crucial.

7.5.3 Origin Attribution of Generated Images

Recent advancements in visual generative models have significantly improved the quality of generated images, raising concerns about their potential misuse. It is critical to develop methods to accurately identify the origin model responsible for generating a given image (Liu et al., 2022a). Especially, the scenarios are especially important and practical where access to the source model is restricted and only a limited number of images from the source model are available (Liu et al., 2022a).

8 Uncertainty

Though modern foundation models possess impressive capabilities across a wide range of domains and tasks, harnessing their power reliably requires a clear understanding of the uncertainties inherent in their outputs.

In the context of foundation models, uncertainty refers to the degree of confidence in the model's predictions or generated content. This uncertainty can arise from multiple sources: epistemic uncertainty, which is due to limited or imperfect knowledge encoded during training (e.g., gaps in the data or model misspecification), and aleatoric uncertainty, which stems from inherent randomness or ambiguity in the data itself (e.g., inherently noisy or ambiguous task definitions). However, it is important to note that this distinction also has limitations (Baan et al., 2023; Gruber et al., 2023; Ulmer, 2024; Mucsányi et al., 2024).

Beyond these, real-world deployment introduces further uncertainty due to distribution shifts, novel scenarios not represented during training, or unforeseen user inputs. Recognizing and managing these different types of uncertainty is essential for responsible and safe use of foundation models in high-stakes applications.

8.1 Sources of Uncertainty

Uncertainty in foundation models can be categorized into aleatoric and epistemic types. Aleatoric uncertainty is primarily influenced by data factors, while epistemic uncertainty is largely affected by modeling decisions. We will discuss the impact of these components in both the training and inference stages, with a comparison presented in Figure 14.

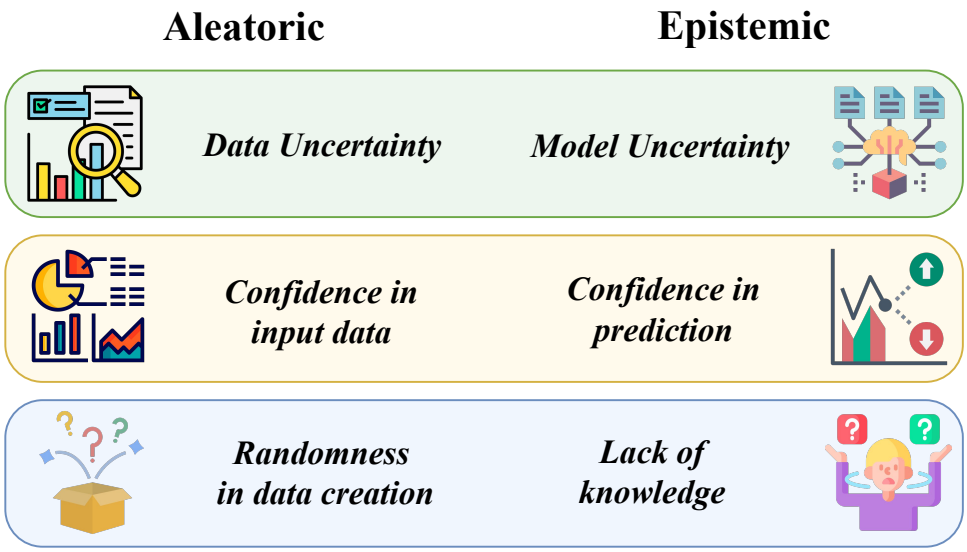

Figure 14: The comparison of aleatoric and epistemic uncertainty in machine learning. Aleatoric uncertainty originates from data variability, while epistemic uncertainty results from model limitations. The former is tied to inherent noise in observations, while the latter is tied to insufficient knowledge in representation.

8.1.1 Data

To begin, we will consider how the nature of language itself introduces uncertainty into the generation process of LLMs (Baan et al., 2023; Ott et al., 2018), which both consume natural language as input data and produce it as output data. Given some input context to a language model, there are usually many possible responses for several reasons. First, the input may be reasonably interpreted to have multiple different meanings. This could be because the context is vague ("She watched the man with binoculars"), very complex (as some reading comprehension questions are, even for humans), or contains spelling or other errors. Besides ambiguity in the input, certain queries may be inherently more open-ended and allow for

many reasonable responses. This might include a request to complete a fictional story, tell a joke, or give a position on some political or social issue. Finally, given a fixed input interpretation, equivalent answers to a query may be expressible in many ways. For example, given the input context "What is the capital of Rwanda?", "Kigali is the capital of Rwanda" and "The capital of Rwanda is Kigali" offer semantically equivalent answers with different surface forms.

The uncertainty in an LLM generation due to natural language data stems both from the training data and the prompt inputted to the model during inference. When the training dataset is small or not sufficiently general, the model may not have the relevant knowledge to effectively process some context (Osband et al., 2023; Hüllermeier & Waegeman, 2021; Lahlou et al., 2023; Pelrine et al., 2023). If the training data contains a large amount of ambiguous language, the trained LLM may reflect this uncertainty in its outputs. Other sources of uncertainty introduced by training data include diverse, conflicting, and outdated information.

In deployment, further uncertainty is introduced by specific text data given as input to the LLM. Queries could be ambiguous (Kuhn et al., 2023a; Kim et al., 2023b; Liu et al., 2023c), and tasks or instructions could be open-ended or underspecified (Tamkin et al., 2022), making it difficult for the model to express an appropriate level of confidence in its response. Also, relevant information could be excluded from the context (Yu et al., 2023d), or users may produce input errors.

8.1.2 Model

In addition to the uncertainty introduced by data in training and inference, the model itself also contributes to the uncertainty in the generation process, given an input prompt. Architecture choices may not reflect the underlying data-generating process. Different modeling techniques like ensembling (Lakshminarayanan et al., 2017; Malinin & Gales, 2021; Glushkova et al., 2021; Wang et al., 2024k) or Bayesian inference (Gal & Ghahramani, 2016; Ott et al., 2018; Xiao & Wang, 2021) can be applied with the hope of accurately characterizing the true posterior probability. However, these methods can be computationally expensive and potentially ineffective (Abe et al., 2022; Ovadia et al., 2019). Besides architecture, the typical optimization objective of producing the most plausible answer (i.e., maximizing observed sequence probability Eikema & Aziz, 2020; Eikema, 2024) may not align to produce the most correct and factual answer (Tian et al., 2023a), and in general, the cross-entropy objective has been shown to lead to overconfidence (Wei et al., 2022a). Finally, though massive pre-trained models are an effective tool for combating the uncertainty introduced by the input context, the popular approach of fine-tuning these LLMs for custom use cases may dilute these generalist capabilities (Yuan et al., 2023a).

8.1.3 Aleatoric vs. Epistemic Uncertainty

Besides identifying how uncertainty may arise due to data and model factors, it may also be useful to characterize uncertainty in LLM responses as either *aleatoric* or *epistemic* (Hüllermeier & Waegeman, 2021; Zhang, 2022). Aleatoric uncertainty, sometimes referred to as data uncertainty, exists due to the inherent randomness in the data-generating process. Additional information cannot be used to reduce aleatoric uncertainty. For example, suppose a language model was asked to predict the probability of heads on the flip of a fair coin. In that case, no additional context or training data would enable a better prediction than 50%. On the other hand, epistemic uncertainty arises precisely because of a lack of knowledge. Epistemic uncertainty may be reduced by incorporating additional data, for instance, by including or prioritizing more informative examples in the training set (Osband et al., 2023; Wang et al., 2023g) or incorporating appropriate few-shot examples in the context (Ye et al., 2023a; Diao et al., 2023; Li & Qiu, 2023; Su et al., 2022; Yu et al., 2023d). Section 8.2.1 discusses more of this work in detail.

8.2 Quantifying and Addressing Uncertainty

While some uncertainty in LLM responses is unavoidable, considerable progress has been made in quantifying and addressing this uncertainty so that models can be deployed responsibly and reliably. Common measures for quantifying uncertainty use notions such as entropy to characterize the uncertainty in response and produce higher scores for outputs that are less likely to be correct. Methods have been developed to recalibrate confidence scores so that they better reflect the true probability of an answer being correct. Addi-

tionally, uncertainty estimates can be used to identify cases where the system should abstain from answering or seek further clarification before providing a response. Researchers have also examined whether LLMs can generate linguistic expressions indicating their uncertainty for a given output (Ott et al., 2018; Si et al., 2023; Kuhn et al., 2023b). Finally, rigorous statistical methods, which are quickly gaining popularity in the deep learning community, have been applied to provide high probability bounds on LLM performance and risk (Ulmer et al., 2024c; Su et al., 2024; Ravfogel et al., 2023). Next, we will highlight important work in these areas, accompanied by an illustration in Figure 15.

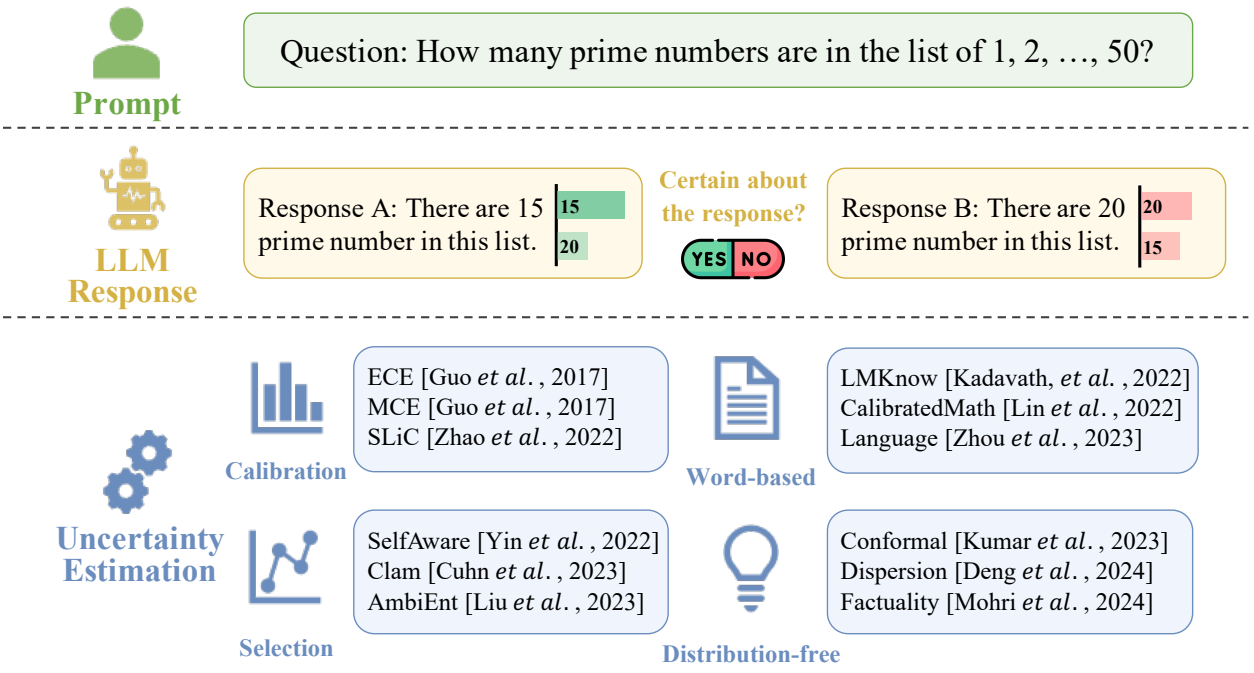

Figure 15: Overview of representative methods for estimating and mitigating uncertainty in foundation models. The illustration highlights how different answers may be generated with varying confidence levels, and categorizes existing approaches into calibration, word-based, selection, and distribution-free methods.

8.2.1 Estimating Uncertainty

One critical ingredient for the reliable deployment of a black-box LLM is the ability to measure the uncertainty in its responses so that appropriate decisions can be made based on its output (Si et al., 2023). Uncertainty in language model output is often quantified in terms of predictive entropy (Kuhn et al., 2023b; Lin et al., 2023b; Malinin & Gales, 2021; Duan et al., 2024a; Wang et al., 2024r). Although predictive entropy can be calculated directly using output class probabilities for classification tasks, measuring the uncertainty in language model generations is a more challenging problem, requiring knowledge of the distribution over all possible sequences.

One popular way to address the challenge of sequence-level uncertainty quantification is to sample many potential generations from the model and use these samples to estimate the underlying distribution. For instance, Fomicheva et al. (2020) use Monte Carlo dropout to draw samples, which are then utilized for quantifying uncertainty in machine translation. Malinin & Gales (2021) instead employ an ensemble with Monte Carlo sampling techniques over the token outputs to produce both token-level and sequence-level uncertainty scores. Such approaches to measuring uncertainty via self-consistency are also studied in (Kadavath et al., 2022; Lin et al., 2022; Si et al., 2023; Diao et al., 2023; Kuhn et al., 2023b; Wang et al., 2024q). While most self-consistency methods focus on the model's natural language outputs, Chen et al. (2024b) offer an effective method for measuring sample consistency in the embedding space. Besides, Stengel-Eskin & Van Durme

(2023a) explore sequence-level uncertainty in semantic parsing through the angle of calibration, where they leverage the minimum confidence across tokens.

In addition to the challenge of sampling from the distribution of possible sequences, there may be many equivalent surface forms of a correct response to a question such as "What is the capital of Germany?" Accordingly, the desirable notion of uncertainty may go beyond the spaces of sequences into the space of semantic meaning. These challenges are outlined and addressed by Kuhn et al. (2023b), wherein a set of sequences are sampled and grouped by semantic equivalence to measure uncertainty over meanings instead of uncertainty over output forms. Lin et al. (2023b) build on this work by proposing more sophisticated semantic uncertainty measures and removing the access requirement to the token-level scores of the potentially black-box model. Also, Duan et al. (2024a) presents an algorithm based on a similar notion that aims to better characterize uncertainty by focusing on the most relevant token.

Finally, some have taken the approach of trying to explicitly identify epistemic uncertainty, which can then be addressed by incorporating additional information. Malinin & Gales (2021) take an ensemble approach and characterizes epistemic or "knowledge" uncertainty using mutual information and the level of disagreement between models in the ensemble. Glushkova et al. (2021) also apply an ensemble-based approach to quantifying epistemic uncertainty in machine translation. The use of more powerful and efficient techniques such as direct uncertainty prediction and heteroscedastic regression are investigated in Zerva et al. (2022) in the context of machine translation; they find these methods perform favorably compared to variance-based baselines such as MC dropout and deep ensembles while being considerably faster. Lahlou et al. (2023) highlight the challenges of using Bayesian techniques or discrepancy-based measures of epistemic uncertainty and proposes a Direct Epistemic Uncertainty Prediction framework, wherein a secondary model is trained to estimate the point-wise generalization error and provides an upper bound on epistemic uncertainty. Their algorithm is shown to be useful in interactive learning environments, where the model can acquire novel examples and continue learning. Similarly, Osband et al. (2023) use an epistemic neural network to identify uncertain data that should be prioritized in fine-tuning, achieving on-par performance while using half as much data as training without prioritization. Hou et al. (2023b) avoid the need to train a separate model to predict the epistemic uncertainty, instead using multiple LLM queries for clarification to rule out data uncertainty so that the remaining uncertainty of each prediction can be prescribed to epistemic uncertainty. As another ICL-based approach, Yadkori et al. (2024a) propose an iterative prompting method to identify when epistemic uncertainty is large and highlight its usefulness in a setting with multiple good responses. More recently, Wang et al. (2024l) introduced Bayesian Low-Rank Adaptation (BLoB), which jointly updates both the mean and covariance of LLM parameters during fine-tuning to enhance uncertainty estimation, and Training-Free Bayesianization (TFB) (Shi et al., 2024a), a post-hoc method that converts pre-trained LoRA adapters into Bayesian models without retraining, offering efficient and accurate uncertainty quantification.

8.2.2 Calibration

One popular method for characterizing a model's predictive uncertainty is concerning (confidence) calibration. For a model to be well-calibrated, its confidence estimates should, on average, reflect the probability of its correct answers. The most common calibration measure in the deep learning literature is Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017), which measures the expected difference between confidence and accuracy over the data distribution. However, since it is impossible to calculate this quantity directly, ECE is typically estimated: data points with similar confidence scores are binned together, and ECE is calculated as the mean absolute difference between average confidence and accuracy over all bins. Other popular measures of calibration error include Maximum Calibration Error (MCE) (Naeini et al., 2015; Guo et al., 2017), Brier Score (Brier, 1950), negative log-likelihood (Hastie et al., 2001) and other novel variants (Ulmer, 2024). Since 0/1 accuracy is often not a suitable metric with respect to LLM performance, Huang et al. (2024d) propose Rank-Calibration Error, which captures whether higher uncertainty scores are associated with worse generations according to continuous metrics like ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002).

While modern neural networks have achieved impressive accuracy across a wide range of tasks and improved calibration relative to simpler methods (Minderer et al., 2021), significant miscalibration remains, usually in the direction of overconfidence (Guo et al., 2017; Wang et al., 2021). To address this remaining calibration

error, post-hoc recalibration methods such as Platt scaling (Platt, 1999), temperature scaling (Guo et al., 2017), or histogram binning (Zadrozny & Elkan, 2001) can be used to refine the confidence estimates of a pre-trained model.

Most of the work in calibration and deep learning has focused on the classification setting, where the softmax probability of a class can be reasonably interpreted as a confidence score. Accordingly, extending techniques for measuring and improving calibration to LLMs in classification (or other settings with single-token answers from a finite, discrete set) is straightforward. For example, Desai & Durrett (2020) find that pre-trained encoder-only transformer models like BERT and RoBERTa are well-calibrated under fine-tuning and that techniques like temperature scaling and label smoothing can be effective in combating poor confidence estimates. Further, Xiao et al. (2022) perform a large-scale analysis of how decisions made along the LLM deployment pipeline, such as model size, architecture, and training objective, affect downstream task calibration on sentiment analysis and NLI. They find that larger models generally give more accurate confidence estimates and that applying temperature scaling and fine-tuning with focal loss may be helpful. Zhao et al. (2021); Zhou et al. (2024b) offer methods to debias answers for a prompt so that confidence scores are calibrated based on the actual input instance under consideration, while Detommaso et al. (2024) introduce the notion of multi-calibration (Úrsula Hébert-Johnson et al., 2018) into the LLM setting by grouping examples based on binary attribute labels produced by the model itself. Kadavath et al. (2022) perform an extensive study of whether LLMs can evaluate the correctness of their own responses across tasks such as multi-choice question answering. They find that self-evaluation improves with model size, although calibration is worse for more complex and out-of-distribution tasks. They also find that popular alignment techniques such as RLHF may hurt the calibration of LLM output probabilities. To handle the case of population shift, e.g., across the distribution of subjects in a sample of MMLU questions, Li et al. (2024g) propose to train a recalibration method that adapts to a new subset of the data given only a few unlabeled examples.

On the other hand, estimating confidence and measuring calibration is less straightforward when tasks are generative or open-ended (for the same sequence-related reasons outlined in Section 8.2.1). Thus much recent LLM calibration research has focused here (Kadavath et al., 2022; Xiao et al., 2022; Singh et al., 2023a; Si et al., 2023; Tian et al., 2023b; Zhao et al., 2022; Mielke et al., 2022; Liu et al., 2024j). In early work highlighting this challenge, Ott et al. (2018) analyze model calibration in the setting of neural machine translation, showing that these models tend to diffuse too much probability mass over the space of possible sequences. One popular avenue for addressing the difficulties of combining calibration and generation is the development of new methods for producing calibrated sequence-level confidence scores. To this end, Chen & Mueller (2023; 2024) combine self-consistency with self-evaluation to produce a confidence score using a method they call BSDetector and find it is more accurate than alternatives in identifying incorrect LLM responses for models like GPT-3 and ChatGPT. Si et al. (2023) measure the calibration of GPT-3 on free-form QA using both the length-normalized language model output probability and self-consistency and finds both methods give more calibrated confidence scores than a supervised BERT baseline. Tian et al. (2023b) study LLM calibration of models aligned with RLHF, finding that these models can verbalize confidence scores that are more reliable than the underlying output probabilities, an approach which is especially useful when the model is behind an API and these probabilities are not available (see Section 8.2.3 for more on verbalized expressions of uncertainty). While most work on the calibration of LLM generations has focused on language tasks like question answering and summarization, Spiess et al. (2024) study the calibration of LLMs for code generation across several tasks, correctness criteria, datasets, and approaches.

In addition to the approaches described above, researchers have also pursued techniques for better quantifying LLM confidence via model training, concerning an external recalibrator or the LLM itself. For instance, Mielke et al. (2022) address conversational agents' overconfidence by training a small auxiliary network to predict the appropriate level of confidence to be expressed. Liu et al. (2024j) offer further work in this direction, proposing to train a new linear layer that predicts a bias term to be added to the language model's output logits. Their approach enables the reordering of candidate generations (as opposed to temperature scaling) and is tested on longer generations including full paragraphs. Kadavath et al. (2022) study whether a language model can be trained to predict the probability that a free-form answer to a question is correct; their experiments show promising results, although generalizing such behavior across distributions remains

challenging. Lin et al. (2022) use fine-tuning to teach a GPT-3 model to express its own uncertainty on various mathematics tasks, finding that responses are generally well-calibrated and remain reasonable under distribution shift. Finally, a supervised fine-tuning step is proposed in Band et al. (2024) to induce linguistic calibration, where model outputs feature confidence estimates that enable downstream decision-makers to make calibrated probabilistic predictions.

8.2.3 Verbalized Uncertainty

Generally, in machine learning, confidence scores are numeric values extracted from a predictive model, for example, based on predicted class probabilities, logit entropy, or ensemble variance. However, the ability of LLMs to generate arbitrary text output enables a paradigm in which language models may express their uncertainty directly in their natural language output.

As an early example of such an approach, Kadavath et al. (2022) verbalize language model calibration by verifying answers using the probability assigned to tokens such as "True" or "IK" ("I know") conditioned on its output or articulating confidence scores using numeric verbalizations such as "30%" or "80%". Their approach shows promise, although it may be difficult to generalize to new tasks or tasks that are difficult to format as multiple-choice. Additionally, Lin et al. (2022) fine-tune a GPT-3 to directly express its confidence in its output using verbalized probabilities (e.g., "61%"), while Tian et al. (2023b) prompt a model directly to output both confidence scores and linguistic markers of confidence (e.g., "highly likely"). Zhou et al. (2023b) study how linguistic markers of certainty, uncertainty, or evidentiality such as "I'm sure...", "I think...", or "Wikipedia says..." affect model confidence. Their findings imply that LLMs are sensitive to epistemic markers in prompts, with more than 80% variation in accuracy, and that expressions of high certainty result in a decrease in accuracy. Their results also suggest that the confidence scores that LLM outputs do not truly reflect epistemic and aleatoric uncertainty in response but instead are based on mimicking language use from the training set. This observation is supported by an extensive study of the ability of black-box models like GPT-4 to verbalize confidence in Xiong et al. (2023). They find the verbalized uncertainty expressions overconfident and difficult to optimize across models and datasets with a single strategy for prompting, sampling, and scoring. In other relevant work, Mielke et al. (2022) train a confidence calibration network to select linguistic expressions of uncertainty that should be included in the output of a conversational agent. Stengel-Eskin et al. (2024) propose LACIE, which splits verbalized uncertainty into explicit markers (e.g., I'm not sure) and implicit markers (e.g., giving details or backstory, stating a person's expertise, etc.). The models are trained to improve calibration by modeling a listener who accepts or rejects answers based on their correctness. The generator is rewarded for providing correct answers that are accepted and penalized for incorrect answers being accepted or correct answers being rejected.

8.2.4 Addressing Uncertain Examples

Selection is another established tool for addressing uncertainty (Geifman & El-Yaniv, 2017; 2019; Fisch et al., 2022; El-Yaniv & Wiener, 2010; Zollo et al., 2024a). We use the term selection broadly to encompass methods that identify inputs that are particularly difficult for the model. We offer interventions like allowing the model to abstain from the prediction (the classic paradigm in selection) or request further information. Selection has been well-studied in the context of language models (Cole et al., 2023; Kamath et al., 2020; Si et al., 2023), and has been shown to improve outcomes concerning hallucination and safety (Tomani et al., 2024). Kamath et al. (2020) investigate selective question answering under domain shift, proposing a novel algorithm that incorporates out-of-distribution data to train a selection model that identifies examples on which the model is likely to err. Gupta et al. (2024) derive a new score based on token-level uncertainty features, to identify examples that should be deferred from a smaller model to a larger model. Uncertainty scoring methods, for example, the semantic entropy-based measures proposed in Lin et al. (2023b), are often evaluated via selection to highlight how such measures are useful for predicting the correctness of LLM responses. Stengel-Eskin & Van Durme (2023b) show that we can recover low-confidence examples in semantic parsing by rephrasing and asking for user confirmation, which is the number of questions the model abstains from while keeping model safety high. To support work on selection in LLMs, Yin et al. (2023b) introduce the *SelfAware* dataset of questions that should be recognized as unanswerable.

Given the opportunity for interactivity provided by the text interface, a significant amount of research has gone towards algorithms to enable the LLM to request further information before responding, particularly in the case of ambiguous questions. For instance, Kuhn et al. (2023a) use few-shot learning to detect ambiguous questions that require clarifying questions, while Kim et al. (2023b) propose a tree-based approach to disambiguating questions and retrieving missing information. As the interest in identifying ambiguous questions in LLMs has grown, there has been an accompanying effort to release public datasets that can be used to evaluate the relevant abilities. Liu et al. (2023c) offer AmbiEnt, a dataset to test an LLM's ability to manage ambiguity in resolving entailment relations, finding their task difficult even for powerful commercial models like GPT-4. Additionally, Tamkin et al. (2022) introduce AmbiBench, a benchmark of ambiguous tasks where the ambiguity is introduced by the task description itself (as opposed to the specific instance of the task). Stengel-Eskin et al. (2023a) create a dataset for identifying and disambiguating instances of the visual question-answering task with MLLMs. Besides, Stengel-Eskin et al. (2023b) introduce a dataset of ambiguous queries and their logical forms and test whether models can recover both interpretations. Also, Saparina & Lapata (2024) introduce a similar ambiguous parsing dataset but with human-sourced SQL queries.

8.2.5 Distribution-free Uncertainty Quantification

As LLMs are increasingly deployed in risk-sensitive domains such as medicine, law, and finance, it may be important to have not only an estimate of the uncertainty in a model's response but also a high probability upper bound on the error rate at test time. Recently, there has been increasing research employing techniques from the Distribution-Free Uncertainty Quantification (DFUQ) family to control the risk of deep learning systems. This line of work generally descends from the literature concerned with conformal prediction (Shafer & Vovk, 2008; Vovk et al., 2005), wherein a threshold on class probabilities is calibrated to produce prediction sets that fulfill some coverage (i.e., recall) guarantee. Angelopoulos & Bates (2021) offer a tutorial on the subject in the context of modern neural network applications, and Kumar et al. (2023b) illustrate the application of conformal prediction to multi-choice question answering with LLMs. To broaden its applicability, Angelopoulos et al. (2023) derive a version of conformal prediction for bounding the expectation of any monotone loss function and studies their method in open-domain question answering. Recent work has offered algorithms for producing bounds on more general loss functions concerning the mean (Angelopoulos et al., 2021), quantile-based risk measures like value at risk (VaR) (Snell et al., 2022), and measures of statistical dispersion like the Gini Coefficient or differences in loss among protected subgroups (Deng et al., 2023c).

While it is straightforward to apply existing DFUQ techniques to classification with LLMs (Snell et al., 2022; Deng et al., 2023c; Kumar et al., 2023b), the question of how best to apply them to generation tasks like summarization, chat, and code remains open. Multiple approaches have been proposed to apply these techniques to language model decoding. For example, Schuster et al. (2022) utilize the Learn Then Test framework (Angelopoulos et al., 2021) to calibrate early exit criteria concerning the number of transformer layers applied to an input. Their goal is to identify when an LLM is sufficiently confident that it can exit the forward pass, and thus reduce the amount of computation used. In the conformal prediction vein, Quach et al. (2023) calibrate a stopping rule to produce a set of candidate generations that with high probability contains a suitable response (while removing redundant candidates), and Deutschmann et al. (2023) incorporate conformal prediction into a novel beam search algorithm. To mitigate the risk of models hallucinating answers, Yadkori et al. (2024b) proposes a conformal abstention procedure using measures of self-consistency that are evaluated by the LLM itself. Finally, Mohri & Hashimoto (2024) enforce factuality in LLMs by using conformal techniques to determine a level of specificity with which a given question can be answered.

As a more general approach, Prompt Risk Control (Zollo et al., 2024b) unites many techniques from the DFUQ family under a single framework for selecting a prompt (e.g., system prompt or set of few-shot examples) based on rigorous upper bounds on rich families of informative risk measures. The authors propose a two-step prompt selection process. First, a set of prompts is validated as producing an acceptable risk for some contextually relevant measure before a final prompt is chosen based on some performance metric, like average reward or accuracy. Prompt Risk Control can be applied to any bounded loss function, such as

top-1 accuracy, ROUGE, or toxicity, and can be used to control risk measures, including tail quantities like value-at-risk or measures of statistical dispersion such as the Gini coefficient.

8.3 Current Limitations and Future Directions

Much work has gone into methods to identify and address uncertainty in foundation model generation. However, existing results and methods are limited and much work remains to be done before these models can be responsibly and reliably deployed.

First, many results in uncertainty quantification in LLMs are produced in limited settings. Experiments are usually performed on tasks like trivia question answering, which can be answered via a single token, word, or short phrase. Further, the tasks under study also often assume that there is only one right answer: there may be no uncertainty in the correct response to "Who won Super Bowl XX?". However, much LLM usage revolves around tasks that require generating long-form responses to open-ended queries, for which multiple reasonable answers exist. Some works have made progress in this direction (Zhang et al., 2024b;c; Yoon et al., 2025), it is unclear whether the results produced in these limited settings offer insight into more complex, uncertain, and sequential settings, such as chat or customer care.

Alongside the difficulty of extrapolating results from simple settings, many of the recently proposed methods may not be suitable for application at scale in many relevant applications. Several recently proposed methods for improved uncertainty quantification of language model generations have come at the expense of generating multiple times for a single query. Soaring inference costs have become a major concern, prompting a surge of research aimed at curbing these expenses. Given that certain methods can increase costs by 2 to 20 times—or more—compared to standard inference, it seems improbable that many LLM users or service providers will adopt such approaches. Furthermore, although modern frontier models have shown some ability to express their uncertainty in words, there is good evidence that any correlation between accuracy and verbalized expressions of confidence is simply a result of spurious features in training data (Zhou et al., 2023b). In addition, it should be noted that these verbalization techniques also usually require extra inference costs, even for the simplest methods, such as scoring $p(\text{True})$ for the generated answer. Finally, although these algorithms have largely not been tested in open-ended tasks and over long generations, it seems probable that new tools will be needed in this setting. For example, consistency-based methods assume that producing diverse samples for a particular query indicates an example for which the model will likely give a poor answer. However, a model that can only produce a single answer to a query such as "Tell me a joke" or "Write me a story" would lack the capabilities to suit many modern LLM use cases.

Overall, it is unclear whether any advanced method for quantifying LLM uncertainty in the zero-shot setting robustly outperforms a baseline sequence entropy score calculated using token probabilities. However, these scores are often unavailable for black-box LLMs behind an API. Furthermore, it is difficult to imagine how best to exploit probabilities taken directly from the language model, since these probabilities do not necessarily relate to the task at hand (McCoy et al., 2023), but instead reflect the cross-entropy objective used in training and plausibility of an answer under the training data distribution (unless the model receives RLHF, which makes accurate uncertainty estimation even more difficult Kadavath et al., 2022; Tian et al., 2023b).

Besides addressing the limitations in methodology and experimental settings mentioned above, future work in this area may benefit from taking a broader view of the challenge of quantifying and addressing uncertainty in large generative models. It could explore how uncertainty can be better quantified and addressed across the entire model development and deployment pipeline, and how interventions and measurements at different points in the pipeline interact and affect downstream outcomes. Also, it may be useful to gain a more thorough understanding of how techniques for selecting, mixing, and filtering training data affect a user's ability to accurately estimate the model's confidence on downstream tasks, whether via token probabilities or verbalizations. As new architectures and pre-training recipes emerge, they should be benchmarked for calibration, not only accuracy. Fine-tuning algorithms, whether supervised or RL, have been shown to worsen models' UQ characteristics, and this phenomenon must be kept in focus as the community iterates on these methods. Finally, given a model that has been pre-trained and fine-tuned and is ready for deployment, we might develop new methods to select system prompts and few-shot exemplars that reduce and control

uncertainty in the wild, ideally with rigorous statistical methods like those provided by DFUQ (Zollo et al., 2024b).

9 Distribution Shift

Foundation models can occasionally produce unacceptable errors when faced with distribution shifts. These models, typically trained on a fixed corpus, require additional adaptation for new tasks. This limitation is particularly challenging in our ever-changing world, where knowledge is constantly shifting due to various factors, such as changes in location or time (Kasai et al., 2023; Kim et al., 2024). For instance, if a model trained before 2023 is asked, "Which team does Messi play for?", it may incorrectly assign a higher probability to Paris Saint-Germain instead of Inter Miami. This example highlights the importance of understanding, detecting, and mitigating distribution shifts in foundation models to improve their reliability.

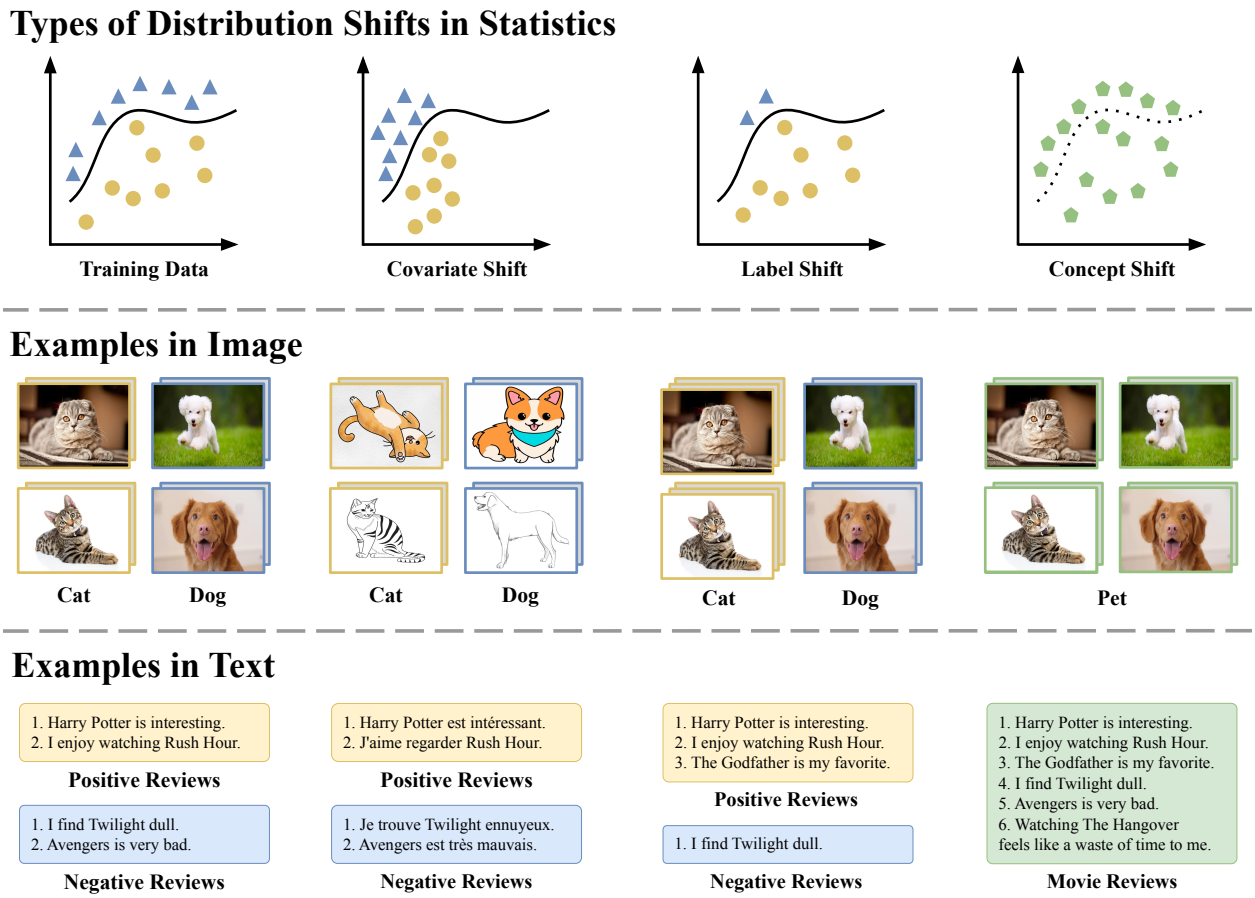

Figure 16: Different types of distribution shifts in the perspectives of (1) statistics, (2) image, and (3) text. The concept shift scenarios show how two distinct classes can merge into a single class when labels change.

9.1 Definition and Categorization

The distribution shift (Lakshminarayanan et al., 2017; Arora et al., 2021; Hupkes et al., 2023) occurs when the independent and identically distributed (i.i.d.) assumption does not hold between the training and test distributions. This divergence between the training distribution p_{train} and the test distribution p_{test} can significantly impact the performance of machine learning models, including foundation models. In essence, distribution shift describes the scenario where $p_{\text{train}} \neq p_{\text{test}}$, which can degrade model performance and reliability.

Based on how the data distribution changes, distribution shifts can be classified into three primary categories, with examples from various domains presented in Figure 16.

- **Covariate Shift.** This term refers to changes in the feature distribution $p(x)$ while the relationship between the features and the labels $p(y|x)$ remains unchanged. This type of shift is prevalent in scenarios where the environment or context of the features change.
- **Label Shift.** It occurs when the distribution of labels $p(y)$ changes, while the conditional distribution of features given labels $p(x|y)$ remains constant. This shift can result from changes in the real-world phenomena being modeled.
- **Concept Shift.** Concept shift, also known as conditional shift or concept drift, happens when the relationship between the features and the labels $p(y|x)$ changes. It reflects the evolution of the underlying problem statement or process over time.

9.2 Out-of-Distribution Detection

Out-of-distribution (OOD) detection involves identifying inputs different from the training distribution (Yang et al., 2021a; Fort et al., 2021), which plays a vital role in enhancing the reliability of foundation models. By flagging unfamiliar data points for further scrutiny, these techniques help mitigate risks and maintain the integrity of the model's performance. Despite the impressive generalization capabilities of today's extremely large foundation models, they remain fundamentally bounded by their training data. When deployed in dynamic open-world environments, such models can still encounter domain shifts, rare edge cases, or adversarial input that leads to overconfident but incorrect predictions. Therefore, OOD detection remains crucial: not only as a safeguard against catastrophic failures in high-stakes applications (e.g., medicine, autonomous driving), but also as a tool to trigger human oversight, guide active learning and preserve trust in automated decision-making systems.

In the context of language models, Liu et al. (2024b) present an empirical investigation into the OOD detection capabilities of LLMs, specifically examining the LLaMA families with different model sizes. The study evaluates common OOD detectors in both zero-shot and fine-tuning scenarios, yielding several significant insights: (i) LLMs inherently serve as effective OOD detectors without requiring fine-tuning. (ii) In-distribution (ID) fine-tuning can boost OOD detection. (iii) Generative fine-tuning demonstrates superior generalization ability because it aligns with the pre-training objectives of LLMs. (iv) A simple cosine distance OOD detector proves to be highly effective, attributed to the isotropic nature of LLM embedding spaces. Furthermore, Zhang et al. (2024a) propose a novel approach for OOD detection, utilizing the likelihood ratio between a pre-trained LLM and its fine-tuned variant. This method leverages the pre-trained LLM's extensive prior knowledge about OOD data, which, when fine-tuned with ID data, can effectively differentiate between ID and OOD samples. Expanding on these findings, Salimbeni et al. (2024) explore the effectiveness of unmerged Low-Rank Adaptor (LoRA) (Hu et al., 2021) weights for OOD detection during the fine-tuning process, further contributing to the growing body of research in this area.

In addition to textual OOD detection, recent advancements have begun to harness the powerful representation capabilities of foundation models in visual OOD detection. Dai et al. (2023b) propose a method to enhance OOD detection by selectively generating information from LLMs. Their method incorporates a consistency-based uncertainty calibration to estimate generation confidence scores and extracts visual objects from images to leverage the world knowledge encoded in LLMs. ODPC (Huang et al., 2024a) utilizes LLMs to generate specific prompts for creating "OOD peer classes," which are synthetic categories constructed from in-distribution (ID) semantics but intentionally placed outside the original label space. These peer classes act as proxy OOD categories during training, enabling the model to learn tighter ID class boundaries and better distinguish unfamiliar samples. This approach serves as an auxiliary modality for detection and introduces a contrastive loss based on OOD peer classes to learn compact ID class representations and clarify boundaries between different classes. EOE (Cao et al., 2024) improves OOD detection by tapping into the expert knowledge and reasoning capabilities of LLMs without requiring actual OOD data. This method is designed to adapt to various open-world scenarios, making it suitable for (i) *far OOD* detection, where the OOD samples come from entirely different domains (e.g., animals vs. vehicles); (ii) *near OOD* detection, where the OOD samples are semantically close but from unseen categories (e.g., unseen dog breeds when trained on other breeds); and (iii) *fine-grained OOD* detection, where differences are subtle and intra-class variation is high (e.g., distinguishing between visually similar medical conditions). In the medical domain,

CARES (Xia et al., 2024a) evaluate the OOD detection capability of medical LLMs, focusing on their ability to detect medical images that differ significantly from those used in the training phase.

9.3 Out-of-Distribution Generalization

OOD generalization, on the other hand, aims to enhance the robustness of foundation models under new, unseen environments (Hendrycks et al., 2021; Liu et al., 2021b; Yang et al., 2023e; Xia et al., 2024b; Nan et al., 2024). This approach improves the model's resilience to variations in input data through diverse techniques. Prior to the era of foundation models, the deep learning community explored a rich set of strategies for OOD generalization, supported by extensive empirical studies. These included (i) *data augmentation*, where transformations were applied to create synthetic training examples (Krizhevsky et al., 2012; Shorten & Khoshgoftaar, 2019); (ii) *adversarial training*, which exposed models to adversarially perturbed inputs (Goodfellow et al., 2014b; Madry et al., 2017); (iii) *label smoothing*, a regularization technique to prevent overconfidence (Szegedy et al., 2016; Müller et al., 2019); (iv) *invariant learning*, which aimed to capture features stable across environments (Arjovsky et al., 2019; Ahuja et al., 2020); and (v) *model ensembles*, which aggregate predictions from multiple models to reduce variance and improve robustness (Lakshminarayanan et al., 2017; Dietterich, 2000). Yuan et al. (2023a) evaluate these commonly used methods for LLMs, leading to important insights and conclusions.

9.3.1 Data Augmentation

Data augmentation (Zhang et al., 2017a; DeVries & Taylor, 2017; Yun et al., 2019) involves creating new training examples through various transformations of the original data. These transformations range from simple operations, such as flipping or rotating images in computer vision tasks, to more complex manipulations by generative models to simulate the data distribution (Li et al., 2020; Trabucco et al., 2023; Islam et al., 2024). In the NLP context, Easy Data Augmentation (EDA) (Wei & Zou, 2019) refers to a set of simple, low-cost textual augmentation operations — synonym replacement, random insertion, random swap, and random deletion — designed to increase lexical diversity without altering overall meaning. EDA was originally shown to be effective for small-scale text classification tasks, but its naive application to LLMs often degrades performance due to distributional shifts in token usage and disruption of learned long-range dependencies. The primary objective of data augmentation is to increase the diversity of the training set, thereby enabling the model to learn more robust features that generalize better to unseen data. However, recent research has shown that applying simple augmentation techniques, such as EDA, to LLMs often leads to performance degradation across most tasks, underscoring the need for more advanced augmentation methods tailored to foundation models.

9.3.2 Adversarial Training

Adversarial training (Madry et al., 2017; Bai et al., 2021) is a robust technique used to improve OOD generalization by exposing models to adversarial examples during the training process. These adversarial examples are inputs deliberately perturbed to mislead the model into making incorrect predictions, despite appearing similar to regular data. In earlier deep learning literature, adversarial training was shown to improve robustness in image recognition (Goodfellow et al., 2014b; Tramèr et al., 2017), and it is now increasingly applied to LLMs and MLLMs.

Mechanistically, adversarial training operates by solving a min-max optimization problem (Madry et al., 2017; Zhang et al., 2019): the inner maximization finds the worst-case perturbation within a certain norm-ball around each input, while the outer minimization updates model parameters to correctly classify these perturbed inputs. By repeatedly training on such challenging examples, the model learns smoother and more stable decision boundaries, which are less sensitive to input shifts, thereby improving robustness and generalization to unseen or distribution-shifted data.

In LLMs, Free Large-Batch (FreeLB) (Zhu et al., 2019), an adversarial training method that adds perturbations to the input data, improves generalization performance in most scenarios. Similarly, Verma et al. (2024) introduce image perturbations in MLLMs through augmentations like noise addition, blurring, and median filtering. Additionally, they craft adversarial questions using conjunctions, disjunctions, and nega-

tions to challenge models' reasoning abilities. Among the tested augmentations, Gaussian Noise Addition is identified as the most detrimental, causing the largest decline in performance. The study also finds that the complexity of questions, especially those with multiple connectives, significantly impacts the models' performance.

9.3.3 Label Smoothing

Label smoothing (Szegedy et al., 2016) is a regularization technique used to improve OOD generalization by preventing the model from becoming overly confident in its predictions (Müller et al., 2019). Unlike traditional training algorithms where models learn to assign a probability of 1 to the correct class and 0 to all others, label smoothing introduces a small probability to incorrect classes. This approach encourages models to maintain a degree of uncertainty in their predictions, potentially improving their ability to generalize to unseen data. In the context of LLMs, however, the effectiveness of label smoothing has been called into question. Yuan et al. (2023a) conducted experiments where they smoothed the hard labels in the training data but observed that this technique did not improve the LLMs' generalization ability.

9.3.4 Invariant Learning

Invariant learning (Arjovsky et al., 2019) plays a crucial role in OOD generalization by capturing invariant representations or predictors across different environments while disregarding spurious correlations. One notable approach of invariant learning involves the use of specialized loss functions, such as Focal Loss (Lin et al., 2017), Dice Loss (Sudre et al., 2017), and Mixup Loss (Zhang et al., 2017a). Focal Loss was originally designed for class-imbalanced detection tasks, down-weighting well-classified examples to focus training on harder cases. Dice Loss, derived from the Sørensen–Dice coefficient, is widely used in segmentation to maximize overlap between predicted and ground-truth regions, thus emphasizing recall. Mixup Loss linearly interpolates both inputs and labels between pairs of examples, encouraging the model to behave linearly in-between training samples and reducing overfitting to spurious patterns. These loss functions differ in their inductive biases — e.g., emphasizing difficult examples, optimizing overlap, or encouraging linearity — but all aim to produce more generalizable decision boundaries that are less sensitive to environment-specific correlations. By applying Focal Loss to the training process of LLMs, these models emphasize hard-to-classify examples and enhance their ability to handle diverse and unfamiliar inputs.

9.3.5 Model Ensemble

Model ensemble (Arbib, 2003; Lakshminarayanan et al., 2017) is a powerful technique for enhancing the robustness and performance of AI models in complex environments. This approach combines predictions from multiple models to produce more accurate and reliable final outputs. Yuan et al. (2023a) evaluated model ensembling but observed limited improvement in generalization ability. However, building on this foundation, recent studies by Jiang et al. (2023) and Wan et al. (2024) have introduced more advanced model ensemble algorithms, improving performance across several downstream tasks.

9.4 Domain Adaptation

Unlike OOD generalization, domain adaptation tailors the model to domain-specific tasks by injecting domain-specific knowledge (Ge et al., 2024; Siriwardhana et al., 2024), including in-context learning (ICL), retrieval-augmented generation (RAG), fine-tuning, test-time training, and model editing. These methods enable foundation models to specialize in particular domains while maintaining their broad capabilities.

9.4.1 In-context Learning

In-context learning (ICL) shows great potential to address the gap between foundation models and domains not covered in their pre-training and fine-tuning data (Dong et al., 2022b; Min et al., 2022). Recently, ICL has gained attention as a transformative approach for foundation models (Bar et al., 2022; Zhang et al., 2023h; Huang et al., 2024e). It demonstrates the ability to adapt to new tasks or distributions without altering model parameters by adding domain-specific input-output pairs to the test example. This augmented input

serves as a guide, helping the model produce desired outputs for new tasks. Consequently, ICL offers a flexible and efficient method for continual adaptation without the need for computationally expensive retraining.

In the field of LLMs, the BOSS benchmark (Yuan et al., 2023a) explores ICL for LLMs by using examples from both ID datasets and the training split of OOD datasets. The findings reveal that fine-tuning domain-specific models is advantageous when sufficient training data is available, while LLMs with ICL perform better in low-resource scenarios. Notably, the effectiveness of ICL varies across models and tasks, highlighting the need for task-specific adaptation strategies. Complementing this research, Reizinger et al. (2024) delve into the intricacies of ICL, focusing on its approximate non-identifiability and the implications for understanding LLMs. Through a combination of mathematical examples and empirical observations, their work demonstrates how this approximate non-identifiability manifests in OOD generalization, providing deeper insights into the behavior of ICL in various contexts.

For MLLMs, Zhang et al. (2024h) demonstrate that ICL can significantly enhance the generalization capabilities, suggesting new approaches to overcome existing limitations. However, their study also investigates the robustness of ICL under various distribution shifts. The findings reveal that ICL is vulnerable to domain shifts, label shifts, and spurious correlation shifts between in-context examples and test data.

9.4.2 Retrieval-augmented Generation

Retrieval-augmented generation (RAG) enhances foundation models by retrieving relevant information from external data sources to supplement input queries or generated outputs (Khandelwal et al., 2019; Min et al., 2020; Asai et al., 2024). This process provides necessary domain knowledge, mitigating distribution shifts and improving generation quality (Gao et al., 2023b; Kang et al., 2024a; Siriwardhana et al., 2023; Zhou et al., 2025a). In practice, RAG techniques are effective and efficient to apply in various unseen tasks with simple adaptation of the retrieval component, requiring minimal or even no additional training (Ram et al., 2023).

In the context of language models, Shao et al. (2024) construct MASSIVEDS, a massively multi-domain database comprising 1.4 trillion tokens of both general web data and domain-specific data. Their findings demonstrate that as the database's size and diversity increase, more distributions are covered during inference, reducing OOD scenarios. To incorporate this domain knowledge without requiring additional training, recent studies (Shi et al., 2023c; Ram et al., 2023) focus on in-context Retrieval-Augmented Language Models (RALMs). These models directly input a concatenation of all retrieved texts as additional context to LLMs. For the choice of retriever, most work (Zhang et al., 2023e; Shao et al., 2023; Neelakantan et al., 2022; Seo et al., 2024) employ an embedding model to decide what to retrieve. However, with the increasing prevalence of LLMs, researchers have begun using the models themselves as retrievers to improve accuracy (BehnamGhader et al., 2024; Ma et al., 2024b; Weller et al., 2024; Liu et al., 2024m; Wang et al., 2023d). In a parallel direction, as these RAG methods may retrieve irrelevant information that even hurt the performance, Zhang et al. (2024g) proposed RAFT to further fine-tune the LLMs to learn to disregard "distractor documents" within the provided context, thereby enhancing the model's ability to focus on relevant information. The effectiveness of these In-Context RALMs has been further demonstrated in several domain-specific tasks (Xu et al., 2024a; Li et al., 2024f; Xiong et al., 2024a; Lozano et al., 2023), showcasing the potential of RAG in addressing real-world distribution shifts.

To extend RAG to multimodal query input (Zhao et al., 2023b), Wei et al. (2023c) create M-BEIR, a multimodal instruction-following benchmark building on existing 10 diverse datasets. UniIR is trained on M-BEIR to take a heterogeneous query to retrieve from a heterogeneous candidate pool with millions of candidates in diverse modalities. Built upon it, UniRAG (Sharifymoghaddam et al., 2024) employs UniIR's CLIP Score Fusion and BLIP Feature Fusion models as retrievers, improving performance in MLLMs. For visual question answering (VQA) tasks, RA-VQA (Lin & Byrne, 2022) proposed a novel framework for joint training of the retriever and the answer generator, and FLMR (Lin et al., 2023a) further improved the retrieval accuracy by combining multi-dimensional embeddings from language and vision models. Similarly, MuRAG (Chen et al., 2022) uses T5 (Raffel et al., 2020) and ViT (Dosovitskiy et al., 2020) for text and image encoding respectively, and retrieval from a large-scale memory bank for knowledge-based VQA. To improve embodied agents, MART (Yue et al., 2024) utilizes interaction data to fine-tune a multimodal retriever based

on preference learning. For image captioning and text-to-image generation tasks, RA-CM3 (Yasunaga et al., 2022) enhances performance by using a pre-trained CLIP model to augment inputs for a CM3 Transformer. These methods effectively address the shift in knowledge representation across modalities. Additionally, domain-specific multimodal RAG solutions have shown promising results in various fields (Xia et al., 2024d; Kumar & Marttinen, 2024; Tao et al., 2024).

9.4.3 Fine-Tuning with New Knowledge

Fine-tuning is a widely adopted method for addressing domain adaptation in foundation models (Reizinger et al., 2024; Yuan et al., 2023a; Kirk et al., 2023). This technique involves adapting pre-trained models to specific downstream tasks by further training them on task-specific datasets. The primary goal is to enhance the model's performance on new, unseen data that may differ from the data it was initially trained on.

The BOSS benchmark (Yuan et al., 2023a) evaluates vanilla fine-tuning for LLMs, which involves directly fine-tuning pre-trained models on ID datasets without any additional processes. This benchmark helps investigate the relationship between performance on ID and OOD datasets by varying factors such as model scale, training steps, available training samples, and tunable parameters. Observations indicate that fine-tuning with the full dataset generally yields superior performance for ID examples, while LLMs employing in-context learning (ICL) paradigms demonstrate better performance on OOD instances. Reizinger et al. (2024) explore the non-identifiability of fine-tuning in LLMs, highlighting its implications for understanding and improving these models. They argue that fine-tuning is non-identifiable, meaning that models with similar fine-tuning performance (such as equivalent test loss) can exhibit markedly different behaviors when applied to real-world tasks.

To address OOD generalization, Kirk et al. (2023) investigate Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), which is typically implemented in three stages: (i) supervised fine-tuning (SFT), where the model is aligned with high-quality human-labeled data; (ii) reward modeling, where a learned reward model predicts preference scores for outputs; and (iii) reinforcement learning, where the base model is optimized against the reward model. While SFT is the first step in RLHF, it can also be viewed as a standalone fine-tuning approach. In their experiments, the authors compare the full RLHF pipeline against using only SFT and find that RLHF generally yields stronger generalization to new, unseen inputs, especially under significant distribution shifts between training and testing data.

Jiang et al. (2024) propose a novel method for fine-tuning LLMs in domains where obtaining large volumes of high-quality, domain-specific data is challenging, such as healthcare or harmless content generation. They re-evaluated the Transformer architecture to identify the most impactful parameter updates. Their analysis revealed that within the self-attention and feed-forward networks of the Transformer architecture, only the attention parameters significantly benefit downstream performance when there is a mismatch between the training and test set distributions. Based on this insight, they proposed Training All parameters but Inferring with only Attention (TAIA), which involves updating all parameters during training but utilizing only the fine-tuned attention parameters during inference. Additionally, recent studies have observed that parameter-efficient fine-tuning (PEFT) methods, such as Low-Rank Adaptor (LoRA), can maintain more general capabilities from the pre-trained distribution while acquiring new knowledge from the fine-tuning data (Biderman et al., 2024).

Beyond parameter-efficient methods, recent work has explored *activation steering* as an alternative or complement to conventional fine-tuning for improving generalization. This line of research modifies model activations at inference time or during light-weight training to achieve desired behavioral shifts without large-scale parameter updates. For example, Lai et al. (2025) propose Joint Localization and Activation Editing, which identifies and edits specific activation subspaces relevant to the target domain, enabling effective low-resource fine-tuning. Similarly, Turner et al. (2023b) introduce Activation Addition, a technique for steering model outputs by adding direction vectors in activation space, allowing the incorporation of new knowledge or behavioral adjustments without gradient-based optimization. These approaches can reduce overfitting to in-distribution features while selectively enhancing capabilities relevant for OOD settings.

For multimodal scenarios, the proposal of EMMA (Yang et al., 2024f) adapts LLMs to the field of embodied multimodal agents. The key technique involves distilling the reflection outcomes of the LLM, which improves

actions derived from analyzing mistakes in text world tasks. It uses these outcomes to fine-tune the vision-language models on analogous tasks in the visual world, which is capable of quickly adapting to the dynamics of the visual world. The cross-modality imitation learning is facilitated by a novel DAgger-DPO algorithm, which ensures that EMMA can generalize to a wide range of new tasks without further guidance. Belyaeva et al. (2023) describe a method to address OOD challenges by developing a framework called HeLM (Health Large Language Model for Multimodal Understanding). HeLM integrates multiple data modalities, learns robust data encodings, and enhances predictive performance through comprehensive data utilization to achieve OOD generalization. Regarding model architectures, Ito et al. (2024) find that models with multiple attention layers or those leveraging cross-attention mechanisms between input domains perform better in their constructed gCOG benchmark. Their study emphasizes that cross-modal attention and deeper attention layers are crucial for integrating multimodal inputs and improving generalization in the presence of distractors and new tasks.

9.4.4 Test-time Training

Test-time training methods view each test instance as an individual learning problem with its own generalization target. This method creates a self-supervised learning task for each test sample and updates the model parameters at test time before making a prediction. For LLMs, Sun et al. (2024a) proposes a new class of sequence modeling layers called Test-Time Training (TTT) layers. These layers transform the hidden state into a machine learning model, with the update rule functioning as a step in self-supervised learning. By aligning the training and test data distributions, these methods significantly enhance model performance when faced with distribution shifts.

9.4.5 Model Editing

All the domain adaptation methods discussed above modify a model's behavior by incorporating new knowledge. This process is closely related to model editing for foundation models (Wang et al., 2023f; Yao et al., 2023c), which aims to rectify specific errors without affecting unrelated inputs. To explore its potential in addressing distribution shifts, we will now provide an overview of model editing approaches, which typically adhere to three essential properties:

- Reliability: The edited model should successfully produce the desired output for the edited sample, such as correctly answering 'Inter Miami" when asked Who does Messi play for?"
- Generality: The corrections made should be consistent across equivalent contexts, for example, accurately responding to 'Which team is Messi in?"
- Locality: The acquired knowledge should be minimally affected, ensuring that unrelated queries like 'Who does LeBron James play for?" remain unaffected.

These properties ensure the reliability, generality, and locality necessary for the effective and efficient correction of foundation model behaviors. Recent studies in model editing (Hewitt et al., 2024; Akyürek et al., 2023) have also demonstrated promising performance in several OOD scenarios. Next, we will delve deeper into four distinct categories of model editing in LLMs (Figure 17), subsequently extending our discussion to address related issues in MLLMs.

Memory-based Model Editing. In memory-based approaches, an external memory, outside the intrinsic architecture of the pre-trained LLM, serves as a repository for edited knowledge. LLM can access and modify this external memory during inference. For example, Language Patch (Murty et al., 2022) performs editing by integrating with a library of patches in natural language, and MemPrompt (Madaan et al., 2022) adopts a growing memory bank as a look-up table to store the edit sample and its corresponding prompts, which is used to alter the prediction of the edit sample. KAFT (Li et al., 2022a) further strengthens the controllability and robustness of LLMs' working memory through counterfactual data augmentations. In this approach, the entity representing the answer in the context is substituted with an alternative

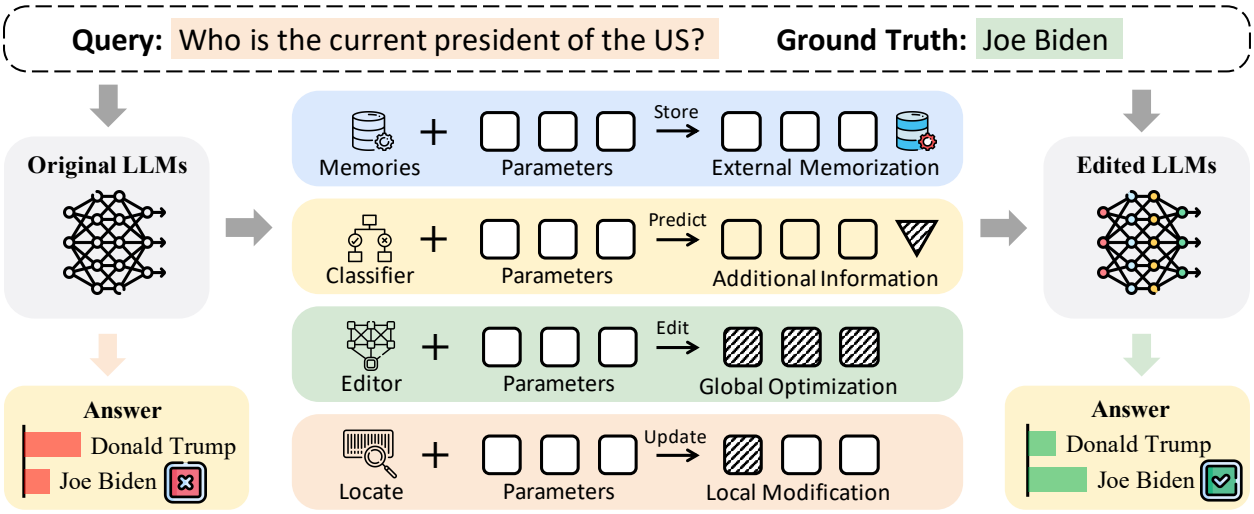

Figure 17: An overview of model editing methods in LLMs. Given an incorrect response from the original model, different editing strategies correct factual errors by modifying or augmenting the model's knowledge.

but still plausible entity. This substitution is intentionally designed to introduce a conflict with the genuine ground truth, thereby incorporating counterfactual and irrelevant contexts to standard supervised datasets. In addition to relying on parameter-based memory, IKE (Zheng et al., 2023b) introduces novel factual information into a pre-trained LLM via in-context learning, where a set of demonstrations will alter the prediction of a target factual detail when the input is influenced by an edit. To solve more complex questions involving chains of facts, MQuAKE (Zhong et al., 2023) enables editing by breaking down each question into iterative subquestions and retrieving the most pertinent fact from the edited fact memory.

Classifier-based Model Editing. The classifier-based model editing paradigm aims to preserve pre-trained parameters while utilizing a classifier to determine whether behavior adjustment is necessary. In this approach, if a sample falls outside the scope of the edit sample, the original model is applied to maintain predictions. Conversely, interventions occur when the sample is within the scope, with the specific interventions varying across different methods. SERAC (Mitchell et al., 2022) employs a scope classifier to determine whether the original model or a new lightweight model should be used for prediction. The new lightweight model is specifically trained for in-scope samples. In contrast, Language Patch (Murty et al., 2022), CaliNET (Dong et al., 2022a), and T-Patcher (Huang et al., 2023d) introduce additional trainable parameters to adapt the original model instead of requiring entirely new models. For example, Language Patch trains a new gating head (acting as a classifier) to combine predictions from the original prediction head and a newly trained interpreter head. CaliNET and T-Patcher insert a residual block into the original model's feed-forward network (FFN) as an adapter. This adapter utilizes an activation operation on hidden states to determine whether the intervention should be activated. When the activations are zero, there will be no change to the original prediction. However, the success of these classifier-based methods heavily relies on the quality of the classifier, which also necessitates a substantial number of unrelated samples for training. Alternatively, GRACE (Hartvigsen et al., 2022) edits a model by adding a retrieval-based adaptor to a chosen layer that enables judicious decisions regarding the utilization of the dictionary for a given input, accomplished via the implementation of a deferral mechanism.

Hypernetwork-based Model Editing. The hypernetwork-based model editing paradigm utilizes an external model, referred to as the editor, to facilitate parameter updates in the models. Knowledge Editor (KE) (Cao et al., 2021) employs a bidirectional LSTM to transform an edit pair, consisting of the edit sample, incorrect prediction, and correct label, into shifting operation parameters (i.e., mask $\mathbf{m}$, offset $\mathbf{b}$, and scaling factor α) for ∇: $\hat{\nabla} = \alpha(\mathbf{m} \odot \nabla) + \mathbf{b}$. Based on KE, SLAG (Hase et al., 2023) further

appends metrics for two types of input texts: (1) those that, while not part of the targeted edit set, align logically with it; and (2) those that share a formal resemblance to edited knowledge, but do not affect the prediction outcomes. However, hyper-networks are generally incapable of updating LLMs due to the massive parameter size. To address this issue, MEND (Mitchell et al., 2021) applies low-rank decomposition to ∇ and utilizes two MLP layers to generate a new low-rank update, $\hat{\nabla}$. This approach is lightweight and efficient, particularly for large models like T5-11B. Moreover, KGEditor (Cheng et al., 2023b) combines the benefits of memory-based methods and hypernetworks to ensure flexibility and further reduce computation costs. In particular, it introduces an additional feed-forward networks (FFNs) layer for knowledge storage. It then employs a bi-directional LSTM to encode embeddings of triples. In this manner, KGEditor becomes an efficient way to edit knowledge graph embeddings. Despite the success of this paradigm, the editors need to undergo a prior training stage. The availability of training data, including edit samples and pre-training data, poses a critical challenge. While these methods employ synthetic edit samples (e.g., selecting hypotheses via beam search except the top-1 for Question-answering tasks Cao et al., 2021), their generalization to realistic mistakes beyond the synthetic sample distribution remains limited.

Knowledge-based Model Editing. The knowledge-based model editing paradigm focuses on identifying a subset of parameters specifically associated with particular pieces of knowledge and only updating those parameters. This approach assumes that knowledge is stored within the FFNs, which function as key-value memories (Geva et al., 2022). Knowledge Neuron (KN) (Dai et al., 2021) attributes knowledge parameters using integrated gradients (Sundararajan et al., 2017), where more salient gradients indicate a greater influence on the knowledge. Building on this idea, Rank-One Model Editing (ROME) (Meng et al., 2022a) uses causal tracing to localize the specific FFN layer whose activation most strongly mediates the recall of a target factual association. Once the target layer is identified, ROME performs a rank-one update to its value projection matrix, effectively replacing the original stored fact with a new subject–object mapping. MEMIT (Meng et al., 2022b) extends ROME by identifying a set of relevant layers (e.g., layers 3–8 for GPT-J) and applying a closed-form multi-layer update. This allows MEMIT to edit multiple facts in parallel while preserving surrounding model behavior. It is important to note that these methods do not establish that these layers are exclusively dedicated to a single piece of knowledge, implying that the layers may be shared across different knowledge domains (Gandikota et al., 2024). To mitigate potential effects on out-of-scope samples, regularization techniques are employed during the neuron/layer updates. For example, MEMIT enforces the model to maintain predictions for several unrelated samples. By adopting a knowledge-based approach, these methods selectively update parameters associated with specific knowledge while minimizing interference with unrelated samples. Based on ROME, BIRD (Ma et al., 2023b) studies the novel problem of Bidirectional Assessment for Knowledge Editing (BAKE), which evaluates the reversibility of edited models in recalling knowledge in the reverse direction of editing and incorporating the bidirectional relationships between subject and object in an edit fact into the updated model weights.

Model Editing in MLLMs. Compared to single-modal model editing, the task of editing MLLMs is more challenging due to their inherent diversity and complexity. Specifically, errors in MLLM outputs can be attributed to the synergistic effects of various modalities. A recent study (Cheng et al., 2023a) introduces a pioneering benchmark for MLLM editing, named MMEdit. This benchmark evaluates three aforementioned key principles: Reliability, Locality, and Generality, and covers two specific sub-tasks: Editing VQA and Editing Image Captioning. Empirical evidence indicates that while current methodologies (Cao et al., 2021; Zheng et al., 2023b; Mitchell et al., 2021) are effective for editing the textual model in MLLMs, they fall short in editing the vision module. Researchers are encouraged to explore innovative techniques for efficient and accurate editing across various modalities and to develop comprehensive benchmarks for evaluating larger MLLMs. The end of 9.4.5 could profit from a discussion on how the different model editing methods differ and what kind of advantages and disadvantages they carry.

Overall, the four paradigms of model editing discussed above differ substantially in their mechanisms and trade-offs:

- *Memory-based methods* store edits externally, avoiding interference with model parameters. They are easy to update and revert but introduce extra retrieval latency and depend on effective memory indexing.

- *Classifier-based methods* preserve original parameters and selectively activate edits only when needed, offering strong locality. However, they rely heavily on a high-quality scope classifier and require ample negative examples to prevent over-triggering.
- *Hypernetwork-based methods* generate parameter updates dynamically from an edit description, enabling flexible and lightweight adaptation. Their main limitations are the need for a pre-trained editor network and reduced scalability to very large models unless combined with low-rank or parameter-efficient techniques.
- *Knowledge-based methods* directly modify internal representations linked to specific facts, often achieving high reliability and generality with minimal changes. Yet, they risk unintended side effects if the targeted layers store multiple pieces of unrelated knowledge, and they require accurate localization of the relevant parameters.

In multimodal settings, these trade-offs can be amplified: memory- and classifier-based methods may generalize more easily across modalities but depend on modality-aware retrieval/classification, while knowledge- and hypernetwork-based methods may offer more precise edits but require sophisticated cross-modal localization strategies. Future research may benefit from hybrid approaches that combine the precision of parameter-based edits with the flexibility and safety of external-memory or classifier gating mechanisms.

9.5 Current Limitations and Future Directions

Foundation models, despite their remarkable capabilities, face several challenges when confronting distribution shifts. These limitations primarily stem from inherent difficulties in OOD detection, generalization, and adaptation. Such challenges significantly impact the reliability and robustness of these models in real-world scenarios. When exposed to data that deviates from their training distribution, these models often exhibit decreased performance (Yuan et al., 2023a; Zhang et al., 2024h), leading to unreliable predictions in dynamic environments where data characteristics frequently change.

While various OOD detection methods have been developed, many struggle with scalability issues, making them less practical for large-scale deployment. Current approaches to OOD generalization and adaption, such as domain adaptation (Yuan et al., 2023a; Kirk et al., 2023; Yang et al., 2024f) and adversarial training (Bai et al., 2021; Yuan et al., 2023a; Verma et al., 2024), demonstrate varying degrees of success across different domains. These methods often require extensive retraining or fine-tuning to handle new domains effectively, a process that can be both resource-intensive and time-consuming. Furthermore, many techniques for improving OOD robustness heavily depend on the availability of large, high-quality datasets (Yuan et al., 2023a; Yang et al., 2024f; Belyaeva et al., 2023; Ito et al., 2024). This dependence poses significant challenges in domains where data is scarce or expensive to obtain. Additionally, for multimodal foundation models, effectively integrating and processing diverse data types remains a complex task. Current editing and generalization methods often fall short in scenarios involving multiple modalities, such as text, images, and audio (Wu et al., 2023c). Last but not least, modern foundation models often undergo continual pre-training and fine-tuning, either horizontally across a sequence of domains or vertically from a general-purpose model to a domain-specific model (Shi et al., 2024b). As a result, they inevitably tend to suffer from catastrophic forgetting, such as *horizontal forgetting* (Shi et al., 2024b) when continually adapting across domains and *vertical forgetting* (Shi et al., 2024b) when continually adapting from more general models to more domain-specific models.

To address these limitations, future research should focus on developing more lightweight OOD detection and generalization methods. These approaches should aim to identify and mitigate distribution shifts in large-scale settings while maintaining low resource requirements. By focusing on efficiency, such methods could be more readily integrated into practical applications, enhancing the robustness and reliability of foundation model systems across diverse real-world scenarios.

To adapt to rapidly evolving environments, we should prioritize the development of continual or even lifelong learning mechanisms for foundation models (Yang et al., 2024a; Shi et al., 2024b; Kim et al., 2024). These mechanisms would enable models to adapt to new data distributions without requiring extensive retraining (Li et al., 2022a) while simultaneously preserving knowledge acquired from previous training data,

including data previously used during pre-training or from previous domains. In other words, they should remain robust against both *vertical* and *horizontal forgetting* (Shi et al., 2024b). This approach could significantly enhance the flexibility and longevity of foundation models in dynamic domains. Additionally, due to the scarcity of data in several domains, developing more data-efficient transfer learning algorithms or creating diverse synthetic data will help models generalize to more practical applications.

To further improve the generality of foundation models, advancing their abilities to handle multimodal data effectively is essential, with unified frameworks that can seamlessly integrate various data types and leveraging techniques like cross-modal learning and multimodal embeddings enhancing performance in complex scenarios (Wu et al., 2023c; Zhang et al., 2024h; Yin et al., 2023c; Yu et al., 2024b). By addressing these limitations and exploring these future directions, we can significantly improve the robustness and reliability of foundation models, ensuring their effective deployment in diverse real-world applications.

10 Explainability

There are substantial existing efforts tailored towards the explainability of foundation models, particularly LLMs. In this section, we demonstrate the literature on the explainability of LLMs from the following aspects: (1) Feature Attribution Methods, i.e., Explaining LLMs with the raw features (words, sentences, syntax); (2) Exploring the inherent knowledge incorporated in LLMs themselves; (3) Discovering the roles and training samples in pre-training, fine-tuning, and few-shot learning. Following an overview of the methods used for model explanation, we dive into the evaluations and applications of explainability in LLMs. The discussion then broadens to include multimodal large language models (MLLMs), emphasizing the ongoing efforts in the field. Figure 18 provides a detailed overview of various methods to explain different foundation model components.

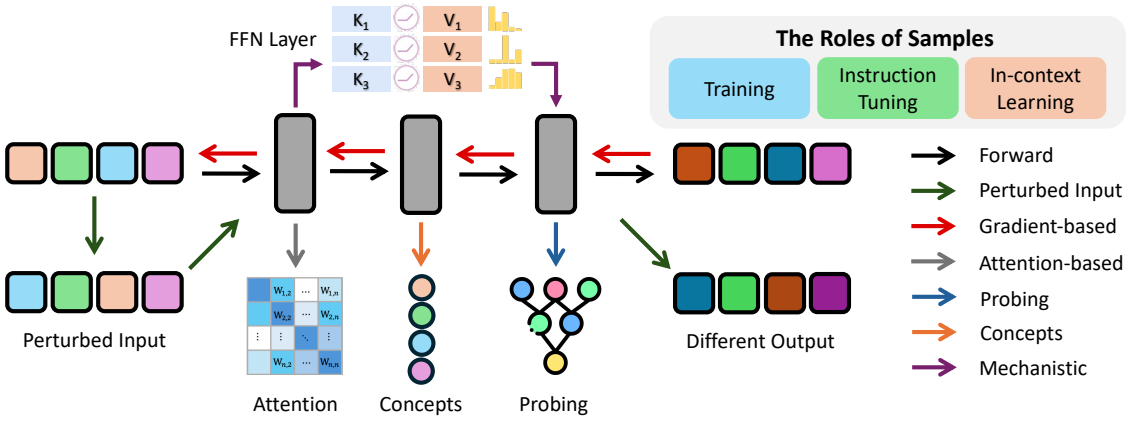

Figure 18: An overview of explainability in foundation models. This figure illustrates various techniques for uncovering how different model inputs and internal components influence model outputs. The right legend highlights the role of samples in different learning stages and the typology of explanation approaches.

10.1 Feature Attribution Methods

When adopting LLMs on downstream tasks, it is important to determine which part of words or tokens in the input contribute most to the prediction. Thus, we need to determine the importance of each part of the input, i.e., explaining the prediction using the raw features. To explore this, there are several important lines of work:

10.1.1 Perturbing the Input for Explanation

To study the effects of the raw features for model prediction, it has been important to perturb part of the input (a piece of text) while monitoring the model output. With this routine, Perturbed Masking (Wu et al., 2020) proposes to perturb a token in the given sentence while monitoring the representation of another token. They further propose span-level perturbation to study the impacts of a certain span within the sentence. While Wu et al. (2020) regard the monitored variable as the representation of the token or span, MICE (Ross et al., 2021) study the roles of inputs for model prediction in classification tasks (i.e., the monitoring variable becomes the model prediction). They present a method to find the edits that could flip the model's prediction, where the edits could serve as contrastive explanations. In addition, perturbing the input to shift the label could create counterfactual examples. Crest (Treviso et al., 2023) proposes a framework to first perturb the input sentence with masks and then edit the masked tokens to obtain counterfactuals. Here, perturbing the sentence with masked tokens is essentially extracting rationales as they are both locating the important tokens for model prediction, though finding the rationales could also be achieved by other methods (Lei

et al., 2016; He et al., 2022). To provide more diverse perturbation types and locations, Polyjuice (Wu et al., 2021) presents a general-purpose counterfactual generator that can generate diverse sets of realistic counterfactuals.

10.1.2 Gradient-based Explanation

Mohebbi et al. (2021) adopt gradient-based attribution methods to provide token-level attribution scores to understand the representation space of BERT (Devlin et al., 2019) better. More advanced difference-from-reference approaches such as Integrated Gradients (IG) are also used to explain the BERT's prediction (Sikdar et al., 2021; Sanyal & Ren, 2021). REAT (Du et al., 2019) decomposes the final prediction of RNNs directly into the additive contribution of each word in the input text. Voita et al. (2021) extend LRP (Montavon et al., 2019) to the Transformers to attribute the relevance score on the source and target contexts in Neural Machine Translation tasks. Wu & Ong (2021) analyze different gradient-based methods for explaining BERT classification results. Recent work has extended gradient-based methods to autoregressive decoder-only language models. For example, Enguehard (2023) introduce Sequential Integrated Gradients, which computes attributions along the generation path by integrating gradients between an empty sequence baseline and the final generated tokens of GPT-2. Kariyappa et al. (2024) propose to approximate token attributions by backpropagating importance through each auto-regressive decoding step, which scales efficiently to long sequence generations.

10.1.3 Attention-based Explanation

Previous works suggest that information could be encoded within the heads of the attention weights (Tenney et al., 2019a), including abundant information (Goldberg, 2019; Voita et al., 2018; Vig & Belinkov, 2019; Raganato & Tiedemann, 2018; Hewitt & Manning, 2019a; Clark et al., 2019a; Zhang et al., 2021b;a), which could be used for both input-level explanation and attention heads pruning (Voita et al., 2019). Multiple tools are proposed to visualize the attention to illustrate the correlations between words for explanation purposes (Vig, 2019; Park et al., 2019; Jaunet et al., 2021; Hoover et al., 2020). Moreover, DeRose et al. (2021) propose Attention Flows to visualize the whole attention flow instead of the visualization of one layer. Some methods combine gradients and attention for explanation (Barkan et al., 2021; Hao et al., 2021), which generally perform better than using attention alone. Abnar & Zuidema (2020) treat self-attention as a flow network across layers to enable post-hoc computation of token-to-token information propagation, which shows higher correlation with gradient-based and ablation-based importance scores compared to raw attention. Though attention scores could be used to understand the large language models, they may not necessarily be capable of identifying the explanations (Jain & Wallace, 2019).

10.2 Exploring the Knowledge in LLMs

Instead of explaining LLMs by highlighting the important tokens or spans in the input, interest increasingly gravitates toward understanding the breadth of knowledge encapsulated by these models. Several key areas of investigation are outlined as follows:

10.2.1 Probing the Representations within LLMs

Probing the model could help us understand deep neural networks (Belinkov, 2022). Early work by Veldhoen et al. (2016) introduced diagnostic classifiers for revealing how neural networks process hierarchical structure by training simple probes on latent representations to test hypotheses about compositional strategies. With such probing philosophies, recent transformer-based works investigate the embeddings and hidden states from various mechanistic components of the network. Kunz & Kuhlmann (2020) show that the token embeddings learned by BERT and ELMo contain rich information about the exact linear context of the token. Belinkov et al. (2017) interpret the representation of different layers in NMT encoders, finding that higher layers have more semantic information. In contrast, lower-layer representations tend to be more suitable for part-of-speech tagging. The fact that language models can capture semantic information and conduct arithmetic operations is also studied in Sorodoc et al. (2020) and Zhou et al. (2024c). Similarly, Clark et al. (2019b); Lin et al. (2019) show that Bert's representations encode surface and positional information in the lower layers,

but more semantic features in higher layers, while Hewitt & Manning (2019b) propose a structural probe showing the syntax trees are embedded in a linear transformation of ELMo and Bert's word representation space. Building on previous probing work, Tenney et al. (2019b) probes word-level contextual representations to investigate how they encode sentence structures. Different from the above methods paying attention to the representation in certain metric spaces (typically Euclidean space), Chen et al. (2021a) consider the probing methods in hyperbolic space, which could better recover tree structures. While these methods could reveal the ability of representations to encode syntactic information, Maudslay & Cotterell (2021) show that syntactic probes may not properly isolate syntax. With a new corpus that is semantically nonsensical but syntactically well-formed, it is shown that syntactic and semantic information are entangled. Further, Zhang et al. (2022) argue that even with the existing works, it remains unclear whether LLMs have understood linguistic knowledge. Thus they probe GPT-3 to show that it has acquired linguistic knowledge in most cases but may still fail when disturbances happen. Apart from exploring the representations, some other works focus on self-attention heads, which could be helpful for heads pruning (Kovaleva et al., 2019; Clark et al., 2019b).

Some methods are designed to be used during the inference of LLMs without training the classifier on the hidden vectors, such as cloze completion or text generation (Petroni et al., 2019; Apidianaki & Soler, 2021; Li et al., 2022b; Ravichander et al., 2020). Though prompts can be designed to reveal the abilities of the LLMs, Zhong et al. (2021) question if the prompt-search methods also learn from the training data, i.e., the training data may contain certain regularities of the underlying fact distribution that could be exploited.

Probing methods are also used to understand the roles of neurons in LLMs. Torroba Hennigen et al. (2020) propose a framework based on a decomposable multivariate Gaussian probe to explore how linguistic information is structured within the representation, showing that most attributes are reliably encoded by only a few neurons. Moreover, some methods propose to probe the internal activations to predict the presence of features in the input, showing the sparse combinations of neurons can represent many features (Gurnee et al., 2023). Recently, OpenAI has shown the possibility of using an advanced LLM (e.g., GPT-4) to explain the neurons in a small model (e.g., GPT-2) (OpenAI, 2023b). Summarize and Score (SASC) (Singh et al., 2023b) proposes to generate candidate explanations to explain the modules from LLMs, which could be more efficient than explaining single neurons. Marks & Tegmark (2023) reveal that LLM representations of true/false statements form distinct linear directions that can be identified via probing and causally intervened upon to flip model outputs. Ji et al. (2025) show that identifying a linear "verbal uncertainty feature" in LLM representations can be manipulated at inference time to reduce hallucinations. Merullo et al. (2025) investigate how pretraining data frequency influences the emergence of linear representations of factual relations and find strong correlations between term co-occurrence counts and probe performance across models.

10.2.2 Explaining LLMs with Concepts

Concept-based explanation refers to mapping the input into concepts and then using a linear classifier to predict the final class with the mapped concepts. As the prediction from the concept to the class is a simple linear classifier, it has the property of explainability even though the mapping from the input to the concepts is not explainable. Pioneering methods in this direction include Concept Activation Vectors (CAVs) (Kim et al., 2018) and Concept Bottleneck Models (Koh et al., 2020). Such a concept-driven framework is widely adopted in visual representation learning where the images are first mapped to the concept space, based on which the classifier makes the decision (Kim et al., 2018; Koh et al., 2020; Yan et al., 2023; Kazmierczak et al., 2023; Chattopadhyay et al., 2023; Zhang et al., 2024j; Huang et al., 2024b). More recently, Wang et al. (2024d) propose Probabilistic Conceptual Explainers (PACEs), drawing inspiration from hierarchical Bayesian deep learning (Wang & Yeung, 2016; 2020; Jordan et al., 1998) and topic models (Blei et al., 2003) to provide concept-based explanations at multiple levels (e.g., datasets, images, and patches) to address key concerns in model interpretation such as faithfulness, stability, and parsimony.

Beyond computer vision, CAVs are also tailored to language models for sentiment classification tasks (Captum, 2022), featuring two concepts: Positive Adjectives and Neutral. Besides, while Captum (2022) define concepts manually, Mu & Andreas (2021) propose to learn the abstractions by analyzing the neurons, where they find that neurons learn shallow lexical heuristics from dataset biases. Wang et al. (2024c) propose Vari-

ational Language Concepts (VALCs) to learn the concept-based explanations in an unsupervised learning manner while enabling neuron editing in the concept space. Turner et al. (2023a) propose to steer the behaviors of language models by curating concept activations and injecting them in the model's hidden layers, and Zou et al. (2023a) propose a unified paradigm for concept interventions in the activation space. Barrault et al. (2024) propose Large Concept Model (LCM) to perform next-sentence-prediction-based autoregressive learning in the conceptual embedding space. In summary, developing concept representation is a crucial step towards interpretable LLMs for diverse tasks. Such interpretability offers a feasible solution for diagnosing, revising, and intervening LLMs.

10.2.3 Mechanistic Interpretability

In each block of a transformer, the self-attention layer projects input tokens or hidden states into query, key, and value vectors that effectively store key–value memories. The main focus of mechanistic interpretability is reverse engineering for retrieving contextual information across those vectors, tokens, and layers, which provides a systematic approach to explaining LLMs (Elhage et al., 2021). The aforementioned study also finds that in-context learning in small models could be explained by specific attention heads, termed "Induction Heads". This mechanism is hypothesized to constitute the mechanism for most "in-context learning" in large transformer models (Olsson et al., 2022). Another line of work focuses on FFN layers. Earlier work (Geva et al., 2020) argue that FFN layers contain most of the information that operates as key-value memories, and more recent works (Yao et al., 2024; Yu et al., 2024a) propose to search neural circuits or salient neurons for parametric knowledge representation. With the localization of the information, we could perform model editing on the relevant matrices in FFN layers (Meng et al., 2022a). In addition, Geva et al. (2022) and Park et al. (2024) analyze the learning dynamics of generative models in the concept space, demonstrating that updates can be decomposed into sub-updates, where each sub-update corresponds to human-interpretable concepts.

10.3 Discovering the Roles of Samples in Training, Fine-tuning, and Few-shot Learning

The development of foundation models encompasses multiple learning stages, including pre-training, fine-tuning, and few-shot learning. During different stages, samples play distinct roles as illustrated in Figure 19.

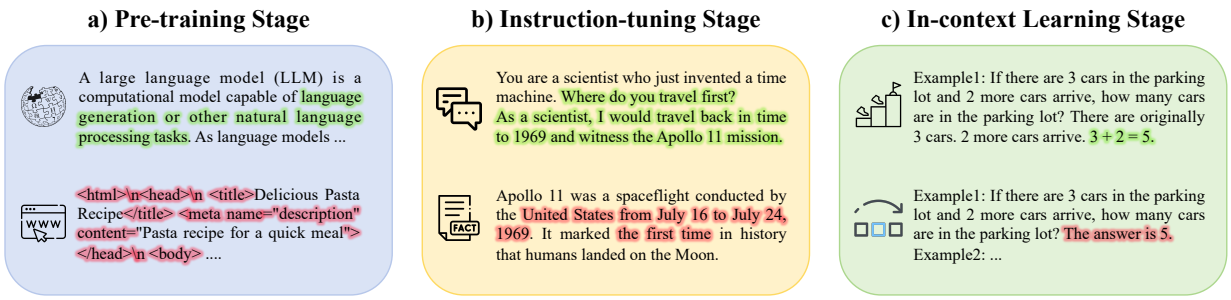

Figure 19: The influence of samples in pre-training, instruction-tuning, and in-context learning stages. We highlight the beneficial and detrimental textual fragments in green and red, respectively.

10.3.1 Influence of Single Example in Training

There is a growing body of work studying the effects of one single example in the training process. SHAP (Shapley et al., 1953) first proposes Shapley values to allocate the contribution of one single player in a coalitional game. TransSHAP (Kokalj et al., 2021) proposes to adapt SHAP to transformers models, Bert specifically, to explain the classification results. Other works measure the effects of the example with the influence of this example on test loss values (Yeh et al., 2018). Influence function, as a statistical technique adapted to deep neural networks by Koh & Liang (2017), approximates how upweighting a single training example would change model parameters and test loss. Recently, Grosse et al. (2023) scale the influence

functions on LLMs with up to 52 billion parameters, and Ruis et al. (2024) leverage influence-based analysis to demonstrate that procedural knowledge in the LLM pretraining sequences drives the emergence of reasoning capabilities.

10.3.2 Influence of Training Stages

The training stages in the current most powerful models include pre-training and instruction tuning. LIMA (Zhou et al., 2023a) analyzes the relative importance of pre-training and instruction-tuning, hypothesizing that the knowledge revealed in the generation primarily comes from the pre-training stage, while instruction-tuning tends to fixate on the style and format of interacting with users, which is tested by using only 1000 curated examples to train Llama-65B to achieve near-GPT-4 performance on a controlled human study. Wu et al. (2023e) explore the instruction recognition and knowledge evolution before and after instruction-tuning, demonstrating that instruction-tuning could better identify the instruction parts from the input and align the knowledge with the user instruction.

10.3.3 Influence of Samples in Few-shot Learning

Few-shot learning in LLMs typically refers to in-context learning (ICL). Li et al. (2023i) investigate ICL's functionality using contrastive demonstration and saliency maps. Wei et al. (2023d) examine how specific examples influence learning outcomes in few-shot scenarios, employing two distinct approaches: ICL with intentionally incorrect labels, and ICL with semantically unrelated labels. They found that large models can better override the input-label mapping learned during the pre-training stage, and small models rely more on semantic priors than large models do. In both settings, they find larger models and those enhanced with In-Context Fine-Tuning perform better. Wu et al. (2023d) focus on how Chain-of-Thoughts (CoT) affects the model behavior, while others try to perturb CoT demonstrations and check the effects on the outcome (Madaan & Yazdanbakhsh, 2022; Wang et al., 2022a). Hahn & Goyal (2023) propose a theory of emergent in-context learning as implicit structure induction to show how compositional structure in pretraining data gives rise to ICL. Xie et al. (2021) provide a complementary perspective by modeling ICL as implicit Bayesian inference where LLMs infer latent concepts shared across prompt examples. From the empirical perspective, Lu et al. (2023) conduct large-scale evaluations of emergent abilities across multiple tasks to conclude that the emergent performance gains can be largely ascribed to ICL mechanisms and parametric knowledge.

10.4 Evaluation of Explainability

The evaluation of explainability usually focuses on two perspectives (Zhao et al., 2023a): (1) Plausibility (also known as persuasiveness by Jacovi & Goldberg, 2020). A plausible explanation seems logical and coherent to the audience, regardless of whether it is correct or accurately reflects the model's reasoning process. Essentially, plausibility measures the quality of the explanation in terms of its persuasiveness and understandability from a human perspective. (2) Faithfulness. A faithful explanation accurately represents the internal workings and decision-making processes of the LLM, demonstrating how well the explanation aligns with what the model is doing when generating the response.

10.4.1 Evaluation of Plausibility

To evaluate the plausibility of the explanations of pre-trained LMs, Shen et al. (2022) propose a benchmark to test LMs abilities in five dimensions: grammar, semantics, knowledge, reasoning, and computation. Another benchmark, HateExplain (Mathew et al., 2021), asks the annotators to highlight part of the text that could justify their decisions, which could serve as the ground-truth explanations. With these ground-truth tokens, we could calculate the metrics such as Accuracy, Macro F1-score, AUROC score (Mathew et al., 2021), and AUPRC (Area Under the Precision-Recall Curve), IOU (Intersection-Over-Union) (DeYoung et al., 2020), etc.

The above metrics could be applied to the explanations with raw features (discussed in Sec 10.1), but they may not be suitable for the explanations based on natural language (Sec 10.2.1) as there would be no ground-truth explanations in this case. To resolve this issue, Chen et al. (2023g) propose to evaluate

the counterfactual simulatability of natural language explanations, i.e., whether humans could predict the model's behavior according to the explanations given by the model. If so, then we say LLMs could explain themselves.

10.4.2 Evaluation of Faithfulness

To evaluate the faithfulness of rationales selected by the model, ERASER (DeYoung et al., 2020) proposes the following metrics: (1) Comprehensiveness, which refers to the probability change of the original predicted class before and after the removal of the predicted rationales, and (2) Sufficiency, which means how much the extracted rationales could support the model to make a prediction. Ideally, the objective is to achieve maximal change of comprehensiveness without compromising the accuracy of predictions when relying solely on the extracted rationales. In addition, TaSc (Chrysostomou & Aletras, 2021) proposes another line of metrics: (1) Decision Flip - Fraction Of Tokens (DFFOT), which measures the fraction of important tokens required to be removed to cause a decision flip. A lower DFFOT indicates a more faithful explanation. (2) Decision Flip - Most Informative Token (DFMIT), where the rate of decision flips caused by removing the most influential tokens is reported for comparison. To further evaluate the faithfulness of the explanations, Liu et al. (2022c) propose a faithfulness violation test, showing that most methods are hindered by the faithfulness violation issue. Although these metrics each have their rationale and applicability, the consistency between these metrics remains questionable. Chan et al. (2022) show that the explanations that achieve the best DFFOT may have the worst Sufficiency score. These metrics are also not suitable for natural language explanations. To solve this issue, for classification tasks, Atanasova et al. (2023) propose two tests: (1) counterfactual input editor for inserting reasons leading to counterfactual predictions; (2) reconstruct inputs from the reasons given by the explanation models and check if they lead to the same predictions. Different from modifying the input and monitoring the output (basically perturbation), REV (Chen et al., 2023b) quantifies the amount of new, label-relevant information in the explanations beyond the information within the input, which can give the measurement without perturbation. For CoT-style explanations, Turpin et al. (2023) find that CoT explanations could be vulnerable towards biasing features in the model inputs, thus being systematically unfaithful. Lanham et al. (2023) monitor how the model predictions change when the input is intervened. They argue that models may produce less faithful reasoning when the models become larger. These researches demonstrate the need for better explanations in the CoT style.

10.5 Applications of Explainability

Gaining the explainability of LLMs has various applications, including diagnosing the model, and improving the model, which can help obtain user trust in the model.

10.5.1 Avoiding Shortcut Learning

Du et al. (2023; 2021) point out that LLMs may rely on shortcut features like data biases, and artifacts to make predictions rather than understanding the meaning, which demonstrates an important challenge in the field of LLMs. Chen & Ji (2022) propose to utilize the explanations revealed by the model to determine if the model is robust or not, as they argue that a robust model should behave consistently between original and adversarial example pairs. Wei et al. (2022b) use chain-of-thought to understand the reasoning process of the model, though the faithfulness needs further exploration (Turpin et al., 2023; Lanham et al., 2023). Li et al. (2024b) design an urban-environment multi-agent simulator based on customizable first-order logic to evaluate the logical reasoning capability of LLMs. These works have identified new challenges in LLMs.

10.5.2 Improving Model Performances

Apart from understanding the model, other works try to improve the model with explanations, which can also help gain user trust. For in-context learning, Lampinen et al. (2022) find that using explanations in the prompts can improve performances, and hand-tuned explanations on a small validation set could even offer substantial improvements. To improve the model's reasoning ability, Nye et al. (2021) find that asking LLMs to emit intermediate computation steps into a "scratchpad" helps with multi-step reasoning tasks. On the other hand, Turpin et al. (2023) demonstrate that chain-of-thought explanations can be systematically

unfaithful. LLM may produce plausible but misleading rationales, which cautions against over-reliance on CoT for performance gains. Besides the training-free prompting, Stacey et al. (2022) supervise the model's attention weights to encourage the model to pay more attention to the words that are present in the explanations, which significantly improves the model performance. In the large language model regime, Mukherjee et al. (2023) propose to train a 13B model with the explanations and reasoning processes provided by GPT4 to improve the reasoning ability of small models. ~~Human feedback can also be incorporated to improve the model.~~ Early works by Hendrycks et al. (2020) show that structured human feedback can steer language model behaviors toward alignment goals. Lee et al. (2022) introduce the XMD framework which shows humans the explanations of model behavior and also updates the model based on the user feedback. Then to improve the model's OOD generalization ability, there is a popular paradigm called Explanation Regularization (ER) which aims to align the model rationales with human-annotated rationales (Liu & Avci, 2019; Rieger et al., 2020; Zaidan & Eisner, 2008; Ghaeini et al., 2019; Huang et al., 2021; Ross et al., 2017; Kennedy et al., 2020), where the effects of ER on OOD generalization is evaluated by ER-TEST (Joshi et al., 2022). As these methods require human-annotated rationales, which might be exhaustive, AMPLIFY (Ma et al., 2023a) proposes to automate the process of rationale generation with the insights from post hoc explanations to provide corrective signals to LLMs. Some other applications include identifying the important instructions to compress the instruction (Yin et al., 2023a). Fernandes et al. (2023) comprehensively summarize a taxonomy of integrating human feedback into natural language generation systems for improving generation quality and explaining model decisions.

10.6 Explainability of MLLMs

Existing works on the explainability of MLLMs primarily focus on CLIP-based image-text alignment models. Early research suggests querying GPT to augment class labels, thereby improving the zero-shot performances of CLIP (Radford et al., 2021). Recent methods utilize CLIP to analyze the composition of images with textual concepts, wherein the concepts are further used for image classification (Chattopadhyay et al., 2024; Yang et al., 2022a) or editing (Chefer et al., 2024; Luo et al., 2025). Such approaches offer additional interpretability (Yan et al., 2023; Yang et al., 2022a) and controllability compared with directly using the CLIP representation for class prediction. Furthermore, Agarwal (2023) investigates the trustworthiness of explanations generated for zero-shot and fine-tuned Vision and Language Models (VLMs), revealing that explanations for zero-shot CLIP classifiers are more faithful than those of the fine-tuned versions. While these works concentrate on image-text alignment models, the explainability of LLM-based image/video understanding models, such as MiniGPT-4 (Zhu et al., 2023c), LLaVA (Liu et al., 2023g), VideoChat (Li et al., 2023d), and LVChat (Wang et al., 2024n), remains underexplored.

10.7 Current Limitations and Future Directions

Despite significant advancements in the field of explainability, several limitations persist that necessitate future attention:

10.7.1 Faithfulness of Raw Features

Current methods, ranging from input perturbation to gradient-based and attention-based techniques, offer explanations for model predictions. However, the faithfulness of these explanations remains questionable. As illustrated in Section 10.4.2, the metrics from different perspectives may vary drastically (Chan et al., 2022). This issue extends to downstream tasks where the model may rely on various biases for predictions (He et al., 2022; Du et al., 2023), potentially compromising their generalization ability if the predictions are not truly faithful.

10.7.2 Understanding How LLMs Store Knowledge

Research focusing on the knowledge stored within LLMs – examining layer representations (Belinkov et al., 2017; Kunz & Kuhlmann, 2020; Rajendran et al., 2024) and analyzing generated content (Liu et al., 2024d; Alivanistos et al., 2022) – has shed some light on the distribution of syntactic versus semantic information across layers. While these works provide the insight that lower layers encode syntactic information and higher

layers possess semantic knowledge, the community is still actively discussing how knowledge is dynamically injected into the model during training and how the injected knowledge is triggered through the inference process. For knowledge injection, Müller-Eberstein et al. (2023) trace representational subspaces during pre-training to capture phases when syntactic subspaces rapidly emerge and disentangle from semantic and reasoning subspaces. Hu et al. (2023a) fit hidden Markov models to training-time metrics (e.g., weight norms, variances) to derive latent states of learning dynamics, which reveal phase transitions during the training. Wal et al. (2025) deploy multiple LLM pre-training runs to demonstrate that training dynamics and knowledge injection stabilize consistently with identifiable outlier behaviors. On the other hand, how to trigger relevant knowledge with the input demonstrations needs further understanding (Liu et al., 2021a; Chen et al., 2023e). Some works argue that the knowledge is mainly stored in MLP layers (Meng et al., 2022a;b); however, it is shown in their papers that attention layers also have slight effects when predicting the facts, especially in earlier layers (See Figure 3 in (Meng et al., 2022b)). Even in the work that explicitly stores knowledge in a memory module (Wang et al., 2024m), how the model processes the knowledge is under-explored.

10.7.3 Transferability of Explanation Across Different Modalities

How to develop unified explanation frameworks that can inherently transfer across modalities has been an emerging topic. As we have discussed above, prior works on Integrated Gradients (Sikdar et al., 2021; Sanyal & Ren, 2021; Enguehard, 2023) are modality-agnostic and can be employed on image classification, text generation, and multimodal VQA tasks. Yet these gradient-based attributions suffer from superficial faithfulness and low robustness to input perturbations (Adebayo et al., 2018). On another track, attention-based methods such as MultiViz (Liang et al., 2022a) can provide visualizations of cross-modal attention and thus enable analysis of pixel-to-token flow in multimodal transformers. However, the credibility of attention as explanation has long been arguable in the community (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Concept interpretation tools such as representation engineering for LLMs (Zou et al., 2023a) have seen the potential to be adapted to vision-language modalities (Liu et al., 2024f). Aggregating multiple explanation levels and addressing the aforementioned limitations will motivate future work in developing robust, faithful, and transferable explanations for multiple modalities. It is also increasingly necessary to develop standardized benchmarks for cross-modal explanations that promote both interpretability and fidelity.

10.7.4 Reliability and Responsibility of Foundation Models from the Explainability Perspective

Without a deep comprehension of foundation models, ensuring their reliability and responsibility is challenging, where explainability has the potential to offer a pathway to address these issues. For instance, identifying the biases in pertaining data and implementing various de-biasing strategies could pose more equitable models (Li et al., 2023h). Moreover, understanding how the model stores knowledge (Liang et al., 2024) can facilitate model editing with the current knowledge and the unlearning of harmful information, achieving up-to-date and safer foundation models (Meng et al., 2022a; Wang et al., 2024e; Zhang et al., 2023a). However, the effectiveness of interpretability methods themselves must be critically assessed to ensure they provide meaningful insights. Adebayo et al. (2018) perform sanity checks for saliency maps and reveal that some widely used saliency methods are independent of both the model and the data, questioning their validity in explaining model behavior. Similarly, Alvarez-Melis & Jaakkola (2018) investigates the robustness of interpretability methods and demonstrates that small perturbations to the input can significantly alter the explanations provided, highlighting the need for more robust interpretability techniques. Moreover, understanding how practitioners use interpretability tools is also crucial. Kaur et al. (2020) explore data scientists' use of interpretability tools and find that mismatches between tool capabilities and user needs can limit their effectiveness in ensuring model reliability and responsibility. They emphasize the importance of designing interpretability tools that align with the practical requirements of users. As the development of increasingly powerful foundation models continues, focusing on both the advancement and the critical evaluation of explainability methods cannot be overstated.

11 AIGC Detection

The advent of foundation models has led to a surge of artificial intelligence-generated content (AIGC) across various modalities, including text (Team et al., 2023; OpenAI, 2023b; Team et al., 2024), images (Ramesh et al., 2021; Zhang et al., 2023d; Esser et al., 2024), audio (Kreuk et al., 2022; Guo et al., 2023c; Huang et al., 2023b; Anastassiou et al., 2024), and video (Kondratyuk et al., 2023; Blattmann et al., 2023; Bar-Tal et al., 2024). While these technologies have unlocked many useful applications, they also pose significant challenges, particularly in terms of content authenticity (Gu, 2024; Li et al., 2024i; Hong & Zhang, 2024). The capacity of foundation models to generate human-like content can be exploited for malicious purposes, including the dissemination of misinformation and identity theft. Consequently, the demand for research focusing on detecting AIGC is on the rise. This section provides a comprehensive overview of current methodologies and techniques for AIGC detection, highlighting the pivotal role this field plays in preserving the integrity of digital information in an era increasingly dominated by foundation models and AI technologies.

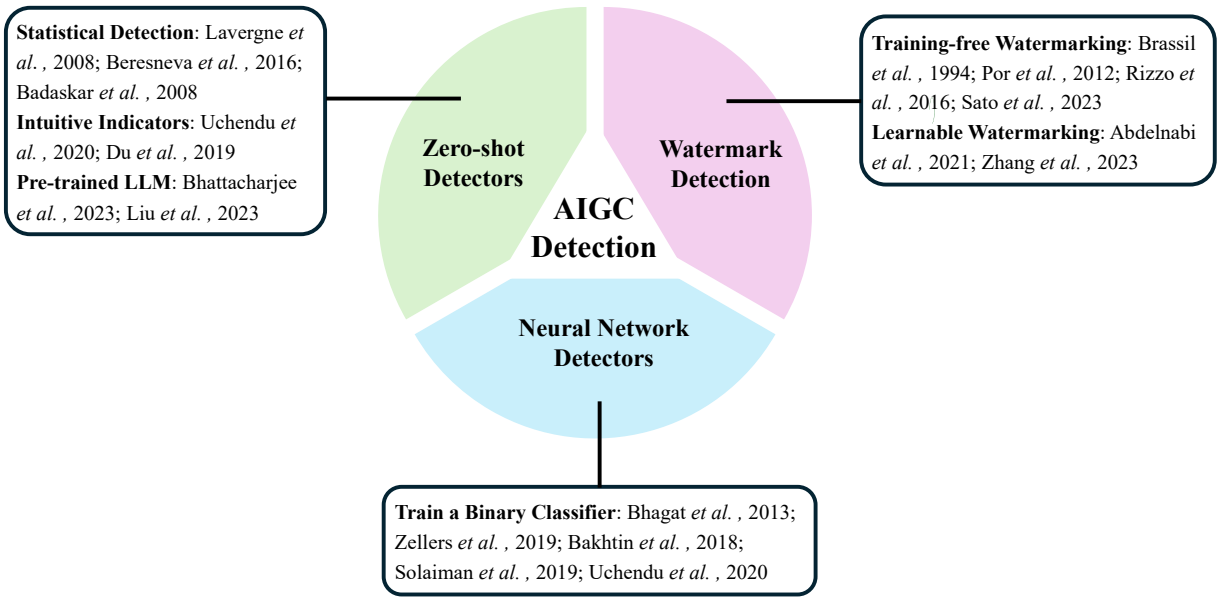

Figure 20: An overview of AIGC detection techniques. We group them into three categories: zero-shot detectors, watermark-based detection, and neural network detectors, each with further subdivisions.

11.1 The AIGC Detection Problem

The task of AIGC detection can be seen as a binary classification problem. In general, we aim to determine whether a given input $x \in \mathcal{X}$, such as an image, text, or audio, is generated by AI models. This can be achieved using a detector $D : \mathcal{X} \to \{0, 1\}$, which can be defined as follows:

$$D(x) = \begin{cases} 1 & \text{if } x \text{ is generated by AI or is partially generated by AI.} \\ 0 & \text{if } x \text{ is created by a human.} \end{cases} \tag{14}$$

The detector D can be broadly categorized into the following types: (i) zero-shot detectors, (ii) watermark detectors, and (iii) learnable detectors. Furthermore, we summarize the representative work for all types in Figure 20, with examples for each type illustrated in Figure 21.

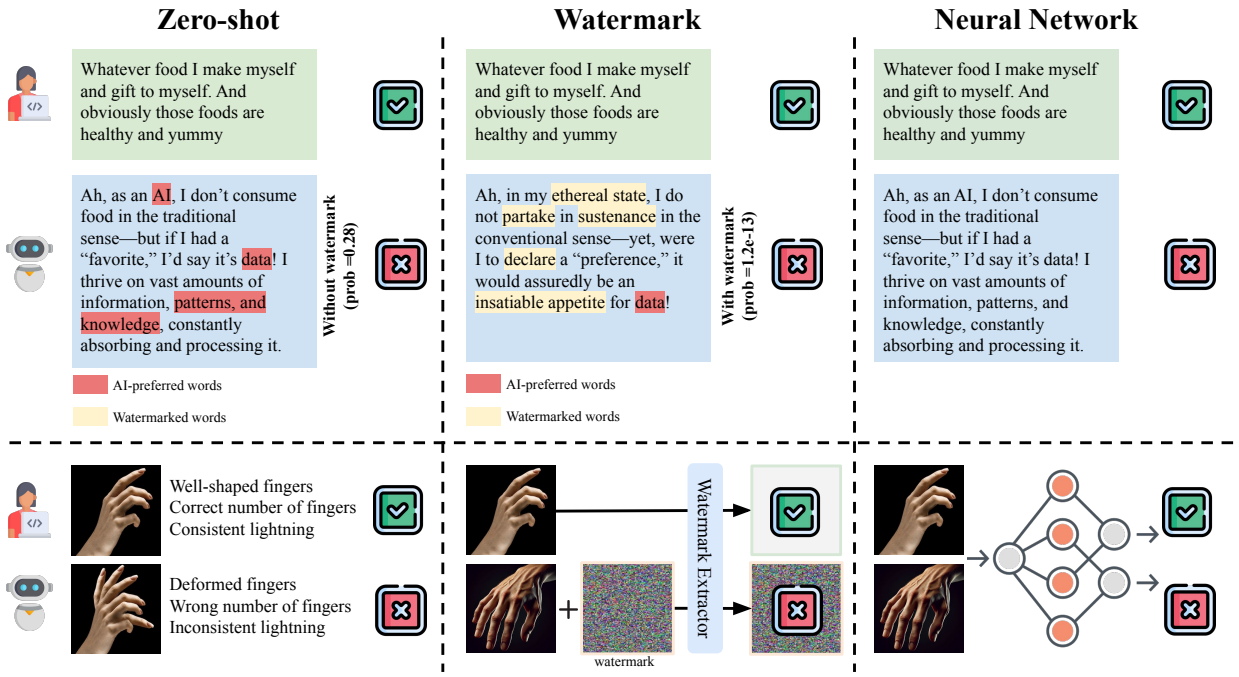

Figure 21: Examples of zero-shot, watermark, and neural network detectors for textual and visual inputs.

11.2 Zero-shot Detectors

The fundamental concept behind zero-shot detectors is to differentiate between AI- and human-generated content based on their intrinsic distinctions, such as the frequency of word occurrence in the generated text, which can be identified and flagged by hand-crafted detectors. That said, zero-shot detectors are arguably the simplest to deploy since they do not require additional training of both the detectors and the foundation models that generate the content.

11.2.1 Statistical Detection

These detectors assume full, or at least partial (e.g., the token logits during generation), access to the foundation model that generated the content. In the text domain, traditional methods usually rely on statistical outlier detection based on different metrics, including entropy (Lavergne et al., 2008), perplexity (Beresneva, 2016), n-gram frequencies (Badaskar et al., 2008), the ratio of perplexity to cross-perplexity (Hans et al., 2024a), which measures how surprising the next token predictions of one model are to another model, and average per-token log probability (Solaiman et al., 2019). We use them to evaluate the given text passage and apply thresholding to assess whether the content is likely AI-generated. However, these approaches are inadequate in the era of foundation models, where AI-generated content becomes more diverse and of high quality.

To this end, several recent studies improve upon these simple ideas and extend them to LLMs. (Gehrmann et al., 2019) propose GLTR, which is centered on the underlying assumption that LLMs overgenerate from a limited subset of the true distribution of natural language, for which they have high confidence. This property is detected by computing, for each token in a text sequence: (i) the probability of generating the token, (ii) the rank of the word, and (iii) the entropy of the generated distribution. These metrics are then compared against those of human writers. In a similar vein, DetectGPT (Mitchell et al., 2023) leverages the empirical observation that AI-generated text tends to lie in negative curvature of the model's log probability function, i.e., the sequence sits at a locally concave region of $f(x) = \log p_\theta(x)$ such that small random paraphrases/perturbations $y \sim q(\cdot|x)$ systematically *decrease* f. Practically, DetectGPT scores

a text by the average log-likelihood drop $\Delta = f(x) - \mathbb{E}_{y\sim q(\cdot|x)}[f(y)]$. A second-order Taylor expansion gives $\mathbb{E}[f(y)] \approx f(x) + \frac{1}{2}\mathrm{tr}\big(H_f(x)\,\mathrm{Cov}_q\big)$, so $\Delta > 0$ implies $\mathrm{tr}\big(H_f(x)\,\mathrm{Cov}_q\big) < 0$ (negative curvature) along the perturbation directions—an effect pronounced for model-generated text but weaker/inconsistent for human-written text. This observation led to follow-up investigations on improving detection efficiency (Deng et al., 2023b) and utilizing conditional probability curvature (Bao et al., 2023). DetectLLM (Su et al., 2023a) employs a similar principle, but scores with log-rank information. However, these approaches rely on thresholding the probability of a given sequence, which requires access to the model's token generation probability distribution. Such a requirement can be too restrictive in many practical scenarios.

To alleviate this, recent detection methods that require only API-level access to the unknown source model are proposed. For instance, (Yang et al., 2023c) utilize the N-Grad divergence between re-prompted and original text to identify AI-generated content in the biology domain. Additionally, recent research has shown that smaller surrogate models can serve as effective proxies for AIGC detection (Mireshghallah et al., 2023; Yang et al., 2023d; Cozzolino et al., 2025). By observing that AI-generated text exhibits lower intrinsic dimensionality compared to human-written text when measured in a representation space of fixed text embeddings (e.g., sentence- or token-level vectors produced by a pretrained encoder), (Tulchinskii et al., 2023) propose to employ a persistence-homology-based intrinsic dimension estimator (PHD) to exploit this property for AIGC detection, estimating local manifold dimension from neighborhoods within the embedding space. This approach does not require API-level access—i.e., it operates in a complete black-box setting.

11.2.2 Intuitive Indicators

These methods use the analytical abilities of humans to identify inconsistencies with prior knowledge in AIGC, thus achieving detection. As a result, these methods provide notable interpretability and credibility in the detection process.

For AI-generated text, (Uchendu et al., 2023) note that a lack of coherence and consistency serves as a strong indicator of AIGC, and emphasize the importance of collaboration among human detectors in improving detection accuracy. Similarly, (Dugan et al., 2022) note the unreliability of relying solely on grammatical errors as a detection strategy. They further showcase that while LLMs frequently commit factual and logical errors, these mistakes are often overlooked by neural network-based detectors but are easily noticed by human detectors. More recently, (Mao et al., 2024) find that LLMs exhibit a greater propensity to alter human-written text compared to AI-generated text when tasked with rewriting. This tendency stems from LLMs' perception of AI-generated text as being of high quality, which results in fewer modifications. They then proposed "geneRative AI Detection viA Rewriting" (RAIDAR) to detect AI-generated content by instructing LLMs to rewrite text and then calculating the edit distance of the output by the Levenshtein Score (Levenshtein et al., 1966).

In vision, the detection of AI-generated images can be done by examining inconsistency with reality. Numerous studies (Borji, 2023; Farid, 2022a) note that AI-generated images often violate physical rules in the real world, such as missing or unnatural reflections and shadows of objects that are inconsistent with natural lighting and environment. In addition, (Farid, 2022b) has noticed that AI-generated images exhibit inconsistency in perspective, such as parallel lines cannot converge at a common vanishing point. For facial images, (Borji, 2023) outlines key cues for detecting AI-generated faces, including symmetry, iris color, pupil shapes, skin, etc., where the generated images tend to depict physiological falsehood.

However, AIGC detection by intuitive indicators are becoming much harder as the capabilities of AIGC models continually improve.

11.2.3 Pre-trained LLMs

Without training, a few studies have investigated the use of pre-trained LLMs to directly identify generated texts either by themselves or by other LLMs. However, it has been observed that the performance of these detection methods is often inferior to statistical and neural network approaches. For example, (Bhattacharjee & Liu, 2024; Liu et al., 2023o) formulate the AIGC detection task in a question-and-answer format, and prompt LLMs with the question to obtain an answer for detection. (Bhattacharjee & Liu, 2024) note that

neither ChatGPT nor GPT-4 could reliably identify text generated by various LLMs, while (Liu et al., 2023o) reveal the poor zero-shot performance of GPT-3.5-turbo in AIGC detection which is close to random guessing.

A recent work (Koike et al., 2023) considers employing in-context-learning (ICL) with pre-trained LLMs for AIGC detection, in which a few labeled examples (context) are integrated into the question prompt as a single input to the model, thereby facilitating the learning of new tasks in context. The results in (Koike et al., 2023) show that using ICL outperforms both traditional zero-shot methods and RoBERTa-based detectors, however, (Liu et al., 2023o) observe no significant improvement in using ICL with GPT-3.5-turbo. It is worth noting that while ICL methods are not strictly zero-shot, they do not require additional training of the detectors. Another recent work (Krishna et al., 2023) proposes a detection mechanism based on retrieval, which involves creating a database of generated text and comparing the semantic similarity of the target text with all the text stored in the database to perform detection. Although this approach is effective and robust against paraphrasing, its requirement of storing LLMs generation may raise privacy concerns.

11.3 Watermark-based Detection

Watermarking injects algorithmically detectable patterns into the AI-generated content while ideally preserving the quality and diversity of AIGC. A watermarking algorithm for AI-generated content detection typically involves three components:

- The *watermark* or message, denoted as m, can be represented as a bit-string in the generated images or as a specific occurrence of words in the generated text. From now on, the term "watermark payload" will be used to refer to the amount of information conveyed by the watermark message.
- An *encoder*, denoted as A, is responsible for embedding the watermark message m into an AI-generated content x, thereby transforming it into a watermarked content $\tilde{x}$.
- A *detector*, denoted as D, is capable of determining the presence of a watermark in either $\tilde{x}$ or x, provided that the content is generated by AI.

In *zero-bit watermarking*, the embedded message m only signifies the presence or absence of a watermark, hence is only used to indicate whether x is generated by AI; whereas in *multi-bit watermarking*, the embedded message m can carry additional detailed, customized information, e.g., the name of the AI model or authorship attribution. We will primarily focus on the first case - using watermarking for AIGC detection.

A watermarking algorithm that is effective for detecting AI-generated content should possess the following key properties:

- It should be algorithmically easy to verify yet remain imperceptible to humans, where ease of verification can refer to the ability to open-sourcing, or a high success rate for detection.
- It should have minimal impact on the quality of AI-generated content. This means that foundation models incorporating the watermark algorithm, potentially during training, should still produce content of similar quality compared to the non-watermarked version.
- It should exhibit high robustness to attacks aimed at removing the watermark or applying semantically invariant transformations to AI-generated content with watermarks. These transformations can range from rephrasing generated text to distorting watermarked images.
- It should demand minimal effort to incorporate the watermark into AI-generated content.

11.3.1 Training-free Watermarking

In training-free watermarking algorithms, the watermark, encoding, and decoding algorithms are all designed based on heuristics, exploiting domain-specific characteristics of the generated content rather than learned through end-to-end training.

Several studies apply various kinds of semantically-invariant transformation directly to *existing* AI-generated text. These include visually imperceptible reformatting such as adding whitespace characters and replacing characters with similar ones in appearance but with a different Unicode representation (Brassil et al., 1994; Por et al., 2012; Rizzo et al., 2016; Sato et al., 2023); lexical-based modifications such as synonym substitution (Munyer & Zhong, 2023; Topkara et al., 2006b; Yang et al., 2023b; Yoo et al., 2023a; Yang et al., 2021b); syntax-based manipulation which alters the arrangement of words and phrases in the text through several predefined types of transformations (Atallah et al., 2001; Meral et al., 2009; Topkara et al., 2006a). Each distinct type of transformation corresponds to a specific message bit, therefore allowing the detection and extraction of watermarks. The immediate advantage of these approaches is that they do not require knowing the identity (i.e., the name of the model) or access to the AI models that generated the content. However, since these methods largely rely on simple semantically invariant transformation, they are easy to spot and hence are vulnerable to watermark attack or removal. Moreover, these manually defined modifications can create abrupt and unnatural modifications to the original text, hence significantly degrading the quality of the generated content.

Instead of encoding watermarks in the existing context *after* generation, it is also possible to encode training-free-based watermarks *during* the content generation process without the need for re-training the models. Consequently, unlike previous approaches discussed, the following methods assume at least the given access to controlling the generation process of the foundation models. The pioneering research of (Kirchenbauer et al., 2023a) first proposes a watermarking framework for LLMs by altering the *logits* for token sampling in a text sequence generation. The algorithm (Kirchenbauer et al., 2023a) works by selecting a randomized set of “green" tokens before generation, and then softly promoting the use of “green" tokens during generation by adding a small bias on the sampling logits of “green" tokens. Detection can be achieved by deploying statistical tests which are essentially based on identifying the unnatural occurrence of “green" tokens in the writing. Follow-up research works expand upon this idea in the directions of preserving quality and semantic meaning of generated content in low-entropy text generation scenarios (Lee et al., 2023b; Wang et al., 2023c), where text quality is vulnerable to such tiny bias towards generating randomly selected “green" tokens; multi-bit watermarking (Yoo et al., 2023b; Fernandez et al., 2023a; Qu et al., 2024); improving the robustness of watermarking against removal attack and post-processing (Kirchenbauer et al., 2023b; Ren et al., 2023a; Zhao et al., 2023c; An et al., 2024); and defending against forgeries of watermarks (Hu et al., 2023c; Wu et al., 2023f). In contrast altering the logits, a line of works (Hou et al., 2023a; Kuditipudi et al., 2023; Christ et al., 2023) alternatively choose to manipulate the *token sampling* process itself directly by encoding a watermark in a pseudo-random number sequence as seeds to guide the sampling of each token or sentence in a text generation sequence. Detection therefore needs to access the correspondence between the tokens generated and the underlying pseudo-random numbers.

Beyond text generation, training-free watermarks have also been applied to AI-generated images. For instance, DaLL · E (Ramesh et al., 2021) always prints a tiny visible color pattern at the bottom right corner of its generated images. To better preserve the visual quality of the generated images, invisible-watermark (Wang, 2020), which is adopted by Stable Diffusion, encodes bits of the watermark message through modifying coefficients of a carefully selected subset of band frequencies of its generated images under discrete wavelet transforms. Detection and decoding of the watermark is thereon achieved through an inverse transformation. In addition, (Wen et al., 2023b) introduce a training-free watermark for diffusion models by embedding watermark signals into the initial latent noise, creating a semantic watermark.

11.3.2 Learnable Watermarking

Although training-free watermarking and detection techniques are straightforward in concept and require minimal effort to deploy, the pre-defined watermarking rules may be too conspicuous, leading to a compromise in the quality of the generated content or making them susceptible to watermark removal and forgery. In this survey, we use “learnable watermarking” to refer to methods that *modify the generation process* to encode a keyed *watermark payload* at training or inference time, and whose verification requires the corresponding key (or a public verifier). This distinguishes them from post-hoc detectors in Section 11.4, which do not assume any embedded signal. To address this issue, a couple of studies (Abdelnabi & Fritz, 2020; Zhang et al., 2023f) propose using learning-based watermark encoding and decoding modules, in which

the training pipeline involves an encoder that first embeds a binary "watermark payload" into the original text followed by decoding for the message from the watermarked text. To preserve coherence and consistency of the generated content, the modules from (Abdelnabi & Fritz, 2020) are trained against an adversary that performs a classification between the original and watermarked text, whereas (Zhang et al., 2023f) regularize the watermarked message by penalizing semantic difference with the original text. (Liu et al., 2023b) embed watermarks into text by adding extra watermark logits to the LLM's sampling logits at each generation step, following (Kirchenbauer et al., 2023a). To ensure both attack robustness and security robustness, each watermark logit is determined by applying a learned transformation (a trained watermark model) on the semantic embedding of all preceding tokens generated using another pre-trained LLM. Two similarity loss and normalization loss are minimized during training to prompt semantic consistency and unbiasedness in the generated watermark logits and facilitate statistical detection. Moreover, in a recent work, (Liu et al., 2023a) propose an unforgettable publicly verifiable watermark algorithm utilizing two different neural networks for watermark generation and detection, thereby preventing exposing key information in the watermark generation phase when made accessible for public detection. Furthermore, the token embedding parameters are shared between the generation and detection networks which improves both training efficiency and detection accuracy. (Yu et al.) proposed SAEMARK, a user-specific watermarking method that embeds personalized signatures without altering logits. SAEMARK uses Sparse Autoencoder (SAE) to extract features from generated texts and selects outputs by matching key-derived feature distributions. Boundary to Section 11.4. Although methods such as ASI and publicly verifiable schemes train neural *verifiers*, we keep them in Learnable Watermarking because they *require* a payload that was intentionally embedded at generation time. By contrast, Section 11.4 covers detectors that operate *without* any embedded watermark or key, treating detection purely as post-hoc content classification.

11.4 Neural Network Detectors

Unlike Section 11.3 (learnable watermarking), approaches in this section do *not* modify the generator and do *not* assume any embedded payload/key. They train post-hoc classifiers—often on human vs. AI corpora—and can operate in black-box settings against unknown generators. Consequently, any methods whose verification relies on a generation-time watermark remain in Section 11.3 rather than here. Another line of work approaches the AIGC detection problem by training a binary classifier using labeled training samples containing both human and AI-generated content. Earlier work focuses on fake review (Bhagat & Hovy, 2013), fake news (Zellers et al., 2019), fake images (Ma et al., 2023c), or small AI models detection (Solaiman et al., 2019; Bakhtin et al., 2019; Uchendu et al., 2020). Subsequently, growing interest in this line of research turns to detecting high-quality content brought by foundation models. Detectors under this category do not require access to model parameters hence can operate under complete black-box settings.

Targeting the problem of machine-generated text detection, numerous studies (Chen et al., 2023j; Guo et al., 2023a; Zhan et al., 2023; Tian, 2023; Yu et al., 2024d) fine-tune a pre-trained LLM, such as T5 (Raffel et al., 2020) or RoBERTa (Liu et al., 2019), on a dataset of pairs of human-written text and AI-written text from mixed sources as a simple solution. Alternatively, several works also consider training a classifier on top of a frozen pre-trained LLM (Chen et al., 2023j; Guo et al., 2023a; Wu et al., 2023a; Verma et al., 2023). In particular, (Chen et al., 2023j; Guo et al., 2023a) have attempted training a logistic regression classifier on text embedding obtained using a pre-trained LLM for detection, however, they find such a method often underperforms the fine-tuning approach. (Wu et al., 2023a) propose LLMDet, which conducts binary classification utilizing a proxy score for perplexity, while (Verma et al., 2023) propose Ghostbuster, which is inspired by statistical detection methods based on analyzing token log-probabilities. Both methods train a logistic regression classifier on top of these selected and hand-crafted features to detect machine-generated text, therefore, no longer requiring direct access to the model token sampling logits, as in their zero-shot counterparts, at test time. Recognizing the similarities between the original AI-generated and the regenerated text produced with ChatGPT, (Yu et al., 2023c) introduce a novel GPT Paternity Test for AI-generated text detection. This method involves utilizing ChatGPT to infer a question based on the input text being examined, followed by supplying a response. Subsequently, a Siamese network (Koch et al., 2015) is trained to assess the similarity between the original and regenerated text, aiding the detection using another trained binary classifier.

One major challenge in training a reliable binary classifier is data scarcity as collecting sufficient data to train the classifier can be challenging, especially in diverse domains where the availability of training samples is a major bottleneck. To alleviate this, (Liu et al., 2023m) consider adopting contrastive learning approaches in addition to the supervised training for detection. Another significant challenge involves tackling paraphrasing attacks (Sadasivan et al., 2023; Krishna et al., 2023). To mitigate this problem, (Hu et al., 2023b) propose to employ an adversarial learning approach to simultaneously train a detector and a paraphraser. Nevertheless, supervised training of a binary classifier tends to overfit their training data, resulting in a decline in performance when faced with cross-domain or unseen data. Additionally, fine-tuning LLM classifiers is limited in facing data generated by different models.

11.5 Current Limitations and Future Directions

Despite significant advancements in the domain of AIGC detection, several limitations still require future attention:

11.5.1 Fairness of AIGC Detection

Although state-of-the-art text detectors generally achieve high accuracy in experimental settings, as discussed by (Liang et al., 2023b), perplexity-based text detectors exhibit a notable bias against text written by non-native speakers. Specifically, these detectors have been observed to misclassify TOEFL essays written by foreign writers more frequently than those by native speakers. This discrepancy may be due to the lower perplexity of non-native essays, which often display less linguistic diversity and richness. This issue may also affect minority languages, which tend to have higher perplexities compared to popular languages like English. Additionally, similar biases might exist in other modalities, such as image detection. Therefore, it is crucial to consider the fairness of detectors when designing future detection methods and to develop efficient methods for evaluating the fairness of AIGC detection methods.

Meanwhile, on the watermarking side, learnable watermarking methods might also exhibit biases toward out-of-distribution data points. For instance, if the watermarking encoder and decoder are trained on English text written by native speakers, the model might also have a higher misclassification rate on essays written by non-native speakers. Therefore, it is crucial to consider fairness in the development of learnable watermarking methods as well.

11.5.2 Robustness of Watermarks

Both text and image watermarks are susceptible to regeneration or post-processing attacks, such as paraphrasing (Kirchenbauer et al., 2023b) or diffusion purification (Zhao et al., 2023c). In contrast, semantic watermarks tend to be more robust against such attacks. However, because semantic watermarks typically require deep neural networks to decode the watermark signals, they are vulnerable to adversarial attacks (Saberi et al., 2023; An et al., 2024). Adversarial perturbations can also be developed to prevent regeneration and post-processing attacks (Liu et al., 2022a). Adversarial attacks remain a significant challenge even for classification tasks. Therefore, designing robust watermarks that can withstand both attacks is challenging and crucial.

11.5.3 Origin Attribution of Generated Images

Recent advancements in visual generative models have significantly improved the quality of generated images, raising concerns about their potential misuse. It is critical to develop methods to accurately identify the origin model responsible for generating a given image (Liu et al., 2022a). Especially, the scenarios are especially important and practical where access to the source model is restricted and only a limited number of images from the source model are available (Liu et al., 2022a).

12 Intersection and Conclusion

In this survey, we comprehensively examine the reliability and responsibility of foundation models, spanning technical and societal considerations. Given our overview of the current research landscape, we identify significant prior research that makes progress on these critical issues. However, outstanding challenges limit the extent to which current models are reliable or responsibly developed, indicating more research is needed as this technology has a broader societal impact.

Moreover, greater attention should be paid to the intersections between different research areas, as the areas covered by this survey are interconnected and influence each other. Instead of addressing challenges in isolation, we advocate for a more holistic approach to ensure the overall reliability and responsible development of foundation models. In conclusion, we emphasize a number of key points of intersection across these domains, highlighting the challenges at these crossroads and outlining potential directions for future research.

12.1 Bias, Fairness, and Security

To improve the security of foundation models, adversarial training (Madry et al., 2017; Bai et al., 2021) is often leveraged, which aims to make models resistant to malicious manipulations by conducting bi-level adversarial games (Tao et al., 2021; Ma et al., 2024a) during training. However, while adversarial training improves robustness, it can unintentionally exacerbate fairness issues (Xu et al., 2021a). For example, robust models may focus on defending against a certain set of features without considering the fact that some demographic groups may be penalized by these features more than others, leading to disproportionate performance degradation for underrepresented or marginalized groups. For instance, (Sap et al., 2019) found that hate speech detection models were biased against African American English (AAE), disproportionately misclassifying non-offensive AAE utterances as hate speech.

On the other hand, security risks such as data poisoning, where malicious attackers corrupt training data, can introduce new biases into foundation models or exacerbate existing biases (Mehrabi et al., 2021b; Guo et al., 2022). Poisoned data can skew a model's learned representation, posing the dual challenge of protecting models from data poisoning while ensuring that the fairness of the model is not compromised, especially when training on large, uncurated datasets.

12.2 Bias, Fairness, and AI-generated Content Detection

Large multimodal and text-to-image models are often biased due to unbalanced or stereotype-laden training data, such as images that reinforce harmful stereotypes or misrepresent certain social or cultural groups (Zhang et al., 2023i; Cho et al., 2023; Luccioni et al., 2023). Detecting biased AI-generated content (AIGC) poses a significant challenge, as detection systems themselves can inherit or amplify biases found in the training data. For instance, zero-shot or neural network detectors might disproportionately mark content created by minority groups as *AI-generated* based on biased data patterns in the training set. It is important to ensure the fairness of these detection capabilities, as biased detection could result in unfair discrimination, such as unfairly enforcing restrictions on content produced by minority groups.

Similarly, watermark-based detectors can raise fairness concerns if watermarks are inconsistently applied across different types of content. For example, if T2I models excessively generate watermarked content associated with certain groups (e.g., images associated with a particular ethnic or gender identity), this content could be more easily flagged or blocked, suppressing content created or represented by those groups.

12.3 Security and Privacy

Privacy concerns often go hand-in-hand with issues of security. For example, models ought not to reveal users' private information, which can be found in pre-training data as well as in interactions with users. Furthermore, certain legal jurisdictions already impose privacy-related laws that impact technologies including foundation models. For example, the EU's GDPR covers a "Right to be Forgotten", which mandates that users have the ability to delete their private information (Zhang et al., 2023a); non-compliance with these regulations could result in legal penalties.

Several methods exist for preventing models from revealing private information, including unlearning (e.g., removing concepts from the model's parametric knowledge) (Liu et al., 2024h; Yao et al., 2023b; Jang et al., 2022; Wu et al., 2025) and introducing prompts (Edemacu & Wu, 2024) or RLHF (Xiao et al., 2023a) to prevent models from revealing information. However, jailbreaking attacks (see Section 5.1.1) can often circumvent prompts and RLHF, and (Patil et al., 2024) show that model editing is also vulnerable to attack, finding that sensitive information could be recovered from models even after deletion when querying multiple times. Thus, making models compliant with existing and future legislation regarding private information is an open challenge, as is robustly defending against adversaries attempting to extract private information.

12.4 Security and AI-generated Content

AIGC detection is naturally framed as an adversarial task, with an attacker attempting to pass AIGC as real content, and a defender attempting to detect it. In such scenarios, the advantage usually lies with the attacker, who can make multiple attempts to test existing defenses. Existing work shows that the robustness of current detection methods is imperfect and they are vulnerable to adversarial attacks (Saberi et al., 2023; An et al., 2024; Kirchenbauer et al., 2023b; Zhao et al., 2023c); improving these detection algorithms remains an area of continuous future work.

Beyond the security of individual watermarking and detection methods, AIGC raises broader questions of *societal* security, i.e. the potential threats that AIGC poses to both public institutions and individuals. Here, extensive documentation exists regarding ongoing threats from AIGC. In the political sphere, AIGC has been employed to disseminate misinformation and erode public trust in political systems and elections (Dmonte et al., 2024; Jingnan, 2024), where AIGC has been used to spread misinformation and sow distrust. Similarly, in public health settings, AIGC has been utilized to generate and spread health-related misinformation (Menz et al., 2024). Other voices have also called attention to the risks associated with AIGC's interactions with individuals. For example, (Greenfield & Bhavnani, 2023) raise the concern that AIGC could harm mental health through hyper-personalization. Moreover, individuals may become more susceptible to personalized fraud attempts, such as voice cloning or sophisticated phishing-style attacks (Begou et al., 2023; Eze & Shamir, 2024; Bunn, 2024; Chasan, 2023). To counteract these malicious use-cases, larger-scale safeguards will likely be needed. This includes implementing public education initiatives to raise awareness about the risks of AIGC and developing strategies to combat its misuse across various domains.

12.5 Uncertainty and Alignment

Given the importance of correctly expressing model uncertainty (as described in Section 8), a growing area of interest is in aligning models to accurately predict their uncertainty (Mielke et al., 2022; Stengel-Eskin et al., 2024) or to abstain from answering in cases of uncertainty (Wen et al., 2024a). This work builds on past work finding that models' internal states often contain meta-knowledge about whether the model can correctly respond to a particular prompt (Kadavath et al., 2022; Mielke et al., 2022; Liu et al., 2023j). Several past efforts have explored aligning models to predict this information based on internal states. (Mielke et al., 2022) train models to express uncertainty linguistically by extracting control codes from their internal states and using them to adjust the model's output. (Ulmer et al., 2024a) train a smaller LLM to predict the uncertainty of a larger LLM. (Stengel-Eskin et al., 2024) use a speaker-listener framework to supervise models, rewarding a generator or speaker model for getting a listener to accept correct answers and reject incorrect ones, while penalizing it for doing the opposite outcomes. Future research directions in this field include addressing uncertainty *not* reflected by the model's internal state, that is, the "unknown unknowns" where the model is unaware of its own ignorance, often when faced with out-of-distribution data or novel concepts, and enhancing models' capacity to resolve uncertainty effectively.

12.6 Hallucination, Uncertainty, Distribution Shift, and Alignment

Uncertainty and distribution shifts are deeply interconnected with hallucinations in foundation models (Xiao & Wang, 2021; Zhou et al., 2024g; Farquhar et al., 2024; Lee et al., 2024a). When these models encounter unfamiliar OOD data, they often lack the ability to detect OOD data accurately, leading to highly over-confident predictions. This unreliability is especially concerning in safety-critical applications. The risk of

hallucination also increases as the model makes inferences based on its prior knowledge (Ren et al., 2023b; Li et al., 2025), which may not be applicable in the new context (Ji et al., 2023; Huang et al., 2025; Zhang et al., 2025). In the multimodal settings, training a model on evolving image-text data presents significant challenges, as distribution shifts in training data increase the potential of the model to forget previously learned knowledge and disrupt modality alignment. This can exacerbate hallucinations of the model by incorrectly aligning the relationships between observations and prior knowledge (Lee et al., 2024a; Zhu et al., 2024a; Jin & Ren, 2024). For example, changes in language usage or visual context can cause the model to lose its ability to associate specific textual descriptions with visual cues, resulting in errors when trying to generalize to unseen data, highlighting the need for effective memory retention mechanisms or continual learning strategies (Lopez-Paz & Ranzato, 2017; Srinivasan et al., 2022; Yoon et al., 2023; Cossu et al., 2024) to maintain seamless alignment between modalities.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121–140, 2020. URL `https://api.semanticscholar.org/CorpusID:221516138`.

Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023.

Taiga Abe, Estefany Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John P Cunningham. Deep ensembles work, but are they necessary? In *Advances in Neural Information Processing Systems*, 2022.

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4):1–35, 2018.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Chirag Agarwal. Intriguing properties of visual-language model explanations. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of distilbert. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 266–272, 2022.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

AIBase. Aibase: Deepseek app surpasses 100 million downloads in one month. *https://www.aibase.com/news/15598*, 2025.

Afra Feyza Akyürek, Eric Pan, Garry Kuwanto, and Derry Wijaya. Dune: Dataset for unified editing. *arXiv preprint arXiv:2311.16087*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. Prompting as probing: Using language models for knowledge base construction, 2022. *URL: http://arxiv. org/abs/2208.11057*, 2022.

Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.

Konstantinos Andriopoulos and Johan Pouwelse. Augmenting llms with knowledge: A survey on hallucination prevention. *arXiv preprint arXiv:2309.16459*, 2023.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *arXiv:2110.01052*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2023.

Anthropic. Claude 3.7 sonnet and claude code, February 2025. URL `https://www.anthropic.com/news/claude-3-7-sonnet`.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Marianna Apidianaki and Aina Garí Soler. ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality, October 2021. URL `http://arxiv.org/abs/2110.06376`. arXiv:2110.06376 [cs].

Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 339–364. IEEE, 2023.

Michael A Arbib. *The handbook of brain theory and neural networks*. MIT press, 2003.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*, 2024.

Mikhail J. Atallah, Victor Raskin, Michael Crogan, Christian F. Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding*, 2001. URL `https://api.semanticscholar.org/CorpusID:37687669`.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness Tests for Natural Language Explanations, June 2023. URL http://arxiv.org/abs/2305.18029. arXiv:2305.18029 [cs].

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*, 2023.

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. Identifying real or fake articles: Towards better language modeling. In *International Joint Conference on Natural Language Processing*, 2008. URL https://api.semanticscholar.org/CorpusID:4324753.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *IJCAI*, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*, abs/2212.08073, 2022b. URL https://arxiv.org/abs/2212.08073.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. Real or fake? learning to discriminate machine from human generated text. *ArXiv*, abs/1906.03351, 2019. URL https://api.semanticscholar.org/CorpusID:182952342.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of language models, 2024.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–123, 2023.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *ArXiv*, abs/2310.05130, 2023. URL https://api.semanticscholar.org/CorpusID:263831345.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021.

Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *CIKM*, pp. 2882–2887. ACM, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. Large Concept Models: Language modeling in a sentence representation space. 2024. URL https://arxiv.org/abs/2412.08821.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pp. 1–6. CEUR-WS, 2018.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Nina Baumgartner, Matthias Stürmer, Matthias Grabmair, Joel Niklaus, et al. Towards explainability and fairness in swiss judgement prediction: Benchmarking on a multilingual dataset. *arXiv preprint arXiv:2402.17013*, 2024.

BBC. "Art is dead Dude" - the rise of the AI artists stirs debate. https://www.bbc.com/news/technology-62788725, 2022.

Nils Begou, Jérémy Vinoy, Andrzej Duda, and Maciej Korczyński. Exploring the dark side of ai: Advanced phishing attack design and deployment using chatgpt. In *2023 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–6. IEEE, 2023.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, April 2022. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00422. URL https://direct.mit.edu/coli/article/48/1/207/107571/Probing-Classifiers-Promises-Shortcomings-and.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1–10, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1001.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pp. 86–102. Springer, 2023.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

Daria Beresneva. Computer-generated text detection using machine learning: A systematic review. In *International Conference on Applications of Natural Language to Data Bases*, 2016. URL https://api.semanticscholar.org/CorpusID:1175726.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. *ArXiv preprint*, abs/2308.09687, 2023. URL https://arxiv.org/abs/2308.09687.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Camiel J Beukeboom and Christian Burgers. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37, 2019.

Rahul Bhagat and Eduard H. Hovy. Squibs: What is a paraphrase? *Computational Linguistics*, 39:463–472, 2013. URL https://api.semanticscholar.org/CorpusID:32452685.

Amrita Bhattacharjee and Huan Liu. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21, 2024.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023.

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes, 2021.

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. Into the laions den: Investigating hate in multimodal datasets, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3(Jan):993–1022, 2003.

Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*, 2017.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Mikel Bober-Irizar, Ilia Shumailov, Yiren Zhao, Robert Mullins, and Nicolas Papernot. Architectural backdoors in neural networks, 2022.

Rishi Bommasani and Percy Liang. Trustworthy social bias measurement, 2022. URL https://arxiv.org/abs/2212.11672.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`.

Conrad Borchers, Dalia Sara Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Rose Kirk. Looking for a handsome carpenter! debiasing gpt-3 job advertisements. *arXiv preprint arXiv:2205.11374*, 2022.

Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

Ali Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *ArXiv*, abs/2304.06470, 2023. URL `https://api.semanticscholar.org/CorpusID:257826680`.

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22:71–85, 2012.

Dylan Bouchard, Mohit Singh Chauhan, David Skarbrevik, Viren Bajaj, and Zeya Ahmad. Langfair: A python package for assessing bias and fairness in large language model use cases, 2025. URL `https://arxiv.org/abs/2501.03112`.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Jack Brassil, Steven H. Low, Nicholas F. Maxemchuk, and Lawrence O'Gorman. Electronic marking and identification techniques to discourage document copying. *Proceedings of INFOCOM '94 Conference on Computer Communications*, pp. 1278–1287 vol.3, 1994. URL `https://api.semanticscholar.org/CorpusID:6540783`.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, 1950.

Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Amy Bunn. Artificial imposters—cybercriminals turn to ai voice cloning for a new breed of scam, 2024. URL `https://www.mcafee.com/blogs/privacy-identity-protection/artificial-imposters-cybercriminals-turn-to-ai-voice-cloning-for-a-new-breed-of-scam/`. [Accessed 09-09-2024].

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348*, 2023a.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*, 2024.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL `https://api.semanticscholar.org/CorpusID:233289412`.

Yihan Cao, Yanbin Kang, and Lichao Sun. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*, 2023b.

Captum. Testing with concept activation vectors (tcav) on sensitivity classification examples and a convnet model trained on imdb dataset. `https://github.com/pytorch/captum/blob/master/tutorials/TCAV_NLP.ipynb`, 2022.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36, 2024.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *ArXiv preprint*, abs/2307.15217, 2023. URL `https://arxiv.org/abs/2307.15217`.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*, 2022.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5029–5038, 2022.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Aliza Chasan. Ethical hacker scams 60 Minutes staffer to show how easy digital theft is, 2023. URL `https://www.cbsnews.com/news/how-digital-theft-targets-people-from-millennials-to-seniors-60-minutes-2023-05-21/`. [Accessed 09-09-2024].

Aditya Chattopadhyay, Kwan Ho Ryan Chan, Benjamin D Haeffele, Donald Geman, and René Vidal. Variational information pursuit for interpretable predictions. *arXiv preprint arXiv:2302.02876*, 2023.

Aditya Chattopadhyay, Kwan Ho Ryan Chan, and Rene Vidal. Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In *The Twelfth International Conference on Learning Representations*, 2024.

Hila Chefer, Oran Lang, Mor Geva, Volodymyr Polosukhin, Assaf Shocher, Michal Irani, Inbar Mosseri, and Lior Wolf. The hidden language of diffusion models. In *ICLR*, 2024.

Bocheng Chen, Advait Paliwal, and Qiben Yan. Jailbreaker in jail: Moving target defense for large language models. In *Proceedings of the 10th ACM Workshop on Moving Target Defense*, pp. 29–32, 2023a.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. Probing BERT in Hyperbolic Spaces, April 2021a. URL `http://arxiv.org/abs/2104.03869`. arXiv:2104.03869 [cs].

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, et al. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*, 2024a.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024b.

Hanjie Chen and Yangfeng Ji. Adversarial training for improving model robustness? look at both prediction and interpretation. In *The 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. Rev: information-theoretic evaluation of free-text rationales. *The 61th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023b.

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness, 2023.

Jiuhai Chen and Jonas Mueller. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*, 2024.

Jiuhai Chen, Lichang Chen, Heng Huang, and Tianyi Zhou. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*, 2023c.

Jiuhai Chen, Rifaa Qadri, Yuxin Wen, Neel Jain, John Kirchenbauer, Tianyi Zhou, and Tom Goldstein. Genqa: Generating millions of instructions from a handful of prompts. *arXiv preprint arXiv:2406.10323*, 2024c.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data, 2023d.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021b.

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023e.

Shuo Chen, Zhen Han, Bailan He, Zifeng Ding, Wenqian Yu, Philip Torr, Volker Tresp, and Jindong Gu. Red teaming gpt-4v: Are gpt-4v safe against uni/multi-modal jailbreak attacks? *arXiv preprint arXiv:2404.03411*, 2024d.

Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4035–4044, 2023f.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? counterfactual simulatability of natural language explanations, 2023g.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224*, 2023h.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback, 2023i.

Yutian Chen, Hao Kang, Vivian Zhai, Liang Li, Rita Singh, and Bhiksha Ramakrishnan. Gpt-sentinel: Distinguishing human and chatgpt generated content. *ArXiv*, abs/2305.07969, 2023j. URL `https://api.semanticscholar.org/CorpusID:258686680`.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 245–255, 2023k.

Zhaorun Chen, Zhuokai Zhao, Wenjie Qu, Zichen Wen, Zhiguang Han, Zhihong Zhu, Jiaheng Zhang, and Huaxiu Yao. Pandora: Detailed llm jailbreaking via collaborated phishing agents with decomposed reasoning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024e.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*, 2024f.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *arXiv preprint arXiv:2407.12784*, 2024g.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024h.

Zhaorun Chen, Mintong Kang, and Bo Li. Shieldagent: Shielding agents via verifiable safety policy reasoning. *arXiv preprint arXiv:2503.22738*, 2025b.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, 2023l.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model. *European Conference on Computer Vision (ECCV) (To appear)*, 2024.

Jiali Cheng and Hadi Amiri. Multimodal machine unlearning. *arXiv preprint arXiv:2311.12047*, 2023.

Siyuan Cheng, Bo Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? *ArXiv*, abs/2310.08475, 2023a. URL `https://api.semanticscholar.org/CorpusID:263908997`.

Siyuan Cheng, Ningyu Zhang, Bo Tian, Zelin Dai, Feiyu Xiong, Wei Guo, and Huajun Chen. Editing language model-based knowledge graph embeddings. *ArXiv*, abs/2301.10405, 2023b. URL `https://api.semanticscholar.org/CorpusID:256231427`.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043–3054, 2023.

Yujin Choi, Jinseong Park, Hoki Kim, Jaewook Lee, and Saeroom Park. Fair sampling in diffusion models through switching mechanism. *arXiv preprint arXiv:2401.03140*, 2024.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *IACR Cryptol. ePrint Arch.*, 2023:763, 2023. URL `https://api.semanticscholar.org/CorpusID:259092330`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

George Chrysostomou and Nikolaos Aletras. Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 477–488, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.40. URL https://aclanthology.org/2021.acl-long.40.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. In *BlackboxNLP@ACL*, pp. 276–286. Association for Computational Linguistics, 2019a.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

CNN. AI won an art contest, and artists are furious. https://www.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html, 2022.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions, 2023.

Competition and Markets Authority. Ai foundation models: initial review. *https://www.gov.uk/cma-cases/ai-foundation-models-initial-review*, 2023.

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492, 2024.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2025.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Knowledge neurons in pretrained transformers. *ArXiv*, abs/2104.08696, 2021. URL https://api.semanticscholar.org/CorpusID:233296761.

Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. *arXiv preprint arXiv:2307.12499*, 2023a.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. Exploring large language models for multi-modal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*, 2023b.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 253–262, 2004.

Google Deepmind. Identifying ai-generated images with synthid, 2023. URL https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*, 2023a.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. Rlprompt: Optimizing Discrete Text Prompts with Reinforcement Learning. In *EMNLP 2022*, pp. 3369–3391, 2022.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*, 2024a.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=vESNKdEMGp.

Zhijie Deng, Hongcheng Gao, Yibo Miao, and Hao Zhang. Efficient detection of llm-generated texts with a bayesian surrogate model. *ArXiv*, abs/2305.16617, 2023b. URL https://api.semanticscholar.org/CorpusID:258947640.

Zhun Deng, Thomas P. Zollo, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Distribution-free statistical dispersion control for societal applications, 2023c.

Joseph F. DeRose, Jiayao Wang, and Matthew Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Trans. Vis. Comput. Graph.*, 27(2):1160–1170, 2021.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers, 2020.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms, 2024.

Nicolas Deutschmann, Marvin Alberts, and María Rodríguez Martínez. Conformal autoregressive generation: Beam search with coverage guarantees, 2023.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7659–7666, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

Shizhe Diao, Zhichao Huang, Ruijia Xu, Xuechun Li, Yong Lin, Xiao Zhou, and Tong Zhang. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*, 2022.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*, 2023.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*, 2021.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Alphaeus Dmonte, Marcos Zampieri, Kevin Lybarger, and Massimiliano Albanese. Classifying human-generated and ai-generated election claims in social media. *arXiv preprint arXiv:2404.16116*, 2024.

Thang Viet Doan, Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness definitions in language models explained. *arXiv preprint arXiv:2407.18454*, 2024.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv preprint*, abs/2304.06767, 2023a. URL `https://arxiv.org/abs/2304.06767`.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. Calibrating factual knowledge in pretrained language models. *ArXiv*, abs/2210.03329, 2022a. URL `https://api.semanticscholar.org/CorpusID:252762125`.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022b.

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11275–11288, 2023b.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023c.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32):e2123433119, 2022.

Mengnan Du, Ninghao Liu, Fan Yang, Shuiwang Ji, and Xia Hu. On attribution of recurrent neural network predictions via additive decomposition. In *WWW*, pp. 383–393. ACM, 2019.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. Towards interpreting and mitigating shortcut learning behavior of nlu models. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM (CACM)*, 2023.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063, 2024a.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024b.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:255125274.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Kennedy Edemacu and Xintao Wu. Privacy preserving prompt engineering: A survey. *arXiv preprint arXiv:2404.06001*, 2024.

Bryan Eikema. The effect of generalisation on the inadequacy of the mode. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 87–92, 2024.

Bryan Eikema and Wilker Aziz. Is map decoding all you need? the inadequacy of the mode in neural machine translation. *arXiv preprint arXiv:2005.10283*, 2020.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking, 2024. URL https://arxiv.org/abs/2312.09244.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits — transformer-circuits.pub. https://transformer-circuits.pub/2021/framework/index.html, December 2021. [Accessed 27-11-2023].

Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. In *ACL (Findings)*, pp. 7555–7565. Association for Computational Linguistics, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024.

Chibuike Samuel Eze and Lior Shamir. Analysis and prevention of ai-based phishing email attacks. *Electronics*, 13(10):1839, 2024.

Hany Farid. Lighting (in)consistency of paint by text. *ArXiv*, abs/2207.13744, 2022a. URL https://api.semanticscholar.org/CorpusID:251135258.

Hany Farid. Perspective (in)consistency of paint by text. *ArXiv*, abs/2206.14617, 2022b. URL https://api.semanticscholar.org/CorpusID:250113700.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. In *Nature*, 2024.

Shanglun Feng and Florian Tramèr. Privacy backdoors: Stealing data with corrupted pretrained models. *arXiv preprint arXiv:2404.00473*, 2024.

Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16352–16362, 2023.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation, 2023.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2023a. URL https://api.semanticscholar.org/CorpusID:260351507.

Pierre Fernandez, Guillaume Couairon, Herv'e J'egou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22409–22420, 2023b. URL https://api.semanticscholar.org/CorpusID:257767463.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification, 2022. URL https://arxiv.org/abs/2208.12084.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation, 2020.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Yao Fu, Hao-Chun Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-Based Prompting for Multi-step Reasoning. *ICLR 2023 poster*, 2022.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024. URL https://arxiv.org/abs/2309.00770.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *arXiv preprint arXiv:2410.02760*, 2024.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv preprint*, abs/2209.07858, 2022. URL https://arxiv.org/abs/2209.07858.

Feng Gao, Liangzhi Shi, Shenao Zhang, Zhaoran Wang, and Yi Wu. Adaptive-gradient policy optimization: Enhancing policy learning in non-smooth differentiable simulations. In *Forty-first International Conference on Machine Learning*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023a.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3816–3830, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023b.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp. 311–319, 2022.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36, 2024.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:182952848.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option, 2019.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings, 2024. URL https://arxiv.org/abs/2411.04420.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

Reza Ghaeini, Xiaoli Z. Fern, Hamed Shahbazi, and Prasad Tadepalli. Saliency learning: Teaching the model where to pay attention. In *NAACL-HLT (1)*, pp. 4016–4025. Association for Computational Linguistics, 2019.

Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.330. URL `http://dx.doi.org/10.18653/v1/2021.findings-emnlp.330`.

Yoav Goldberg. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287, 2019.

Eric Goldman. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Google. Gemini 2.5 pro: Our most intelligent ai model, March 2025. URL `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/`. Accessed: April 6, 2025.

David Greenfield and Shivan Bhavnani. Social media: Generative ai could harm mental health. *Nature*, 617 (7962):676, 2023.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Roger B. Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamile Lukosiute, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large language model generalization with influence functions. *CoRR*, abs/2308.03296, 2023.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Jindong Gu. Responsible generative ai: What to generate and what not. *arXiv preprint arXiv:2404.05783*, 2024.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqian Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, Xiaochun Cao, and Philip Torr. A survey on transferability of adversarial examples across deep neural networks. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL `https://openreview.net/forum?id=AYJ3m7BocI`.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. In *ICML*, 2024b.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *ArXiv preprint*, abs/2308.08998, 2023. URL https://arxiv.org/abs/2308.08998.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *ArXiv*, abs/2301.07597, 2023a. URL https://api.semanticscholar.org/CorpusID:255998637.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*, 2023b.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares-López, Alexandre Ramé, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct Language Model Alignment from Online AI Feedback. *arXiv*, abs/2402.04792, 2024a.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024b.

Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation, 2024c. URL https://arxiv.org/abs/2411.10915.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023c.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond, 2024.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Michael Hahn and Navin Goyal. A theory of emergent in-context learning as implicit structure induction, 2023.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*, 2021.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024a.

Abhimanyu Hans, Yuxin Wen, Neel Jain, John Kirchenbauer, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, et al. Be like a goldfish, don't memorize! mitigating memorization in generative llms. *arXiv preprint arXiv:2406.10209*, 2024b.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. *Advances in neural information processing systems*, 36, 2024a.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024b.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *AAAI*, pp. 12963–12971. AAAI Press, 2021.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with grace: Lifelong model editing with discrete key-value adaptors. *ArXiv*, abs/2211.11031, 2022. URL `https://api.semanticscholar.org/CorpusID:253735429`.

Peter Hase, Mona T. Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srini Iyer. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL `https://api.semanticscholar.org/CorpusID:258378150`.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2001. ISBN 9780387952840. URL `https://books.google.com/books?id=VRzITwgNV2UC`.

Nan He, Hanyu Lai, Chenyang Zhao, Zirui Cheng, Junting Pan, Ruoyu Qin, Ruofan Lu, Rui Lu, Yunchen Zhang, Gangming Zhao, Zhaohui Hou, Zhiyuan Huang, Shaoqing Lu, Ding Liang, and Mingjie Zhan. Teacherlm: Teaching to fish rather than giving the fish, language modeling likewise, 2023.

Zexue He, Yu Wang, Julian J. McAuley, and Bodhisattwa Prasad Majumder. Controlling bias exposure for fair interpretable predictions. In *EMNLP (Findings)*, pp. 5854–5866. Association for Computational Linguistics, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *NAACL-HLT (1)*, pp. 4129–4138. Association for Computational Linguistics, 2019a.

John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL `https://aclanthology.org/N19-1419`.

John Hewitt, Sarah Chen, Lanruo Lora Xie, Edward Adams, Percy Liang, and Christopher D Manning. Model editing with canonical examples. *arXiv preprint arXiv:2402.06155*, 2024.

Yusuke Hirota, Jerone TA Andrew, Dora Zhao, Orestis Papakyriakopoulos, Apostolos Modas, Yuta Nakashima, and Alice Xiang. Resampled datasets are not enough: Mitigating societal bias beyond single attributes. *arXiv preprint arXiv:2407.03623*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2(4):5, 2024a.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp, 2024b.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

Yan Hong and Jianfu Zhang. Wildfake: A large-scale challenging dataset for ai-generated images detection. *arXiv preprint arXiv:2402.11843*, 2024.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exbert: A visual analysis tool to explore learned representations in transformer models. In *ACL (demo)*, pp. 187–196. Association for Computational Linguistics, 2020.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *ArXiv*, abs/2310.03991, 2023a. URL `https://api.semanticscholar.org/CorpusID:263831179`.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling, 2023b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models, 2025. URL `https://arxiv.org/abs/2501.03262`.

Michael Y. Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. Latent state models of training dynamics, 2023a.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M ^{4} i: Multi-modal models membership inference. *Advances in Neural Information Processing Systems*, 35:1867–1882, 2022a.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2225–2240, 2022b.

Xiaobing Hu, Pinyu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *ArXiv*, abs/2307.03838, 2023b. URL `https://api.semanticscholar.org/CorpusID:259501842`.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. *ArXiv*, abs/2310.10669, 2023c. URL `https://api.semanticscholar.org/CorpusID:264172471`.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text, 2018. URL https://arxiv.org/abs/1703.00955.

Alyssa Huang, Peihan Liu, Ryumei Nakada, Linjun Zhang, and Wanrong Zhang. Safeguarding data in multimodal ai: A differentially private approach to clip training. *arXiv preprint arXiv:2306.08173*, 2023a.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022.

K Huang, G Song, Hanwen Su, and Jiyan Wang. Out-of-distribution detection using peer-class generated by large language model. *arXiv preprint arXiv:2403.13324*, 2024a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025.

Qihan Huang, Jie Song, Mengqi Xue, Haofei Zhang, Bingde Hu, Huiqiong Wang, Hao Jiang, Xingen Wang, and Mingli Song. LG-CAV: Train any concept activation vector with language guidance. In *NeurIPS*, 2024b.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023b.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. Exploring distantly-labeled rationales in neural network models. In *ACL/IJCNLP (1)*, pp. 5571–5582. Association for Computational Linguistics, 2021.

Shuo Huang, William MacLean, Xiaoxi Kang, Qiongkai Xu, Zhuang Li, Xingliang Yuan, Gholamreza Haffari, and Lizhen Qu. Nap^2: A benchmark for naturalness and privacy-preserving text rewriting by learning from human. *arXiv preprint arXiv:2406.03749*, 2024c.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration, 2024d.

Yue Huang, Qihui Zhang, Lichao Sun, et al. Trustgpt: A benchmark for trustworthy and responsible large language models. *ArXiv preprint*, abs/2306.11507, 2023c. URL https://arxiv.org/abs/2306.11507.

Yunpeng Huang, Yaonan Gu, Jingwei Xu, Zhihong Zhu, Zhaorun Chen, and Xiaoxing Ma. Securing reliability: A brief overview on enhancing in-context learning for foundation models. *arXiv preprint arXiv:2402.17671*, 2024e.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. Transformer-patcher: One mistake worth one neuron. *ArXiv*, abs/2301.09785, 2023d. URL https://api.semanticscholar.org/CorpusID:256194369.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://dx.doi.org/10.1007/s10994-021-05946-3.

Timour Igamberdiev and Ivan Habernal. Dp-bart for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*, 2023.

Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *ArXiv preprint*, abs/2303.05398, 2023. URL https://arxiv.org/abs/2303.05398.

Alvi Md Ishmam and Christopher Thomas. Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24820–24830, 2024.

Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27611–27620. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.02608. URL http://dx.doi.org/10.1109/CVPR52733.2024.02608.

Takuya Ito, Soham Dan, Mattia Rigotti, James Kozloski, and Murray Campbell. On the generalization capacity of neural networks during generic multimodal reasoning. *arXiv preprint arXiv:2401.15030*, 2024.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*, 2021.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*, 2023a.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023b.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Theo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. Visqa: X-raying vision and language reasoning in transformers. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):976–986, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations, 2025.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy*, 2022.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

Shuyang Jiang, Yusheng Liao, Ya Zhang, Yu Wang, and Yanfeng Wang. Taia: Large language models are out-of-distribution data learners. *arXiv preprint arXiv:2405.20192*, 2024.

Haibo Jin, Ruoxi Chen, Jinyin Chen, and Haohan Wang. Quack: Automatic jailbreaking large language models via role-playing, 2024. URL https://openreview.net/forum?id=1zt8GWZ9sc.

Xisen Jin and Xiang Ren. What will my model forget? forecasting forgotten examples in language model refinement. In *International conference on machine learning*, 2024.

H. Jingnan. X's chatbot can now generate ai images. a lack of guardrails raises election concerns. *NPR*, August 16 2024. URL https://www.npr.org/2024/08/16/nx-s1-5078636/x-twitter-artificial-intelligence-trump-kamala-harris-election.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pp. 105–161. Springer, 1998.

Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. Er-test: Evaluating explanation regularization methods for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3315–3336, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift, 2020.

Masahiro Kaneko and Danushka Bollegala. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11954–11962, 2022.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *arXiv preprint arXiv:2311.09114*, 2023.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-rag: Certified generation risks for retrieval-augmented language models. *arXiv preprint arXiv:2402.03181*, 2024a.

Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36, 2024b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Sanjay Kariyappa, Freddy Lécué, Saumitra Mishra, Christopher Pond, Daniele Magazzeni, and Manuela Veloso. Progressive inference: Explaining decoder-only sequence classification models using intermediate predictions, 2024.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36, 2023.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. CLIP-QDA: an explainable concept bottleneck model. *CoRR*, abs/2312.00110, 2023.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In *ACL*, pp. 5435–5442. Association for Computational Linguistics, 2020.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

Daniel Khashabi, Shane Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, et al. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. *arXiv preprint arXiv:2112.08348*, 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), June 2018. URL `http://arxiv.org/abs/1711.11279`. arXiv:1711.11279 [stat].

Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. 2023a.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models, 2023b.

Jinhwa Kim, Ali Derakhshan, and Ian G. Harris. Robust safety classifier for large language models: Adversarial prompt shield. *arXiv preprint arXiv:2311.00172*, 2023c.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881*, 2023d.

Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. Carpe diem: On the evaluation of world knowledge in lifelong language models. In *The North American Chapter of the Association for Computational Linguistics*, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, 2023a. URL `https://api.semanticscholar.org/CorpusID:256194179`.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. *ArXiv*, abs/2306.04634, 2023b. URL `https://api.semanticscholar.org/CorpusID:259095643`.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pp. 1–30. Lille, 2015.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *ArXiv*, abs/2307.11729, 2023. URL `https://api.semanticscholar.org/CorpusID:260091573`.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Enja Kokalj, Blaz Skrlj, Nada Lavrac, Senja Pollak, and Marko Robnik-Sikonja. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In *EACL (Hackashop)*, pp. 16–21. Association for Computational Linguistics, 2021.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Fei Kong, Jinhao Duan, RuiPeng Ma, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024b.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the Dark Secrets of BERT, September 2019. URL `http://arxiv.org/abs/1908.08593`. arXiv:1908.08593 [cs, stat].

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *ArXiv*, abs/2303.13408, 2023. URL `https://api.semanticscholar.org/CorpusID:257687440`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Kuaishou Team. Kling. `https://klingai.kuaishou.com/`, 2024. Accessed: 2024-12-09.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *ArXiv*, abs/2307.15593, 2023. URL `https://api.semanticscholar.org/CorpusID:260315804`.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous questions with generative language models, 2023a.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023b.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023a.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering, 2023b.

Surender Suresh Kumar, M.L. Cummings, and Alexander Stimpson. Strengthening llm trust boundaries: A survey of prompt injection attacks surender suresh kumar dr. m.l. cummings dr. alexander stimpson. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, 2024.

Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations. *arXiv preprint arXiv:2403.10153*, 2024.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023a.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023b.

Jenny Kunz and Marco Kuhlmann. Classifier probes may just learn from linear context features. In *COLING*, pp. 5136–5146. International Committee on Computational Linguistics, 2020.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction, 2023.

Wen Lai, Alexander Fraser, and Ivan Titov. Joint localization and activation editing for low-resource fine-tuning. *arXiv preprint arXiv:2502.01179*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 537–563, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. Detecting fake content with relative entropy scoring. In *Pan*, 2008. URL `https://api.semanticscholar.org/CorpusID:12098535`.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. Xmd: An end-to-end framework for interactive explanation-based debugging of nlp models, 2022.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *ArXiv preprint*, abs/2309.00267, 2023a. URL `https://arxiv.org/abs/2309.00267`.

Jaewoo Lee, Jaehong Yoon, Wonjae Kim, Yunji Kim, and Sung Ju Hwang. Stella: Continual audio-video pre-training with spatio-temporal localized alignment. In *International Conference on Machine Learning*, 2024a.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. *ArXiv*, abs/2305.15060, 2023b. URL https://api.semanticscholar.org/CorpusID:258865409.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Rationalizing neural predictions. In *EMNLP*, pp. 107–117. The Association for Computational Linguistics, 2016.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pp. 707–710. Soviet Union, 1966.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bowen Li, Zhaoyu Li, Qiwei Du, Jinqi Luo, Wenshan Wang, Yaqi Xie, Simon Stepputtis, Chen Wang, Katia Sycara, Pradeep Ravikumar, Alexander Gray, Xujie Si, and Sebastian Scherer. Logicity: Advancing neuro-symbolic ai with abstract urban simulation. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024b.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *ArXiv preprint*, abs/2307.11760, 2023b. URL https://arxiv.org/abs/2307.11760.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models, 2024c. URL https://arxiv.org/abs/2402.10946.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Surinder Kumar. Large language models with controllable working memory. *ArXiv*, abs/2211.05110, 2022a. URL https://api.semanticscholar.org/CorpusID:253420654.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12006–12016, 2024d.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023c.

Jialu Li, Jaemin Cho, Yi-lin Sung, Jaehong Yoon, and Mohit Bansal. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. *arXiv preprint arXiv:2403.06952*, 2024e.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. Probing via Prompting, July 2022b. URL http://arxiv.org/abs/2207.01736. arXiv:2207.01736 [cs].

Jiarui Li, Ye Yuan, and Zehua Zhang. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*, 2024f.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023d.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*, 2023e.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In *ACL Long Papers*, 2025.

Xi'ang Li, Jinqi Luo, and Rabih Younes. Activitygan: generative adversarial networks for data augmentation in sensor-based human activity recognition. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '20 Adjunct, pp. 249–254, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380768. doi: 10.1145/3410530.3414367. URL `https://doi.org/10.1145/3410530.3414367`.

Xiang Lisa Li, Urvashi Khandelwal, and Kelvin Guu. Few-shot recalibration of language models, 2024g.

Xiaonan Li and Xipeng Qiu. Finding support examples for in-context learning, 2023.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating unsafe content generation in text-to-image models. *arXiv preprint arXiv:2404.06666*, 2024h.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *ICLR*, 2022c.

Yansong Li, Zhixing Tan, and Yang Liu. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*, 2023f.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023g.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023h.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles' heels of fake images with large multimodal models. *arXiv preprint arXiv:2404.13306*, 2024i.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. *arXiv preprint arXiv:2406.19234*, 2024j.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. *arXiv preprint arXiv:2411.02902*, 2024k.

Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps, 2023i.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 20763–20786, 2023a.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pp. 6565–6576. PMLR, 2021.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models, 2022a.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022b.

Qiyao Liang, Ziming Liu, Mitchell Ostrow, and Ila R Fiete. How diffusion models learn to factorize and compose. In *NeurIPS*, 2024.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7), 2023b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023a.

Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*, 2019.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2023b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Aiwei Liu, Leyi Pan, Xuming Hu, Shuang Li, Lijie Wen, Irwin King, and Philip S. Yu. An unforgeable publicly verifiable watermark for large language models. 2023a. URL `https://api.semanticscholar.org/CorpusID:260333928`.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. *ArXiv*, abs/2310.06356, 2023b. URL `https://api.semanticscholar.org/CorpusID:263830310`.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity, 2023c.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024a.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are llms at out-of-distribution detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8211–8222, 2024b.

Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. In *ACL (1)*, pp. 6274–6283. Association for Computational Linguistics, 2019.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023d.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1, 2023e.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023f.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023g.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021a.

Jiang Liu, Chun Pong Lau, and Rama Chellappa. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*, 2023h.

Jiang Liu, Chen Wei, Yuxiang Guo, Heng Yu, Alan Yuille, Soheil Feizi, Chun Pong Lau, and Rama Chellappa. Instruct2attack: Language-guided semantic adversarial attacks. *arXiv preprint arXiv:2311.15551*, 2023i.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021b.

Jihao Liu, Xin Huang, Jinliang Zheng, Boxiao Liu, Jia Wang, Osamu Yoshie, Yu Liu, and Hongsheng Li. Mm-instruct: Generated visual instructions for large multimodal model alignment. *arXiv preprint arXiv:2406.19736*, 2024c.

Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lei Hou, and Juanzi Li. Probing structured semantics understanding and generation of language models via question answering. *arXiv preprint arXiv:2401.05777*, 2024d.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, 2023j.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023k.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. *European Conference on Computer Vision (ECCV) (to appear)*, 2024e.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering, 2024f.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *ICML*, 2024g.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024h.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4711–4728, 2024i.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023l.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. Coco: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Conference on Empirical Methods in Natural Language Processing*, 2023m. URL https://api.semanticscholar.org/CorpusID:264406273.

Xin Liu, Muhammad Khalifa, and Lu Wang. Litcab: Lightweight language model calibration over short- and long-form responses, 2024j.

Xinwei Liu, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, and Xiaochun Cao. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2022a.

Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pp. 280–289. IEEE, 2022b.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023n.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking attention-model explainability through faithfulness violation test, 2022c.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *ArXiv*, abs/2304.07666, 2023o. URL https://api.semanticscholar.org/CorpusID:258179137.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Zeyuan Liu, Ziyu Huan, Xiyao Wang, Jiafei Lyu, Jian Tao, Xiu Li, Furong Huang, and Huazhe Xu. World models with hints of large language models for goal achieving. *arXiv preprint arXiv:2406.07381*, 2024k.

Zhihan Liu, Hao Hu, Shenao Zhang, Hongyi Guo, Shuqi Ke, Boyi Liu, and Zhaoran Wang. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency. *arXiv preprint arXiv:2309.17382*, 2023p.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *arXiv preprint arXiv:2405.16436*, 2024l.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*, 2024m.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pp. 8–23. World Scientific, 2023.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024a.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pp. 189–202, 2020.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning?, 2023.

Taiming Lu, Lingfeng Shen, Xinyu Yang, Weiting Tan, Beidi Chen, and Huaxiu Yao. It takes two: On the seamlessness between reward and policy model in rlhf. *arXiv preprint arXiv:2406.07971*, 2024b.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, volume 2, 2023.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. *arXiv preprint arXiv:2302.00539*, 2023.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023a.

Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De La Torre. Zero-shot model diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and Rene Vidal. Pace: Parsimonious concept engineering for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Hancheng Min, Chris Callison-Burch, and René Vidal. Concept lancet: Compositional representation transplant for diffusion-based image editing. In *CVPR*, 2025.

Saiyue Lyu, Margarita Vinaroz, Michael F Liu, and Mijung Park. Differentially private latent diffusion models. *arXiv preprint arXiv:2305.15759*, 2023.

Jiachen Ma, Anda Cao, Zhiqing Xiao, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. *arXiv preprint arXiv:2404.02928*, 2024a.

Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, Himabindu Lakkaraju, et al. Post hoc explanations of language models can improve language models. *arXiv preprint arXiv:2305.11426*, 2023a.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. Untying the reversal curse via bidirectional language model editing. *ArXiv*, abs/2310.10322, 2023b. URL `https://api.semanticscholar.org/CorpusID:264146289`.

Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *AdvML Frontiers workshop at 40th International Conference on Machine Learning*, 2023c.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model safety. *arXiv preprint arXiv:2502.05206*, 2025.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2421–2425, 2024b.

Zhe Ma, Xuhong Zhang, Qingming Li, Tianyu Du, Wenzhi Chen, Zonghui Wang, and Shouling Ji. Could it be generated? towards practical analysis of memorization in text-to-image diffusion models. *arXiv preprint arXiv:2405.05846*, 2024c.

Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve gpt-3 after deployment. *ArXiv*, abs/2201.06009, 2022. URL `https://api.semanticscholar.org/CorpusID:246016194`.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=S37hOerQLB`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. Raidar:geneRative AI Detection viA Rewriting. In *ICLR*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.

Rowan Hall Maudslay and Ryan Cotterell. Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing, June 2021. URL `http://arxiv.org/abs/2106.02559`. arXiv:2106.02559 [cs].

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*, 2019.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. 2024.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve, 2023.

Michal Měchura. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 168–173, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021a.

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8930–8938, 2021b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022a.

Kevin Meng, Arnab Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *ArXiv*, abs/2210.07229, 2022b. URL https://api.semanticscholar.org/CorpusID:252873467.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Bradley D Menz, Natansh D Modi, Michael J Sorich, and Ashley M Hopkins. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. *JAMA internal medicine*, 184(1):92–96, 2024.

Hasan Mesut Meral, Bülent Sankur, A. Sumru Özsoy, Tunga Güngör, and Emre Sevinç. Natural language watermarking via morphosyntactic alterations. *Comput. Speech Lang.*, 23:107–125, 2009. URL https://api.semanticscholar.org/CorpusID:1192689.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models, 2025.

Midjourney. Midjourney. https://midjourney.com/, 2023. Accessed: 2023.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks, 2021. URL https://arxiv.org/abs/2106.07998.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1816–1826, 2022.

Fatemehsadat Mireshghallah, Justus Mattern, Sicun Gao, R. Shokri, and Taylor Berg-Kirkpatrick. Smaller language models are better black-box machine-generated text detectors. *ArXiv*, abs/2305.09859, 2023. URL https://api.semanticscholar.org/CorpusID:258740888.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. *ArXiv*, abs/2110.11309, 2021. URL https://api.semanticscholar.org/CorpusID:239050360.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. *ArXiv*, abs/2206.06520, 2022. URL https://api.semanticscholar.org/CorpusID:249642147.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256274849.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. Exploring the role of BERT token representations to explain sentence probing results. In *EMNLP (1)*, pp. 792–806. Association for Computational Linguistics, 2021.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees, 2024.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: An overview. In *Explainable AI*, volume 11700 of *Lecture Notes in Computer Science*, pp. 193–209. Springer, 2019.

Jesse Mu and Jacob Andreas. Compositional Explanations of Neurons, February 2021. URL http://arxiv.org/abs/2006.14032. arXiv:2006.14032 [cs, stat].

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37:50972–51038, 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Travis J. E. Munyer and Xin Zhong. Deeptextmark: Deep learning based text watermarking for detection of large language model generated text. *ArXiv*, abs/2305.05773, 2023. URL https://api.semanticscholar.org/CorpusID:258588289.

Shikhar Murty, Christopher D. Manning, Scott M. Lundberg, and Marco Tulio Ribeiro. Fixing model bugs with natural language patches. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://api.semanticscholar.org/CorpusID:249147353.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training, 2023.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Krishna Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. PII-compass: Guiding LLM training data extraction prompts towards the target PII via grounding. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pp. 63–73, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Yang Nan, Huichi Zhou, Xiaodan Xing, and Guang Yang. Beyond the hype: A dispassionate look at vision-language models in medical scenario. *arXiv preprint arXiv:2408.08704*, 2024.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

Ali Naseh, Kalpesh Krishna, Mohit Iyyer, and Amir Houmansadr. Stealing the decoding algorithms of language models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1835–1849, 2023.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

Debora Nozza, Federico Bianchi, Dirk Hovy, et al. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads — transformer-circuits.pub. `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`, March 2022. [Accessed 27-11-2023].

OpenAI. ChatGPT. `https://openai.com/blog/chatgpt`, 2023a. Accessed: September 10, 2023.

OpenAI. Gpt-4 technical report, 2023b.

OpenAI. Moderator overview. `https://platform.openai.com/docs/guides/moderation/overview`, 2024a. Accessed: June 29, 2024.

OpenAI. Sora. `https://openai.com/sora`, 2024b. Accessed: Feburary 15, 2024.

OpenAI. Gpt-4.5, February 2025. URL `https://openai.com/index/introducing-gpt-4-5/`. Accessed: April 6, 2025.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, and et al. Openai o1 system card, 2024. URL `https://arxiv.org/abs/2412.16720`.

Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks, 2023.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3956–3965. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/ott18a.html`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019.

Anwesan Pal, Radhika Bhargava, Kyle Hinsz, Jacques Esterhuizen, and Sudipta Bhattacharya. The empirical impact of data sanitization on language models. *arXiv preprint arXiv:2411.05978*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. Sanvis: Visual analytics for understanding self-attention networks. In *2019 IEEE Visualization Conference (VIS)*, pp. 146–150. IEEE, 2019.

Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In *NeurIPS*, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2021.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL `https://arxiv.org/abs/2306.01116`.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018. URL `https://arxiv.org/abs/1802.05365`.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases?, September 2019. URL http://arxiv.org/abs/1909.01066. arXiv:1909.01066 [cs].

Buu Phan, Marton Havasi, Matthew Muckley, and Karen Ullrich. Understanding and mitigating tokenization bias in language models. *arXiv preprint arXiv:2406.16829*, 2024.

Renjie Pi, Tianyang Han, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv preprint arXiv:2401.02906*, 2024.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. Synchromesh: Reliable code generation from pre-trained language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=KmtVD97J43e.

Lip Yee Por, Koksheik Wong, and Kok Onn Chee. Unispach: A text-based data hiding method using unicode space characters. *J. Syst. Softw.*, 85:1075–1082, 2012. URL https://api.semanticscholar.org/CorpusID:17690312.

Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. In *NeurIPS*, 2023.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with" gradient descent" and beam search. *ArXiv preprint*, abs/2305.03495, 2023. URL https://arxiv.org/abs/2305.03495.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, volume 1, 2023a.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *ArXiv preprint*, abs/2310.03693, 2023b. URL https://arxiv.org/abs/2310.03693.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*, 2022.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Jiang, Chengfei Lv, and Huajun Chen. AutoAct: Automatic agent learning from scratch for QA via self-planning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3003–3021, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.165.

Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. Tool learning with foundation models. *ArXiv preprint*, abs/2304.08354, 2023a. URL https://arxiv.org/abs/2304.08354.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023b.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1488–1497, 2021.

Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for ai-generated text via error correction code. *arXiv preprint arXiv:2401.16820*, 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, and et al. Scaling language models: Methods, analysis & insights from training gopher, 2022. URL `https://arxiv.org/abs/2112.11446`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv preprint*, abs/2305.18290, 2023. URL `https://arxiv.org/abs/2305.18290`.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Alessandro Raganato and Jörg Tiedemann. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*. The Association for Computational Linguistics, 2018.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Kumar Ravikumar. From causal to concept-based representation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, 2016. URL `https://api.semanticscholar.org/CorpusID:11816014`.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.

Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *ArXiv preprint*, abs/2306.04488, 2023. URL https://arxiv.org/abs/2306.04488.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. Warp: On the benefits of weight averaged rewarded policies. *arXiv preprint arXiv:2406.16768*, 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models, 2024. URL https://arxiv.org/abs/2401.12187.

Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. In *ICLR*, 2024.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. Conformal nucleus sampling. *arXiv preprint arXiv:2305.02633*, 2023.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 88–102, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.starsem-1.10.

Shaina Raza, Ananya Raval, and Veronica Chatrath. Mbias: Mitigating bias in large language models while retaining context, 2024. URL https://arxiv.org/abs/2405.11290.

Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Position: Understanding llms requires more than statistical generalization. In *Forty-first International Conference on Machine Learning*, 2024.

Navid Rekabsaz and Markus Schedl. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2065–2068, 2020.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing. *ArXiv*, abs/2311.08721, 2023a. URL https://api.semanticscholar.org/CorpusID:265213008.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation, July 2023b. URL http://arxiv.org/abs/2307.11019. arXiv:2307.11019 [cs].

Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8116–8126. PMLR, 2020.

Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. Content-preserving text watermarking through unicode homoglyph substitution. *Proceedings of the 20th International Database Engineering & Applications Symposium*, 2016. URL https://api.semanticscholar.org/CorpusID:11689200.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Alexis Ross, Ana Marasovic, and Matthew E. Peters. Explaining NLP models via minimal contrastive editing (mice). In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 3840–3852. Association for Computational Linguistics, 2021.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *IJCAI*, pp. 2662–2670. ijcai.org, 2017.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct Nash Optimization: Teaching Language Models to Self-Improve with General Preferences. *arXiv*, 2024.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models, 2024.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*, 2023.

Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *ArXiv*, abs/2303.11156, 2023. URL `https://api.semanticscholar.org/CorpusID:257631570`.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.

Etienne Salimbeni, Francesco Craighero, Renata Khasanova, Milos Vasic, and Pierre Vandergheynst. Beyond fine-tuning: Lora modules boost near-ood detection and llm security. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024.

Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models. In *EMNLP (1)*, pp. 10285–10299. Association for Computational Linguistics, 2021.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

Irina Saparina and Mirella Lapata. Ambrosia: A benchmark for parsing ambiguous questions into database queries. *arXiv preprint arXiv:2406.19073*, 2024.

R. Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. Embarrassingly simple text watermarks. *ArXiv*, abs/2310.08920, 2023. URL `https://api.semanticscholar.org/CorpusID:264128148`.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. *arXiv preprint arXiv:1910.00292*, 2019.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv preprint*, abs/1707.06347, 2017. URL `https://arxiv.org/abs/1707.06347`.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling, 2022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In *conference on fairness, accountability and transparency*, pp. 48–48. PMLR, 2018.

Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*, 2024.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. Scaling retrieval-based language models with a trillion-token datastore. *arXiv preprint arXiv:2407.12854*, 2024.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*, 2023.

Lloyd S Shapley et al. A value for n-person games. 1953.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. Unirag: Universal retrieval augmentation for multi-modal large language models. *arXiv preprint arXiv:2405.10311*, 2024.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K. Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan, and Ethan Perez. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Junhong Shen, Hao Bai, Lunjun Zhang, Yifei Zhou, Amrith Setlur, Shengbang Tong, Diego Caples, Nan Jiang, Tong Zhang, Ameet Talwalkar, et al. Thinking vs. doing: Agents that reason by scaling test-time interaction. *arXiv preprint arXiv:2506.07976*, 2025.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*, 2024.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. *arXiv preprint arXiv:2310.05199*, 2023a.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023b.

Yaozong Shen, Lijie Wang, Ying Chen, Xinyan Xiao, Jing Liu, and Hua Wu. An Interpretability Evaluation Benchmark for Pre-trained Language Models, July 2022. URL `http://arxiv.org/abs/2207.13948`. arXiv:2207.13948 [cs].

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 38154–38180. Curran Associates, Inc., 2023c. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/77c33e6a367922d003ff102ffb92b658-Paper-Conference.pdf`.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

Ying Sheng, Shiyi Cao, Dacheng Li, Banghua Zhu, Zhuohan Li, Danyang Zhuo, Joseph E Gonzalez, and Ion Stoica. Fairness in serving large language models. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 965–988, 2024.

Haizhou Shi, Yibin Wang, Ligong Han, Huan Zhang, and Hao Wang. Training-free bayesianization for low-rank adapters of large language models. *arXiv preprint arXiv:2412.05723*, 2024a.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024b.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298*, 2023a.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023b.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023c.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning. *arXiv preprint arXiv:2306.17194*, 2023.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable, 2023.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural NLP models. In *ACL/IJCNLP (1)*, pp. 865–878. Association for Computational Linguistics, 2021.

Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. The confidence-competence gap in large language models: A cognitive study, 2023a.

Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023b.

Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023.

Shamane Siriwardhana, Mark McQuade, Thomas Gauthier, Lucas Atkins, Fernando Fernandes Neto, Luke Meyers, Anneketh Vij, Tyler Odenthal, Charles Goddard, Mary MacCarthy, et al. Domain adaptation of llama3-70b-instruct through continual pre-training and model merging: A comprehensive evaluation. *arXiv preprint arXiv:2406.14971*, 2024.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9211, 2022.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv*, abs/2408.03314, 2024.

Jake C. Snell, Thomas P. Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models. *ArXiv*, abs/1908.09203, 2019. URL `https://api.semanticscholar.org/CorpusID:201666234`.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 47783–47803. Curran Associates, Inc., 2023a. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/9521b6e7f33e039e7d92e23f5e37bbf4-Paper-Conference.pdf`.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, June 2023b.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2998–3009, 2023a.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 377–390, 2020.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. *ArXiv preprint*, abs/2306.17492, 2023b. URL `https://arxiv.org/abs/2306.17492`.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. Probing for Referential Information in Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4177–4189, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.384. URL `https://aclanthology.org/2020.acl-main.384`.

Claudio Spiess, David Gros, Kunal Suresh Pai, Michael Pradel, Md Rafiqul Islam Rabin, Amin Alipour, Susmit Jha, Prem Devanbu, and Toufique Ahmed. Calibration and correctness of language models for code, 2024.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. *arXiv preprint arXiv:2503.19206*, 2025.

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 35:29440–29453, 2022.

Joe Stacey, Yonatan Belinkov, and Marek Rei. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 11349–11357, 2022.

Elias Stengel-Eskin and Benjamin Van Durme. Calibrated interpretation: Confidence estimation in semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:1213–1231, 2023a.

Elias Stengel-Eskin and Benjamin Van Durme. Did you mean...? confidence-based trade-offs in semantic parsing. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in vqa, 2023a.

Elias Stengel-Eskin, Kyle Rawlins, and Benjamin Van Durme. Zero and few-shot semantic parsing with ambiguous inputs. In *The Twelfth International Conference on Learning Representations*, 2023b.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*, 2024.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4584–4596, 2023.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners, 2022.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*, 2024.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827. URL https://aclanthology.org/2023.findings-emnlp.827.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all, 2023b.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pp. 240–248. Springer, 2017.

Yang Sui, Huy Phan, Jinqi Xiao, Tianfang Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Disdet: Exploring detectability of backdoor attack on diffusion models. *arXiv preprint arXiv:2402.02739*, 2024.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. Evaluating large language models on controlled generation tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3155–3168, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.190. URL https://aclanthology.org/2023.emnlp-main.190.

Weisong Sun, Yuchen Chen, Guanhong Tao, Chunrong Fang, Xiangyu Zhang, Quanjun Zhang, and Bin Luo. Backdooring neural code search. *arXiv preprint arXiv:2305.17506*, 2023b.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024a.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 13088–13110, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.775.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *International Conference on Machine Learning (ICML)*, 2017.

Harini Suresh and John V. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2019. URL https://api.semanticscholar.org/CorpusID:235436386.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. Task ambiguity in humans and language models, 2022.

Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1541–1558, 2021.

Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, and Songcan Chen. Better safe than sorry: Preventing delusive adversaries with adversarial training. *Advances in Neural Information Processing Systems*, 34: 16209–16225, 2021.

Yitian Tao, Liyan Ma, Jing Yu, and Han Zhang. Memory-based cross-modal semantic alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Pittawat Taveekitworachai, Febri Abdullah, Mustafa Can Gursesli, Mury F Dewantoro, Siyuan Chen, Antonio Lanata, Andrea Guazzini, and Ruck Thawonmas. Breaking bad: Unraveling influences and risks of user inputs to chatgpt for game story generation. In *International Conference on Interactive Digital Storytelling*, pp. 285–296. Springer, 2023.

Claude3 Team. Claude3. *https://www.anthropic.com/news/claude-3-family*, 2024a.

Claude3.5 Team. Claude3.5. *https://www.anthropic.com/news/claude-3-5-sonnet*, 2024b.

Claude4 Team. Claude4. *https://www.anthropic.com/news/claude-4*, 2025a.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemini2 Team. Gemini2. *https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/*, 2024c.

GPT5 Team. Gpt5. *https://openai.com/index/introducing-gpt-5/*, 2025b.

Grok4 Team. Grok4. *https://x.ai/news/grok-4*, 2025c.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL `https://arxiv.org/abs/2501.12599`.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025d. URL `https://qwenlm.github.io/blog/qwq-32b/`.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *ACL (1)*, pp. 4593–4601. Association for Computational Linguistics, 2019a.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. *ArXiv*, abs/1905.06316, 2019b. URL https://api.semanticscholar.org/CorpusID:108300988.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. *arXiv preprint arXiv:2306.04597*, 2023.

David Thiel. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory*, 2023.

Edward Tian. GPTzero: An ai text detector., 2023. URL https://gptzero.me/.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality, 2023a.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023b.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Rubèn Tito, Khanh Nguyen, Marlon Tobaben, Raouf Kerkouche, Mohamed Ali Souibgui, Kangsoo Jung, Lei Kang, Ernest Valveny, Antti Honkela, Mario Fritz, et al. Privacy-aware document visual question answering. *arXiv preprint arXiv:2312.10108*, 2023.

Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, pp. 1–15, 2021.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in llms improves safety and reduces hallucinations, 2024.

Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, et al. Mj-video: Fine-grained benchmarking and rewarding video preferences in video generation. *arXiv preprint arXiv:2502.01719*, 2025.

Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. *arXiv preprint arXiv:2306.12105*, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024a.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024b.

Mercan Topkara, Umut Topkara, and Mikhail J. Atallah. Words are not enough: sentence level natural language watermarking. In *Workshop on Medical Cyber-Physical Systems*, 2006a. URL https://api.semanticscholar.org/CorpusID:5854860.

Umut Topkara, Mercan Topkara, and Mikhail J. Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Workshop on Multimedia & Security*, 2006b. URL https://api.semanticscholar.org/CorpusID:3061822.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic Probing through Dimension Selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 197–216, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.15. URL https://aclanthology.org/2020.emnlp-main.15.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Marcos V. Treviso, Alexis Ross, Nuno Miguel Guerreiro, and André Martins. CREST: A joint framework for rationalization and counterfactual text generation. In *ACL (1)*, pp. 15109–15126. Association for Computational Linguistics, 2023.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, S. Barannikov, Irina Piontkovskaya, Sergey I. Nikolenko, and Evgeny Burnaev. Intrinsic dimension estimation for robust detection of ai-generated texts. *ArXiv*, abs/2306.04723, 2023. URL https://api.semanticscholar.org/CorpusID:259108779.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023a.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv e-prints*, pp. arXiv–2308, 2023b.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

UBS. Ubs: Chatgpt may be the fastest growing app of all time. *https://aibusiness.com/nlp/ubs-chatgpt-is-the-fastest-growing-app-of-all-time*, 2024.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://api.semanticscholar.org/CorpusID:221835708.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Ting-Hao 'Kenneth' Huang, and Dongwon Lee. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2023. URL https://api.semanticscholar.org/CorpusID:257913864.

Dennis Ulmer. On uncertainty in natural language processing. *arXiv preprint arXiv:2410.03446*, 2024.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only. *arXiv preprint arXiv:2403.05973*, 2024a.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*, 2024b.

Dennis Ulmer, Chrysoula Zerva, and André FT Martins. Non-exchangeable conformal language generation with nearest neighbors. *arXiv preprint arXiv:2402.00707*, 2024c.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *NIPS 2016 Workshop on Cognitive Computation*, 2016.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao'Kenneth' Huang, and Shomir Wilson. Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*, 2023.

Aayush Atul Verma, Amir Saeidi, Shamanthak Hegde, Ajay Therala, Fenil Denish Bardoliya, Nagaraju Machavarapu, Shri Ajay Kumar Ravindhiran, Srija Malyala, Agneet Chatterjee, Yezhou Yang, et al. Evaluating multimodal large language models across distribution shifts and augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5314–5324, 2024.

Vivek Kumar Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. *ArXiv*, abs/2305.15047, 2023. URL `https://api.semanticscholar.org/CorpusID:258865787`.

Jesse Vig. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR workshop: Debugging machine learning models*, volume 23, 2019.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *BlackboxNLP@ACL*, pp. 63–76. Association for Computational Linguistics, 2019.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In *ACL (1)*, pp. 1264–1274. Association for Computational Linguistics, 2018.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Elena Voita, Rico Sennrich, and Ivan Titov. Analyzing the source and target contributions to predictions in neural machine translation. In *ACL/IJCNLP (1)*, pp. 1126–1140. Association for Computational Linguistics, 2021.

Ellen M Voorhees. Natural language processing and information retrieval. In *International summer school on information extraction*, pp. 32–48. Springer, 1999.

Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting for linear protocols. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

Oskar Van Der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs, 2025.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, 2023.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models. *arXiv preprint arXiv:2401.10491*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui,

Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022a.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11809–11820. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf.

Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: A framework and some existing methods. *TDKE*, 28(12):3395–3408, 2016.

Hao Wang and Dit-Yan Yeung. A survey on bayesian deep learning. *CSUR*, 53(5):1–37, 2020.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. Eviledit: Backdooring text-to-image diffusion models in one second. In *ACM Multimedia 2024*, 2024a. URL https://openreview.net/forum?id=ibEaSS6bQn.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024b.

Hengyi Wang, Shiwei Tan, Zhiqing Hong, Desheng Zhang, and Hao Wang. Variational language concepts for interpreting foundation language models. In *EMNLP*, 2024c.

Hengyi Wang, Shiwei Tan, and Hao Wang. Probabilistic conceptual explainers: Towards trustworthy conceptual explanations for vision foundation models. In *ICML*, 2024d.

Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy. *arXiv preprint arXiv:2501.09431*, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards codable watermarking for injecting multi-bit information to llm. 2023c. URL https://api.semanticscholar.org/CorpusID:260334887.

Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*, 2023d.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *arXiv preprint arXiv:2402.05813*, 2024e.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14743–14777, Bangkok, Thailand and virtual meeting, August 2024f. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.878.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024g.

Qingquan Wang. Invisible watermark, 2020. URL https://github.com/ShieldMnt/invisible-watermark.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11817–11831, Singapore, December 2023e. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.791. URL https://aclanthology.org/2023.findings-emnlp.791.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. *arXiv preprint arXiv:2401.04350*, 2024h.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 2023f.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han Xiao, Zixin Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, Yuchen Eleanor Jiang, and Wangchunshu Zhou. Weaver: Foundation models for creative writing, 2024i. URL https://arxiv.org/abs/2401.17268.

Xiyao Wang, Wichayaporn Wongkamjan, Ruonan Jia, and Furong Huang. Live in the moment: Learning dynamics model adapted to evolving policy. In *International Conference on Machine Learning*, pp. 36470–36493. PMLR, 2023g.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024j.

Xiyao Wang, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl. In *The Twelfth International Conference on Learning Representations*, 2024k. URL https://openreview.net/forum?id=MSe8YFbhUE.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023h.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. In *NeurIPS*, 2024l.

Yifan Wang and Vera Demberg. A parameter-efficient multi-objective approach to mitigate stereotypical bias in language models. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza (eds.), *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 1–19, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.gebnlp-1.1. URL https://aclanthology.org/2024.gebnlp-1.1/.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self-generated instructions. corr, abs/2212.10560, 2022. doi: 10.48550. *ArXiv preprint*, abs/2212.10560, 2022c. URL `https://arxiv.org/abs/2212.10560`.

Yu Wang, Xiusi Chen, Jingbo Shang, and Julian McAuley. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024m.

Yu Wang, Zeyuan Zhang, Julian McAuley, and Zexue He. Lvchat: Facilitating long video comprehension. *arXiv preprint arXiv:2402.12079*, 2024n.

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images. *Conference On Language Modeling (COLM)*, 2024o.

Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Democratizing reasoning ability: Tailored learning from large language model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1948–1966, Singapore, December 2023i. Association for Computational Linguistics. URL `https://aclanthology.org/2023.emnlp-main.120`.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. *arXiv preprint arXiv:2410.12735*, 2024p.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. *arXiv preprint arXiv:2407.00499*, 2024q.

Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Huaxiu Yao, Yue Zhang, Ren Wang, Kaidi Xu, and Xiaoshuang Shi. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *arXiv preprint arXiv:2402.14259*, 2024r.

Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. *arXiv preprint arXiv:2407.04215*, 2024s.

Futa Waseda and Antonio Tejero-de Pablos. Leveraging many-to-many relationships for defending against visual-language adversarial attacks. *arXiv preprint arXiv:2405.18770*, 2024.

WashingtonPost. He made a children's book using AI. Then came the rage. `https://www.washingtonpost.com/technology/2023/01/19/ai-childrens-\book-controversy-chatgpt-midjourney/`, 2022.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023a.

Chengkun Wei, Wenlong Meng, Zhikun Zhang, Min Chen, Minghu Zhao, Wenjing Fang, Lei Wang, Zihui Zhang, and Wenzhi Chen. Lmsanitator: Defending prompt-tuning against task-agnostic backdoors. *arXiv preprint arXiv:2308.13904*, 2023b.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023c.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization, 2022a.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023d.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4, 2023e.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. Promptriever: Instruction-trained retrievers can be prompted like language models. *arXiv preprint arXiv:2409.11136*, 2024.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*, 2024a.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3481–3495, 2024b.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51008–51025. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a00548031e4647b13042c97c922fadf1-Paper-Conference.pdf.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023b.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=84n3UwkH7b.

Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. Privacy backdoors: Enhancing membership inference through poisoning pre-trained models. In *NeurIPS*, 2024d.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. *arXiv preprint arXiv:2407.02814*, 2024.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation, 2019.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Llmdet: A third party large language models generated text detection tool. In *Conference on Empirical Methods in Natural Language Processing*, 2023a. URL https://api.semanticscholar.org/CorpusID:258865367.

Ruijia Wu, Yuhang Wang, Huafeng Shi, Zhipeng Yu, Yichao Wu, and Ding Liang. Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework. *arXiv preprint arXiv:2305.03980*, 2023b.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023c.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *ArXiv*, abs/2307.13339, 2023d. URL https://api.semanticscholar.org/CorpusID:260155139.

Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *ACL/IJCNLP (1)*, pp. 6707–6723. Association for Computational Linguistics, 2021.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv preprint arXiv:2310.00492*, 2023e.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models. *arXiv*, 2024.

Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. Dipmark: A stealthy, efficient and resilient watermark for large language models. *ArXiv*, abs/2310.07710, 2023f. URL https://api.semanticscholar.org/CorpusID:263834753.

Yixin Wu, Rui Wen, Michael Backes, Ning Yu, and Yang Zhang. Model stealing attacks against vision-language models. 2022.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *AAAI*, 2025.

Yuanwei Wu, Xiang Li, Yixin Liu, Pan Zhou, and Lichao Sun. Jailbreaking gpt-4v via self-adversarial attacks with system prompts. *arXiv preprint arXiv:2311.09127*, 2023g.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *ArXiv preprint*, abs/2306.01693, 2023h. URL https://arxiv.org/abs/2306.01693.

Zhengxuan Wu and Desmond C. Ong. On explaining your explanations of BERT: an empirical study with sequence classification. *CoRR*, abs/2101.00196, 2021.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *ACL*, pp. 4166–4176. Association for Computational Linguistics, 2020.

xAI. Grok 3 beta — the age of reasoning agents, February 2025. URL https://x.ai/news/grok-3. Accessed: April 6, 2025.

Peng Xia, Di Xu, Lie Ju, Ming Hu, Jun Chen, and Zongyuan Ge. Lmpt: Prompt tuning with class-specific embedding loss for long-tailed multi-label visual recognition. *arXiv preprint arXiv:2305.04536*, 2023.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024a.

Peng Xia, Ming Hu, Feilong Tang, Wenxue Li, Wenhao Zheng, Lie Ju, Peibo Duan, Huaxiu Yao, and Zongyuan Ge. Generalizing to unseen domains in diabetic retinopathy with disentangled representations. *arXiv preprint arXiv:2406.06384*, 2024b.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024c.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024d.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=c93SBwz1Ma.

Zhen Xiang, Yi Zeng, Mintong Kang, Chejian Xu, Jiawei Zhang, Zhuowen Yuan, Zhaorun Chen, Chulin Xie, Fengqing Jiang, Minzhou Pan, et al. Clas 2024: The competition for llm and agent safety. In *NeurIPS 2024 Competition Track*, 2024b.

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, et al. Large language models can be good privacy protection learners. *arXiv preprint arXiv:2310.02469*, 2023a.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11407–11427, 2023b.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis, 2022.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024a.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024b.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2021.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227, 2023a.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, pp. 1–11, 2023b.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024a.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024b.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pp. 11492–11501. PMLR, 2021a.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. A survey on green deep learning, 2021b. URL `https://arxiv.org/abs/2111.05193`.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D Wang, Joyce C Ho, Chao Zhang, and Carl Yang. Bmretriever: Tuning large language models as better biomedical text retrievers. *arXiv preprint arXiv:2404.18443*, 2024a.

Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. Equal long-term benefit rate: Adapting static fairness notions to sequential decision making. *arXiv preprint arXiv:2309.03426*, 2023.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024b.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*, 2024c.

Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization. In *AISTATS*, 2025b.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm, 2024a.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024b.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, and Julian J. McAuley. Learning concise and descriptive attributes for visual recognition. *CoRR*, abs/2308.03685, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *ArXiv preprint*, abs/2309.03409, 2023a. URL https://arxiv.org/abs/2309.03409.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14203–14214, 2025b.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021a.

Shu Yang, Muhammad Asif Ali, Cheng-Long Wang, Lijie Hu, and Di Wang. Moral: Moe augmented lora for llms' lifelong learning. *arXiv preprint arXiv:2402.11260*, 2024a.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. *Advances in Neural Information Processing Systems*, 36, 2024b.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. Tracing text provenance via context-aware lexical substitution. *ArXiv*, abs/2112.07873, 2021b. URL https://api.semanticscholar.org/CorpusID:245144237.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang rui Liu, Yuang Qi, Jie Zhang, Han Fang, and Neng H. Yu. Watermarking text generated by black-box language models. *ArXiv*, abs/2305.08883, 2023b. URL https://api.semanticscholar.org/CorpusID:258714683.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *ArXiv*, abs/2305.17359, 2023c. URL https://api.semanticscholar.org/CorpusID:258960101.

Xianjun Yang, Kexun Zhang, Haifeng Chen, Linda Ruth Petzold, William Yang Wang, and Wei Cheng. Zero-shot detection of machine-generated codes. *ArXiv*, abs/2310.05103, 2023d. URL https://api.semanticscholar.org/CorpusID:263831381.

Xinyu Yang, Huaxiu Yao, Allan Zhou, and Chelsea Finn. Multi-domain long-tailed learning by augmenting disentangled representations. *Transactions on Machine Learning Research*, 2023e.

Xinyu Yang, Zichen Wen, Wenjie Qu, Zhaorun Chen, Zhiying Xiang, Beidi Chen, and Huaxiu Yao. Memorization and privacy risks in domain-specific large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024c.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024d.

Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. *arXiv preprint arXiv:2403.01446*, 2024e.

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26275–26285, 2024f.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2024g.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*, 2022a.

Zhou Yang, Zhipeng Zhao, Chenyu Wang, Jieke Shi, Dongsun Kim, DongGyun Han, and David Lo. What do code models memorize? an empirical study on large language models of code. *arXiv preprint arXiv:2308.09932*, 2023f.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024h.

Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. *arXiv preprint arXiv:2210.04492*, 2022b.

Huaxiu Yao, Yiping Wang, Linjun Zhang, James Zou, and Chelsea Finn. C-mixup: Improving generalization in regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning (ICML)*, pp. 25407–25437. PMLR, 2022b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022c.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, may 2023. *ArXiv preprint*, abs/2305.10601, 2023a. URL `https://arxiv.org/abs/2305.10601`.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023b.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023c.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. In *NeurIPS*, 2024.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*, 2024.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning, 2023a.

Xiaoyu Ye, Hao Huang, Jiaqi An, and Yongtao Wang. Duaw: Data-free universal adversarial watermark against stable diffusion customization. *arXiv preprint arXiv:2308.09889*, 2023b.

Chih-Kuan Yeh, Joon Sik Kim, Ian En-Hsu Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks. In *NeurIPS*, pp. 9311–9321, 2018.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. *arXiv preprint arXiv:2306.01150*, 2023a.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know?, 2023b.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2023c.

Kiyoon Yoo, Wonhyuk Ahn, Jiho Jang, and No Jun Kwak. Robust multi-bit natural language watermarking through invariant features. In *Annual Meeting of the Association for Computational Linguistics*, 2023a. URL https://api.semanticscholar.org/CorpusID:259129912.

Kiyoon Yoo, Wonhyuk Ahn, and No Jun Kwak. Advancing beyond identification: Multi-bit watermark for language models. *ArXiv*, abs/2308.00221, 2023b. URL https://api.semanticscholar.org/CorpusID:262903996.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. *arXiv preprint arXiv:2505.14489*, 2025.

Jaehong Yoon, Sung Ju Hwang, and Yue Cao. Continual learners are incremental model generalizers. In *International Conference on Machine Learning*, 2023.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*, 2024.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Jiahao Yu, Xingwei Lin, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023a.

Liu Yu, Yuzhou Mao, Jin Wu, and Fan Zhou. Mixup-based unified framework to overcome gender bias resurgence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1755–1759, 2023b.

Mengxia Yu, De Wang, Qi Shan, Colorado Reed, and Alvin Wan. The super weight in large language models, 2024a.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. *arXiv preprint arXiv:2402.05889*, 2024b.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024c.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38, 2022.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Neng H. Yu. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. *ArXiv*, abs/2305.12519, 2023c. URL https://api.semanticscholar.org/CorpusID:258833423.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. Text fluoroscopy: Detecting llm-generated text through intrinsic features. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15838–15846, 2024d.

Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. Cold-start data selection for few-shot language model fine-tuning: A prompt-based uncertainty propagation approach, 2023d.

Zhuohao Yu, Xingru Jiang, Weizheng Gu, Chang Gao, Yidong Wang, Shikun Zhang, and Wei Ye. Saemark: Steering personalized multilingual llm watermarks with sparse autoencoders.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations, 2023a.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *ArXiv preprint*, abs/2304.05302, 2023b. URL https://arxiv.org/abs/2304.05302.

Junpeng Yue, Xinru Xu, Börje F Karlsson, and Zongqing Lu. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. *arXiv preprint arXiv:2410.03450*, 2024.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*, 2022.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031. IEEE, October 2019. doi: 10.1109/iccv.2019.00612. URL http://dx.doi.org/10.1109/ICCV.2019.00612.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 1, 05 2001.

Omar Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *EMNLP*, pp. 31–40. ACL, 2008.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *ArXiv*, abs/1905.12616, 2019. URL https://api.semanticscholar.org/CorpusID:168169824.

Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. Disentangling uncertainty in machine translation evaluation, 2022.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. *arXiv preprint arXiv:2305.04175*, 2023.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nBjmMF2IZU.

Haolan Zhan, Xuanli He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. G3detector: General gpt-generated text detector. *ArXiv*, abs/2305.12680, 2023. URL https://api.semanticscholar.org/CorpusID:258832418.

Andi Zhang, Tim Z Xiao, Weiyang Liu, Robert Bamler, and Damon Wischik. Your finetuned large language model is already a powerful out-of-distribution detector. *arXiv preprint arXiv:2404.08679*, 2024a.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms. *arXiv preprint arXiv:2403.20279*, 2024b.

Caiqi Zhang, Ruihan Yang, Zhisong Zhang, Xinting Huang, Sen Yang, Dong Yu, and Nigel Collier. Atomic calibration of llms in long-form generations. *arXiv preprint arXiv:2410.13246*, 2024c.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv preprint arXiv:2307.03941*, 2023a.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 110–120, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a.

Jie Zhang, Florian Kerschbaum, Tianwei Zhang, et al. Backdooring textual inversion for concept censorship. *arXiv preprint arXiv:2308.10718*, 2023b.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 993–999, 2023c.

Lining Zhang, Mengchen Wang, Liben Chen, and Wenxin Zhang. Probing GPT-3's Linguistic Knowledge on Semantic Tasks. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 297–304, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.blackboxnlp-1.24`.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023d.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*, 2023e.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. Remark-llm: A robust and efficient watermarking framework for generative large language models. *ArXiv*, abs/2310.12362, 2023f. URL `https://api.semanticscholar.org/CorpusID:264305916`.

Shenao Zhang. Conservative dual policy optimization for efficient model-based reinforcement learning. *Advances in neural information processing systems*, 35:25450–25463, 2022.

Shenao Zhang, Li Shen, and Lei Han. Learning meta representations for agents in multi-agent reinforcement learning. *arXiv preprint arXiv:2108.12988*, 2021a.

Shenao Zhang, Li Shen, Zhifeng Li, and Wei Liu. Structure-regularized attention for deformable object representation. *arXiv preprint arXiv:2106.06672*, 2021b.

Shenao Zhang, Wanxin Jin, and Zhaoran Wang. Adaptive barrier smoothing for first-order policy gradient with contact dynamics. In *International Conference on Machine Learning*, pp. 41219–41243. PMLR, 2023g.

Shenao Zhang, Boyi Liu, Zhaoran Wang, and Tuo Zhao. Model-based reparameterization policy gradient methods: Theory and practical algorithms. *Advances in Neural Information Processing Systems*, 36, 2024d.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024e.

Shenao Zhang, Sirui Zheng, Shuqi Ke, Zhihan Liu, Wanxin Jin, Jianbo Yuan, Yingxiang Yang, Hongxia Yang, and Zhaoran Wang. How can llm guide rl? a value-based approach. *arXiv preprint arXiv:2402.16181*, 2024f.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024g.

Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5189–5202, 2023h.

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024h.

Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *arXiv preprint arXiv:2302.03675*, 2023i.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv preprint arXiv:2406.07057*, 2024i.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*, 2023j.

Yiming Zhang and Daphne Ippolito. Prompts should not be seen as secrets: Systematically measuring prompt extraction attack success. *arXiv preprint arXiv:2307.06865*, 2023.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining. *arXiv preprint arXiv:2404.09387*, 2024j.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *International conference on machine learning*, pp. 4006–4015. PMLR, 2017b.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pp. 1–45, 2025.

Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, Rui Li, Pei Ke, Hao Sun, Lei Sha, Zhifang Sui, Hongning Wang, and Minlie Huang. Shieldlm: Empowering llms as aligned, customizable and explainable safety detectors. *arXiv preprint arXiv:2402.16444*, 2024k.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024l.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *CoRR*, abs/2309.01029, 2023a.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*, 2023b.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models, 2021.

Xingyi Zhao, Depeng Xu, and Shuhan Yuan. Defense against backdoor attack on pre-trained language models via head pruning and attention normalization. In *Forty-first International Conference on Machine Learning*, 2024a.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *ArXiv*, abs/2306.17439, 2023c. URL `https://api.semanticscholar.org/CorpusID:259308864`.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. *ArXiv preprint*, abs/2210.00045, 2022. URL `https://arxiv.org/abs/2210.00045`.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *ArXiv preprint*, abs/2305.10425, 2023d. URL `https://arxiv.org/abs/2305.10425`.

Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*, 2023e.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhangp Zidong Dup Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? *CVPR*, 2024b.

Zhengyue Zhao, Xiaoyun Zhang, Kaidi Xu, Xing Hu, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Adversarial contrastive decoding: Boosting safety alignment of large language models via opposite prompt optimization. *arXiv preprint arXiv:2406.16743*, 2024c.

Boyang Zheng, Chumeng Liang, Xiaoyu Wu, and Yan Liu. Understanding and improving adversarial attacks on latent diffusion model. *arXiv preprint arXiv:2310.04687*, 2023a.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *ArXiv*, abs/2305.12740, 2023b. URL `https://api.semanticscholar.org/CorpusID:258832407`.

Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual Probing Is [MASK]: Learning vs. Learning to Recall, December 2021. URL `http://arxiv.org/abs/2104.05240`. arXiv:2104.05240 [cs].

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL `https://api.semanticscholar.org/CorpusID:258865984`.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: less is more for alignment. *CoRR*, abs/2305.11206, 2023a.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024a.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering, 2024b.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, and Zhenhao Li. Trustrag: Enhancing robustness and trustworthiness in rag. *arXiv preprint arXiv:2501.00879*, 2025a.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models, 2023b.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use fourier features to compute addition, 2024c.

Wangchunshu Zhou and Ke Xu. Learning to compare for better training and evaluation of open domain natural language generation models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 9717–9724, Apr. 2020. doi: 10.1609/aaai.v34i05.6521. URL https://ojs.aaai.org/index.php/AAAI/article/view/6521.

Wangchunshu Zhou, Tao Ge, Chang Mu, Ke Xu, Furu Wei, and Ming Zhou. Improving grammatical error correction with machine translation pairs. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 318–328, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.30. URL https://aclanthology.org/2020.findings-emnlp.30.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. Self-adversarial learning with comparative discrimination for text generation. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=B1l8L6EtDS.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27395–27411. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/zhou22n.html.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Xiangru Tang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. Agents: An open-source framework for autonomous language agents, 2023c. URL https://arxiv.org/abs/2309.07870.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. Controlled text generation with natural language instructions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42602–42613. PMLR, 23–29 Jul 2023d. URL https://proceedings.mlr.press/v202/zhou23g.html.

Wangchunshu Zhou, Yixin Ou, Shengwei Ding, Long Li, Jialong Wu, Tiannan Wang, Jiamin Chen, Shuai Wang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, and Yuchen Eleanor Jiang. Symbolic learning enables self-evolving agents, 2024d. URL https://arxiv.org/abs/2406.18532.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*, 2024e.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024f.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024g.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024h.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, et al. Anyprefer: An automatic framework for preference data synthesis. In *International Conference on Learning Representations (ICLR)*, 2025b.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *ArXiv preprint*, abs/2211.01910, 2022b. URL https://arxiv.org/abs/2211.01910.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2023e.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, 2023a.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023b.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023c.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *International conference on machine learning*, 2024a.

Liuwan Zhu, Rui Ning, Jiang Li, Chunsheng Xin, and Hongyi Wu. Seer: Backdoor detection for vision-language models through searching target text and image trigger jointly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7766–7774, 2024b.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*, 2023d.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*, pp. 10–18. PMLR, 2014.

Thomas P Zollo, Todd Morrill, Zhun Deng, Jake C Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. *arXiv preprint arXiv:2311.13628*, 2023.

Thomas P Zollo, Zhun Deng, Jake Snell, Toniann Pitassi, and Richard Zemel. Improving predictor reliability with selective recalibration. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=Aoj9H6jl6F. Expert Certification.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models, 2024b.

Thomas P. Zollo, Nikita Rajaneesh, Richard Zemel, Talia B. Gillis, and Emily Black. Towards effective discrimination testing for generative ai, 2024c. URL https://arxiv.org/abs/2412.21052.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models, 2024.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses, 2018.