On the Challenges and Opportunities in Generative AI

Anonymous authors
Paper under double-blind review

Abstract

The field of deep generative modeling has grown rapidly in the last few years. With the availability of massive amounts of training data coupled with advances in scalable unsupervised learning paradigms, recent large-scale generative models show tremendous promise in synthesizing high-resolution images and text, as well as structured data such as videos and molecules. However, we argue that current large-scale generative AI models exhibit several fundamental shortcomings that hinder their widespread adoption across domains. In this work, our objective is to identify these issues and highlight key unresolved challenges in modern generative AI paradigms that should be addressed to further enhance their capabilities, versatility, and reliability. By identifying these challenges, we aim to provide researchers with insights for exploring fruitful research directions, thus fostering the development of more robust and accessible generative AI solutions.

1 Introduction

The past few years have demonstrated the immense potential of large-scale generative models to create powerful AI tools capable of impacting society profoundly. Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023; Rae et al., 2021) and their dialogue agents, such as ChatGPT (OpenAI, 2023) and Llama 3 (Grattafiori et al., 2024) have enabled the development of highly effective text generation systems that produce coherent, contextually relevant, and user-tailored outputs across a wide range of use cases. Similarly, advancements in diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020; Ho et al., 2020) have led to groundbreaking advancements in image synthesis tasks, such as large-scale text-to-image generation (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022; Esser et al., 2024). These successes show that highly effective AI systems can be built using a relatively straightforward recipe: combining simple generative modeling paradigms (Larochelle & Murray, 2011; Sohl-Dickstein et al., 2015) with successful network architectures (Vaswani et al., 2017; Dosovitskiy et al., 2020; Ronneberger et al., 2015), training on large-scale datasets, and the incorporation of preferences via human feedback (Ouyang et al., 2022; Ziegler et al., 2019). The impact of generative AI has not been limited to text and image generation applications. It has fueled accelerated progress across a variety of research fields and practical applications, spanning from biology (Jumper et al., 2021) to weather forecasting (Ravuri et al., 2021), from code generation (Chen et al., 2021b; Li et al., 2022b) to video creation (Yang et al., 2023c; Ho et al., 2022b; Singer et al., 2022; Brooks et al., 2024), audio synthesis (Borsos et al., 2023; Liu et al., 2023a), and even artistic and musical composition (Huang et al., 2023b).

With the current advancements and excitement surrounding generative AI, a question naturally arises: Are we on the brink of an AI utopia? Are we close to developing what we might call a *perfect generative model*? For the purpose of this survey, we define such a model as a single architecture that (i) can approximate the joint data distribution of any modality, (ii) provides calibrated uncertainty and causal consistency, and (iii) delivers controllable outputs that satisfy stringent requirements on robustness, safety, efficiency and societal alignment. We argue in this paper that the answer is a resounding *no*; rather, the realization of such a model, one that would fundamentally transform the field of AI is still hampered by substantial theoretical, practical, and ethical challenges, and incremental advances alone are unlikely to close the gap in the near term.

Amidst the excitement and anticipation surrounding this new wave of Deep Generative Models (DGMs),[1] it is easy to overlook the new set of challenges they introduce. Unlike many of the traditional machine learning models, DGMs generate outputs in very high-dimensional spaces, which introduces several technical complexities. These include significantly increased computational demands, a need for larger datasets to accurately capture the underlying data distribution, and challenges in effectively evaluating and interpreting the generated outputs. And while significant progress has been made in improving interpretability and computational efficiency for traditional models (Marcinkevics & Vogt, 2020), these existing methods are frequently ill-suited for DGMs, at least in part because of the complex and high-dimensional nature of their outputs. Consequently, there is a pressing need for the development of a new set of techniques and tools tailored to these models, particularly to enable efficient inference, interpretability and quantization. These challenges lead us to conclude that scaling up current paradigms is *not* in isolation the ultimate path towards a perfect generative model. While increasing model size and training data can enhance performance on benchmarks, it does little to address the fundamental shortcomings of DGMs, such as their inefficiency, lack of inclusivity, limited transparency, and barriers to usability—particularly in high-stakes domains where reliability and fairness are paramount.

This work offers a collection of views and opinions from different communities about these key unresolved challenges in generative AI, with the ultimate goal of guiding future research toward what we perceive are the most critical and promising areas. Concretely, we discuss key challenges in (a) broadening the *scope and adaptability* of DGMs, i.e., their ability to robustly generalize across different domains and modalities (Section 2); (b) improving their *efficiency and resource utilization*, i.e., to lower the memory and computational requirements and enhance accessibility and sustainability in their adoption (Section 3); and, finally, (c) addressing *ethical and societal concerns* that are crucial for responsible deployment (Section 4).

This paper emerged as a result of the Dagstuhl Seminar on *Challenges and Perspectives in Deep Generative Modeling*[2] held in Spring 2023. By outlining a comprehensive roadmap of the current state and open challenges in generative AI, this paper aims to serve as an integrated entry point for researchers and practitioners. Although many existing surveys offer in-depth reviews of specific subfields, such as robustness, causal modeling, or modality-specific techniques, they are often narrowly focused, making it difficult to grasp the broader landscape. In contrast, our goal is to synthesize high-level insights across domains, highlight emerging research directions, and guide readers toward foundational and topic-specific work. In doing so, our goal is to foster a more robust, inclusive, and accessible development of generative AI systems.

2 Expanding Scope and Adaptability

State-of-the-art leaderboard rankings show the remarkable progress in model performance that has been made by scaling DGMs to massive datasets and model sizes (for instance, in text and high-resolution image synthesis). However, automatic evaluations on popular benchmark datasets cannot be our only measure of model success (Bender et al., 2021); such evaluations often fail to capture the nuanced limitations of DGMs, such as potential biases, inabilities to generalize to inputs from underrepresented or specialized distributions, and difficulties with aligning outputs with specific domain requirements. Understanding these inherent and often hidden constraints is essential for ensuring that DGMs can be reliably applied to various

[1] In this paper, we refer to Generative AI as a collection of large-scale DGMs and use the term DGM henceforth.
[2] `https://www.dagstuhl.de/23072`

Sub-challenge	**Typical failure mode**	**Promising mitigation avenues**	**Reference materials**	**Research-question candidates (paper §)**
Robust generalisation to OOD inputs (§2.1)	Large performance drop on unseen domains	Retrieval-augmented generation (RAG); Group DRO training (e.g., Sagawa* et al., 2020)	**Datasets:** WILDS; ImageNet-C **Surveys:** Shen et al. (2021); Yang et al. (2023b); Li et al. (2023c)	*Q1*: Can retrieval-augmented generators help close the WILDS gap without model retraining? *Q2*: Can we provide an inductive bias for foundation models via architectural modifications or training objectives that predisposes them to accurately capture tail events?
Resilience to adversarial perturbations (§2.1)	Imperceptible noise fools the model	Gaussian noise smoothing; Certified defense methods; Adversarial finetuning for diffusion	**Benchmarks:** RobustBench Leaderboard; NSFW Adversarial Benchmark **Surveys:** Sun et al. (2023b)	*Q1*: Will certifiably-robust diffusion samplers scale to ImageNet? *Q2*: How can we unify adversarial training for text & images in a single multi-modal framework?
Mitigating learning of spurious correlations (§2.1 / 2.2)	Model predicts based on background cues, not capturing meaningful relationships	Counterfactual data augmentation; Invariant Risk Minimization-based approaches	**Datasets:** Waterbirds; Colored-MNIST **Surveys:** Ye et al. (2024)	*Q1*: How can we reliably detect and quantify hidden spurious cues encoded by foundation-model features? *Q2*: Which causal probes best expose hidden biases in large latent spaces?
Capturing causal dependencies (§2.2)	Models generate statistically plausible but causally impossible outcomes	Causal masking in attention layers; SCM-guided loss functions; Interventional training objectives	**Benchmarks:** CausalBench **Surveys:** Komanduri et al. (2024)	*Q1*: How can causal invariance objectives be integrated into generative model training to improve robustness to distribution shifts? *Q2*: Can tractable surrogate objectives approximate interventional likelihood, enabling scalable causal DGM training?
Accounting for implicit assumptions (§2.3)	(Implicitly) assumed characteristics of data-generating distribution do not persist under domain shift	Domain-expert knowledge integration; Statistical assumption testing (e.g., independence, stationarity checks)	—	*Q1*: How can we quantify the degree of modeling-assumption violation before deploying a generative model? *Q2*: What learning objectives are most robust when the true data-generating process lies outside the model family?
Cross-modal transfer in specialized domains (§2.4)	Failure/inability to link signals across modalities (e.g. ECG $\leftrightarrow$ notes)	Contrastive multimodal pre-training (e.g., Raghu et al., 2022)	**Datasets:** MMIST-CCRCC; GMAI-MMBench **Surveys:** Shaik et al. (2024)	*Q1*: How can physiological constraints be incorporated into multi-modal pre-training objectives for medical domains? *Q2*: Which training/fine-tuning methods can achieve sufficient cross-modal alignment in low-resource clinical settings?

Table 1: Research challenges summary table - Expanding Scope and Adaptability

real-world tasks, where data characteristics, domain-specific constraints, and measures of success may differ significantly from those in standardized benchmarks (Durall et al., 2020; Daunhawer et al., 2022; Xu et al., 2024). This section analyzes some of these challenges in the context of large-scale DGMs from the lens of their generalization capabilities (Section 2.1) and the lack of transparency in their underlying modeling assumptions (Section 2.3). We examine these fundamental challenges and provide research directions that could broaden the adaptability of DGMs to promote long-term progress in the field. We also discuss two promising avenues that have the potential to greatly enhance the scope of generative models: (i) integrating causal learning (Section 2.2) and (ii) the development of a versatile, generalist agent capable of handling heterogeneous data types (Section 2.4).

2.1 Generalization and Robustness

To ensure reliability across various domains, DGMs must generalize effectively under shifts to the data-generating distribution of inputs, often referred to as *out-of-distribution (OOD) robustness*, and be resilient to minor variations in the input, a necessary component of the broader notion of *adversarial robustness*. Without proper generalization, generative models may produce unrealistic or biased outputs,[3] limiting their practical utility and trustworthiness in real-world applications.

While large-scale generative models show some promise in achieving OOD robustness (Wang et al., 2023a), these models still face challenges in accurately capturing rare events or responding to adversarial inputs (Zhu et al., 2024), a difficulty that lies in effectively modeling the *tail* of information (Kandpal et al., 2023), i.e., the information that appears rarely or only once in the dataset used to (pre)train the model. This limitation indicates a gap in their ability to fully represent the vast and diverse spectrum of real-world scenarios, especially those that are less common but equally significant. Retrieval-augmented language models represent a promising approach for integrating rare or specialized knowledge into model outputs, effectively addressing challenges that cannot be resolved solely by scaling up training datasets (Kandpal et al., 2023). In the vision domain, test-time approaches, such as Generalized Diffusion Adaptation (Tsai et al., 2024), present a promising avenue towards attaining OOD robustness.

DGMs are also prone to adversarial vulnerability, often due to the presence of highly predictive but non-robust features that are used as *shortcuts* for prediction (Du et al., 2023a; Puli et al., 2023; Webson & Pavlick, 2022). This behavior poses a significant threat to various downstream scenarios, especially those of safety-critical applications (Poursaeed et al., 2021; Wang et al., 2023a). Several approaches to mitigate the effect of shortcut learning are based on model refinement or on dataset refinement, also known as data-centric approaches (Whang et al., 2021; Zha et al., 2025). In the former, work has been done towards improving robustness via adversarial training (Zou et al., 2023b; Choi et al., 2025), feature masking during training (Asgari et al., 2022), ensembling (Clark et al., 2019), contrastive learning (Choi et al., 2022), and the direct integration of prior knowledge (Ilyas et al., 2019). The latter includes improving the quality of the data used by large-scale models during training, such as through augmentation (Zhang et al., 2018), labeling (Kutlu et al., 2020)) and inference techniques—for example, employing prompt engineering (Wallace et al., 2019) or data slicing (Chung et al., 2019).

However, in most applications, foundation models are often adapted to specific tasks and downstream datasets. Standard fine-tuning techniques often overemphasize the target task, leading to catastrophic forgetting (Thanh-Tung & Tran, 2020) and a loss in the general robustness of the upstream model (Suprem & Pu, 2022). Therefore, a significant challenge is to develop robust adaptation methods that adequately solve the target task but still maintain the beneficial robustness properties of the upstream model (e.g., robustness to distribution shifts of the target dataset) (Balaji et al., 2020; Du et al., 2023a; Han et al., 2021; Liu et al., 2020). These same issues come into play when developing smaller and more efficient models for the sake of economization of DGM inference and memory costs—which we discuss in greater detail in Section 3. In this context, it is important to develop robust distillation methods that do not sacrifice the robustness of the model (Du et al., 2023b; Zi et al., 2021). We argue that two particularly promising approaches to obtaining robust and interpretable models are embedding causal structure and explicitly encoding human priors into the training process—topics we examine in the following sections.

2.2 Causal Generative Models

Going beyond learning mere statistical correlations and understanding how underlying factors influence the generative process is the main objective of learning a causal structure of data (Pearl & Mackenzie, 2018). Structural Causal Models (SCMs) provide the mathematical foundation for this endeavor, representing causal relationships through directed acyclic graphs paired with structural equations that encode how variables causally influence one another. Such knowledge can be used to reason about hypothetical scenarios in the world, understand the effect of interventions, and perform counterfactuals (Pearl, 2019), thus facilitating

[3]Here, we use the terminology *biased outputs* to refer to systematic deviations in model outputs caused by imbalances or inaccuracies in the training data and/or modeling process. These outputs then do not accurately reflect the true underlying data distribution or are skewed in ways that perpetuate inaccuracies, stereotypes, or unfair conceptions about certain outcomes.

informed decision-making. Although there have been attempts to develop methods for learning the optimal generative structure of deep latent variable models from data (He et al., 2019; Manduchi et al., 2023), current generative models often neglect the underlying causal dependencies in their generative processes, making them prone to shortcut learning and spurious associations (Gururangan et al., 2018; McCoy et al., 2019).

Causal generative models have the potential to offer distribution-shift robustness, fairness, and interpretability (Schölkopf et al., 2021; Wang & Jordan, 2021). They are either focused on causal representation learning, which discovers causally related latent variables, or controllable counterfactual generation, which, instead, focuses on learning a mapping between data and known causal variables. For a detailed review of the topic, we refer to (Komanduri et al., 2024). Current open challenges include but are not limited to, scalable and robust causal discovery from observational data (Reizinger et al., 2023; Zhou et al., 2022; Montagna et al., 2024), identifiability of deep generative models under weaker forms of supervision (Ahuja et al., 2023; Locatello et al., 2020; von Kügelgen et al., 2024), lack of benchmark datasets and metrics to evaluate counterfactual quality (Monteiro et al., 2023), strong assumptions that are often violated in real-world applications (Komanduri et al., 2024), and, finally, the integration of diffusion models, a field that is currently under-explored but has tremendous growth potential (Mittal et al., 2021; Pandey et al., 2022; Sanchez & Tsaftaris, 2022; Sanchez et al., 2023). We suggest that the integration of causal principles in DGMs could pave the way for the development of more robust, interpretable, and actionable generative AI systems (Zhou et al., 2023).

2.3 Accounting for Implicit Assumptions

Silent Assumptions. Current generative models often make use of implicit assumptions and inductive biases. Many of these, such as translational equivariance in CNNs or locality in audio diffusion models, are principled and empirically validated. Others, however, persist mainly for computational convenience[4] or historical precedent, even when their validity for specific applications remains unexamined or is blatantly known to be wrong (Zhao et al., 2018). As one example, the algorithms used in machine learning often assume that data are drawn independently. In reality, data points are often correlated, such as in time-series data or through repeated measurements from the same individual (Jiang & Nguyen, 2007; Kirchler et al., 2023). As another example, most generative models assume that latent distributions can be modeled on simple topological structures. However, latent distributions typically benefit from more expressive approaches (Stimper et al., 2022), suggesting the assumptions of their simplicity may be ill-founded.

We argue that convenience should not be the driving factor behind modeling assumptions. While the impact of model misspecifications on downstream applications in DGMs are not yet well understood, we have preliminary evidence suggesting their effects are undesirable: In traditional statistical analyses, such misspecifications are observed to have immense impacts (Cardon & Palmer, 2003); more recently, empirical studies have revealed systematic biases in deep generative models that may stem from inadequate modeling assumptions (Zhao et al., 2018), and models that rely heavily on the training data distribution have been observed to exhibit bias and decreased performance if not properly corrected by meaningful modeling assumptions (Fortuin, 2022).

There has been some progress towards developing methods that allow practitioners to encode more precise and complex modeling assumptions. As concrete examples, random effects (Jiang & Nguyen, 2007)—the paradigm used by traditional statistical methods to model data dependencies—have been adapted to work with neural models (Simchoni & Rosset, 2023). In normalizing flows, data dependencies can be incorporated directly into the likelihood objective (Kirchler et al., 2023), an approach that might be extended to other probabilistic approaches such as VAEs and diffusion models (Sutter et al., 2023). Causal models can also be integrated to directly model data dependencies and perform counterfactual inference (Pawlowski et al., 2020)—which we discuss in more detail in Section 2.2. Notably, these methods have yet to achieve widespread adoption, despite addressing issues that are prevalent and influential in many applications. We believe that further research into the effects of implicit modeling assumptions and methods that allow a wider range of modeling assumptions are promising and impactful directions for the field.

[4]By "convenience" we mean design choices adopted because they are easy to implement or tractable to optimize, not because they have been shown to match the true structure of the data.

Incorporation of Prior Knowledge. Recent major breakthroughs in deep generative models (DGMs) have primarily been achieved in settings where models could be trained on internet-scale data (OpenAI, 2023; Rombach et al., 2022). However, many real-world applications, such as drug design (Vamathevan et al., 2019), material engineering (Wei et al., 2019), personalized medicine (MacEachern & Forkert, 2021), and protein biochemistry (Bonetta & Valentino, 2020), often have much smaller datasets due to the high cost of data generation. In these areas, domain experts often possess troves of detailed prior knowledge, which could potentially be used to enable more data-efficient learning in generative AI models. Indeed, it has been shown in the context of VAEs that incorporating domain prior knowledge can significantly improve model performance (Fortuin et al., 2020; Jazbec et al., 2021) and even unlock their use for tasks that were previously impossible (Fortuin et al., 2019; Manduchi et al., 2021; 2022).

There are multiple routes via which prior knowledge can be encoded in generative AI systems (Dash et al., 2022). One straightforward way to incorporate domain knowledge is in Bayesian settings through the choice of prior distribution; such distributions can explicitly encode known properties of the target data. For example, an informed prior can reflect physiological constraints in medicine or chemical properties in materials science (Sam et al., 2024), taking a step towards ensuring that the learned model aligns with real-world principles. Beyond priors, domain knowledge can guide architectural design by suggesting specialized network components or hierarchical structures that reflect known relationships within the data (Andreas et al., 2016b; Shen et al., 2019; Bronstein et al., 2021) or can encourage models to process data in a more human-like manner for the sake of interpretability (Andreas et al., 2016a; McCoy et al., 2020; Vu et al., 2023; Lu et al., 2023). Finally, constraints embedded in either the model specification or the training algorithm can further ground generative models in real-world processes, leading to improved performance and trustworthiness (Raissi et al., 2019; Ren et al., 2020; Dash et al., 2021; Mohan et al., 2023). Each of these approaches can equip our models with helpful inductive biases that aid data-efficient learning.

While designing future models with domain-informed inductive biases holds great promise, it may not be so straightforward in practice. For example, while VAEs are Bayesian models and, therefore, offer a natural paradigm for specifying a prior distribution over their latent space, many other DGMs lack such explicit mechanisms for encoding prior information. We consider diffusion models as a concrete example. At first, it might seem that the diffusion process's Gaussian sampling distribution is comparable to the Gaussian latent prior in a VAE, suggesting a straightforward route for specifying priors for these models. However, this property of the diffusion process arises from the central limit theorem rather than from precise knowledge about the nature of the underlying data-generating distribution. Recent works have attempted to enhance the space of diffusion priors through auxiliary dimensions (Pandey & Mandt, 2023; Singhal et al., 2023). Unfortunately, neither of those approaches offers nearly the same flexibility of prior specification as the Bayesian priors in the latent space of VAEs, so further research into priors for diffusion models is sorely needed. In general, there are several caveats and hurdles that must be considered when designing future models with domain-informed inductive biases. First, by definition, biases constrain or push our models towards certain solutions. If the underlying bias does not capture every facet of the real-world process—an especially common concern in areas like biology, where core mechanisms remain poorly understood—it may inadvertently limit the model's expressivity or lead to systematic errors. In other words, it may cause models to fail to learn important patterns that fall outside the imposed structure. Moreover, adding constraints often introduces computational challenges: physically or biologically inspired restrictions might be non-differentiable or otherwise difficult to incorporate into standard training pipelines, leading to more complex optimization procedures or increasing computational overhead. Thus, while biasing models with domain knowledge can significantly improve data efficiency and performance, careful consideration of both the correctness of those biases and the technical feasibility of their implementation is essential.

2.4 Foundation Models for Heterogeneous Data Types

While there has been tremendous progress in large-scale foundation models for modalities like text and images, as the scope of application widens to encompass a broader range of data modalities, a variety of challenges emerge. These challenges are particularly pronounced in specialized fields such as healthcare and chemistry (Raghupathi & Raghupathi, 2014; Korshunova et al., 2022).

In healthcare, generation based on diverse data types—including imaging, health records, and genomics—poses challenges in interoperability, data privacy, and security (Moor et al., 2023a). Time series generation, in particular, requires addressing irregularly sampled data, missing values, seasonality, and long-term dependencies (Steinberg et al., 2021). In chemistry, physics, and chemical engineering, generative models have huge potential, not just for molecule, drug, and material design, but also in data augmentation, property prediction, and reaction prediction (Winter et al., 2019; Ahmad et al., 2022; Castro Nascimento & Pimentel, 2023; Hu et al., 2020). Data in these fields are often sparse, heterogeneous and correlated. On the other hand, they provide a vast body of physical and chemical domain knowledge, ranging from (strict) laws of nature and boundary conditions to (soft) empirical correlations and human experience. Therefore, developing hybrid (ML + domain knowledge) foundation models is a particular challenge. While there has been some recent progress toward this goal, e.g., in the realms of physics (Jirasek et al., 2022; Jirasek & Hasse, 2023; Howard et al., 2022) and medicine (Moor et al., 2023b; Xia et al., 2024), there is still much work to be done (Venkatasubramanian, 2019).

We argue that an overarching goal of the generative modeling field is to build general models, which can be applied to different applications spanning various data type, by integrating information from diverse sources and understanding complex relationships across different types of data (Li et al., 2023a; Reed et al., 2022; Driess et al., 2023). In downstream tasks that involve acting within the physical world, one way to operationalize such generality is through *embodied agents*, which may need to integrate perception, reasoning, and planning across vision and language, along with executing physical actions and environmental interactions. While datasets for natural language or images are relatively accessible (Hausknecht et al., 2020; Li et al., 2023b; 2024f), a comparable dataset for control tasks is lacking. Generative simulation is one route we identify to achieve this potential usecase for cross-domain generative models (Xian et al., 2023; Fan et al., 2022).

3 Optimizing Efficiency and Resource Utilization

Efforts to scale deep generative models (DGMs) for tasks like language modeling and text-to-image synthesis often involve training large models with billions of parameters, which demands significant computational resources. This leads to practical issues such as high energy costs (Wu et al., 2022) and expensive inference, limiting access for many users. This further raises environmental concerns due to the energy consumption required to fuel modern tensor processing hardware (Strubell et al., 2020). Training PaLM leads to 271 tons of CO2e effective emissions (Chowdhery et al., 2022) and training GPT-3 emits approximately double (Patterson et al., 2022). Therefore, there is a clear need to reduce the memory and computational requirements of large-scale DGMs to enhance accessibility and sustainability (Bender et al., 2021).

In this context, we discuss the efficiency-related challenges in current DGMs. We focus on minimizing training and inference costs (3.1), as well as highlighting challenges in designing evaluation metrics for DGMs (3.2), which greatly affect the computational resources needed for model selection and tuning.

3.1 Efficient Training and Inference

Network Architecture. Optimizing the network architecture, which forms the backbone of modern machine learning, is crucial for efficient training and inference in DGMs. While we have seen recent improvements in model quality (OpenAI, 2023; Touvron et al., 2023; Peebles & Xie, 2023), there is still a dearth of systematic comparative studies of architectural components' contributions to generative model performance. For instance, several popular LLMs like PaLM (Chowdhery et al., 2022) and Llama (Touvron et al., 2023) still largely reuse the original transformer architecture from Vaswani et al. (2017) with some additional modifications (Shazeer, 2020; Su et al., 2024; Zhang & Sennrich, 2019). A modification of particular importance has been that of the self-attention (Bahdanau et al., 2015) mechanism; in the original architecture, this operation incurred a computational cost that scaled quadratically in the context length. This made inference computationally expensive, especially for long-context modeling. Several recent works have proposed attention variants that provide faster inference times (Tay et al., 2022b). For example, Flash Attention (Dao et al., 2022) employs hardware optimizations and efficient memory management techniques to reduce the effective computational overhead of attention from quadratic to linear in the context length; Flash Atten-

Sub-challenge	**Typical failure mode**	**Promising mitigation avenues**	**Reference materials**	**Research-question candidates (paper §)**
Efficient attention mechanisms (§3.1)	Long contexts needed in certain settings and context length limited by computational demands	Hardware-aware attention computations (e.g., FlashAttention-2); Sub-quadratic attention alternatives	**Datasets:** LongBench **Surveys:** Tay et al. (2022a)	*Q1*: Can selective-state SSMs (e.g. Mamba) fully replace transformers for text generation? *Q2*: Can better retrieval methods in RAG systems mitigate the need for longer context windows in LLMs?
Low-bit quantization without quality loss (§3.1)	Sharp accuracy drop below 4-bit	Activation-aware weight quantization; Quantization-aware training and fine-tuning	**Surveys:** Zeng et al. (2025)	*Q1*: Do activation-aware quantization methods preserve model calibration after preference-based fine-tuning (e.g., RLHF)? *Q2*: What theoretical limits bound post-training quantization of diffusion models?
Fast sampling for diffusion models (§3.1)	Hundreds of network evaluations needed per sample	Progressive Distillation; Consistency Models; Model quantization	**Benchmarks:** FID/IS on CIFAR-10/ImageNet **Surveys:** Shen et al. (2025)	*Q1*: Can we design models that achieve one-step generation while maintaining diffusion models' training stability and sample quality?
Reliable quality metrics (§3.2)	Automatic evaluation metrics do not correlate with human perception of quality	Generative models for quality assessment (e.g., LLM-as-Judge, Auto-J); Sample-based metrics (e.g., Feature-Likelihood Score)	**Benchmarks:** JudgeBench **Surveys:** (Betzalel et al., 2024)	*Q1*: How can learned reward models be made reproducible enough to serve as public benchmarks? *Q2*: Is a unified multi-modal MAUVE variant feasible? *Q3*: How reliable are LLM-as-Judge scores in OOD settings?
Compute-aware model selection (§3.2)	Grid-search training runs are prohibitive for today's large models	Scaling-law extrapolation; Zero-cost proxies	**Benchmarks:** NAS-Bench-101; NATS-Bench **Surveys:** Li et al. (2024a) White et al. (2022)	*Q1*: Can information-theoretic complexity measures serve as reliable proxies for validation performance when direct evaluation on private data is prohibited? *Q2*: Can zero-cost proxies and scaling law extrapolations be effectively combined to provide a stronger indication of optimal models than their individual signals?

Table 2: Research challenges summary table - Optimizing Efficiency and Resource Utilization

tion 2 (Dao, 2024) takes these optimizations a step further, bringing attention computations close to the achievable bounds on fast matrix multiplication. Grouped Query Attention (Ainslie et al., 2023) proposes a structural change to the standard attention mechanism, where queries[5] are divided into distinct groups that are then processed independently and simultaneously. We see several promising research directions for reducing the computational needs of large-scale generative models, including specialized methods that make popular network architectures more computationally efficient (e.g., the attention variants discussed above), early-exit designs that allow models to make predictions without running the forward pass through the full network (Chen et al., 2024b) and alternative autoregressive sequence-modeling frameworks with favorable properties like scalability and linear complexity in the context length (Gu & Dao, 2023; Gu et al., 2021).

Similarly, several popular large-scale text-to-image diffusion models like DALL-E 2 (Ramesh et al., 2022) and StableDiffusion (Rombach et al., 2022) largely reuse the popular UNet (Ronneberger et al., 2015) backbone from Ho et al. (2020), which has high memory costs. Therefore, we believe that a principled study of the impact of different network components in large-scale generative models is crucial for efficient training and inference. Some recent works (Hoogeboom et al., 2023; Karras et al., 2023; Peebles & Xie, 2023; Podell et al., 2024) already explore architectural design choices for reducing diffusion model sizes, thereby improving training dynamics while enabling faster inference with a lower memory footprint.

[5]In the attention mechanism, a query is a transformed vector representation derived from input tokens.

Model Quantization. The goal of model quantization is to reduce the precision of model weights and activations, enabling faster, memory-efficient training and inference, ideally without losing performance on downstream tasks. The most common quantization approaches are Post-Training Quantization (PTQ), which applies quantization to a pre-trained large model to enable faster and memory-efficient inference, and Quantization-Aware Training (QAT), which involves training a quantized model from scratch (Krishnamoorthi, 2018).

Despite some progress in developing PTQ and QAT methods for LLMs (Dettmers et al., 2022; Liu et al., 2023b; Xiao et al., 2023; Yao et al., 2022; Dettmers et al., 2023) and large-scale text-to-image diffusion models (Li et al., 2023d), the existing methods are far from perfect. For instance, OPTQ (Frantar et al., 2023), a PTQ-based approach, can perform inference for a quantized LLM (in this case OPT (Zhang et al., 2022)) with 175B parameters on a single A100 GPU with 80GB of memory without degradation in accuracy. Though impressive, even this quantized model would likely have limited utility on a consumer-grade GPU device, let alone on standard edge devices. Similarly, QAT-based approaches can often achieve lower bitrates but trade off additional training for this efficiency. This can be a major computational bottleneck for large generative models. While some recent work suggests preliminary success in this direction (Lin et al., 2024a), we believe that investigating the impact of model quantization at low bitrates in large-scale generative models is a crucial direction for the practical deployment of these models.

Design Challenges. The current dominant modeling paradigms in generative AI, such as diffusion models (Ho et al., 2020) and LLMs (OpenAI, 2023), demonstrate remarkable sample quality. However, the design of the generative processes in these approaches can cause significant challenges. Diffusion models, for instance, rely on an iterative, multi-stage denoising process, which slows down inference considerably. Generating high-quality samples often requires hundreds to thousands of network function evaluations (NFEs) (Ho et al., 2020; Song et al., 2020). Similarly, LLMs employ an autoregressive structure that generates tokens sequentially, resulting in slow inference due to the left-to-right generation process. These challenges contrast with alternative generative models like VAEs and GANs, which require only a single NFE for sample generation. However, these models suffer from other drawbacks, such as blurry sample generation in VAEs (Dosovitskiy & Brox, 2016) and mode collapse in GANs (Arjovsky et al., 2017).

To address the inefficiencies in diffusion models, researchers have explored multiple complementary approaches to speed up inference. Some notable approaches include: developing training-free samplers (Song et al., 2021; Liu et al., 2022a; Lu et al., 2022; Zhang & Chen, 2023; Karras et al., 2022; Pandey et al., 2024), designing better diffusion processes (Singhal et al., 2023; Dockhorn et al., 2022; Pandey & Mandt, 2023; Karras et al., 2022), and combining other model families with diffusion models (Pandey et al., 2022; Zheng et al., 2023b; Xiao et al., 2022; Yang & Mandt, 2023). Additionally, training a diffusion model in the latent space of a lossy transform (Vahdat et al., 2021; Rombach et al., 2022) not only improves memory requirements and sampling efficiency but also provides access to a more interpretable low-dimensional latent representation. A lossy transform (such as VQ-GAN (Esser et al., 2021)) can drastically reduce data dimensionality while retaining the perceptually relevant details of high-resolution images. Designing more efficient lossy compression operations in the context of diffusion models has received less attention in the community and is an important direction for further work (Yang et al., 2023d; Havasi et al., 2019; Yang et al., 2020). Despite these advances, sampling from diffusion models remains computationally challenging, typically requiring 25-50 NFEs to generate high-quality samples. While approaches based on progressive distillation (Salimans & Ho, 2022; Meng et al., 2023) can further speed up inference, they trade off additional training for faster sampling. Therefore, there is a need for DGMs that inherit all the advantages of diffusion models while supporting one-step sample generation by design (e.g., see consistency models (Song et al., 2023; Song & Dhariwal, 2024) for recent work in this direction).

In the case of LLMs, in addition to an expensive self-attention operation in transformer-based autoregressive models, sequential token generation in a left-to-right fashion in these models makes inference more expensive. Indeed, it is one reason that (sub)word tokenization—the pre-processing of text into pre-defined units—is still an essential part of these pipelines. Notably, tokenization itself introduces a strong inductive bias into language modeling: the model is constrained to work with the predefined units set by the tokenization scheme. The representation of the data that the model learns is inherently shaped by these units, limiting

the model's flexibility. Token-free approaches have been proposed to allow for the joint optimization of text segmentation alongside other parameters, but in practice, they are often computationally infeasible with attention-based architectures because handling raw character sequences at scale magnifies the already expensive attention mechanism (Xue et al., 2022). Dynamic tokenization schemes (Ahia et al., 2024, e.g.,) present an interesting research direction, as they allow for predicting token boundaries at inference time, but they likewise can introduce significant computational overhead. Tokenization remains a key design choice that shapes both the performance and efficiency of modern generative models and whose further optimization is constrained by needs for efficiency (Rust et al., 2021; Toraman et al., 2023; Ali et al., 2024). Methods such as speculative decoding (Leviathan et al., 2022) are one approach that can help reduce the bottleneck caused by left-to-right generation. Non-autoregressive models also offer an interesting alternative for sequence modeling. For example, diffusion models amortize the computational cost of generating sequences across all tokens simultaneously (Dieleman et al., 2022; Wu et al., 2023; Li et al., 2022a). However, these models inherently lack the inductive bias for contextual generation, which has been shown to work well empirically for sequential modeling tasks. This affects their performance in downstream tasks that might require long-context modeling, such as video synthesis (Yang et al., 2023c; Ho et al., 2022a). While diffusion models can be incorporated within the autoregressive framework for such tasks, the resulting models can be very expensive during inference (due to the cost of synthesizing a single token using diffusion across multiple tokens). Therefore, we identify a potential tradeoff between long context modeling and efficient inference, with the diffusion and autoregressive modeling paradigms falling on the opposite ends of this tradeoff. Hence, designing generative modeling paradigms that can optimally balance this tradeoff remains challenging.

3.2 Evaluation Metrics

Evaluation metrics are crucial in guiding research, as the conclusions derived from empirical studies depend greatly on the chosen metrics. In modern ML, evaluation metrics are additionally a key component in hyperparameter tuning and model selection; their design thus affects computational resources required during large-scale training. However, designing robust and meaningful evaluation metrics for DGMs is challenging for several reasons.

Evaluation Metric Design. Many generative models are probabilistic, making likelihood-based metrics a seemingly natural choice for evaluating their performance. These metrics have been widely utilized in the literature due to their alignment with the probabilistic frameworks of such models. However, empirical evidence suggests that likelihood-based metrics often do not provide an accurate assessment of generation quality (Theis et al., 2016). In particular, they often fail to correlate with human judgments of sample quality (Kolchinski et al., 2019; Pimentel et al., 2023). Moreover, many popular generative models do not even allow for tractable likelihood computation. Other automatic evaluation metrics are thus necessary for evaluating the quality of generated samples.

Several notable evaluation metrics for generative models follow the general paradigm of comparing the distribution of generated samples to that of train/test samples (Sajjadi et al., 2018; Pillutla et al., 2024; Jiralerspong et al., 2023). For instance, the Fréchet inception distance (Heusel et al., 2017), which is widely used for evaluating image synthesis models, takes this approach (Salimans et al., 2016; Bińkowski et al., 2018). However, these metrics are far from perfect. First, robust computations of these metrics require a large set of samples (around 50k for image generation models). This can be computationally demanding for generative models with a sequential inference process, like diffusion and autoregressive models. Even given large sample sizes, these metrics have still demonstrated issues with robustness. For example, FID can be sensitive to minor perturbations in the input data (Parmar et al., 2022) (see Chong & Forsyth (2020) for additional discussion on sources of bias associated with FID and Borji (2022) for more related evaluation metrics). Second, these methods typically rely on an external pretrained model, e.g., the GPT-2 family of language models (Radford et al., 2019) or a classifier network trained on ImageNet (Deng et al., 2009). This property makes the metric effective for evaluating sample quality within the domain of the pretrained model's data but seemingly causes it to overlook significant features or overemphasize arbitrary ones in other domains (Kynkäänniemi et al., 2023; Pimentel et al., 2023).

Recent works have attempted to improve upon the above-mentioned shortcomings. For example, Jayasumana et al. (2024) propose the use of embeddings from the CLIP model (Radford et al., 2021), which aligns images and text in a shared embedding space, in order to make a more robust evaluation metric for image synthesis models. Several evaluation metrics for text generation systems use LLMs to score or rank samples, either via prompting (Li et al., 2024g) or explicit fine-tuning on the task of evaluation (Li et al., 2024e). These metrics correlate remarkably well with human judgments (Kocmi & Federmann, 2023) and offer more fine-grained assessments of text generation systems—providing feedback at the individual sample level and taking into account user-specified criteria (Jiang et al., 2024a). These methods make progress towards broader applicability across domains and greater alignment with human evaluations but also demonstrate an increased reliance on generative models for the evaluation of other generative models. This circular dependency introduces the risk of amplifying existing model biases (Fang et al., 2024) and narrowing the diversity of model-generated content (Doshi & Hauser, 2024; Gambetta et al., 2024), as the underlying paradigm of such evaluation metrics should lead to the favoring of outputs that align with the characteristics and biases of the models used for assessment. Some works have proposed (either explicitly or implicitly) rewarding sample diversity in evaluation metrics (Zhu et al., 2018; Alihosseini et al., 2019; Jiralerspong et al., 2023), which would alleviate the latter problem. However, there is often a quality-diversity trade-off (Caccia et al., 2019; Zhang et al., 2021; Naeem et al., 2020), where a model that generates high-quality samples might have low diversity across its samples and vice versa. Further, the quantification of diversity is in itself a difficult task (Tevet & Berant, 2021).

Subjective aspects in generation. A major challenge underlying the evaluation of generation quality is the subjective nature of sample attributes, such as realism, fluency, and style. While human inspection is typically the gold standard for evaluating generative models (Denton et al., 2015; Zhou et al., 2019; Saharia et al., 2022), in many cases, human judges disagree over several attributes, such as which samples have better quality (Clark et al., 2021) or what is considered realistic in the target domain (e.g., medical images or industrial optical inspection). This challenge is even present for conditional synthesis tasks when the set of suitable outputs is limited by constraints from the input. For example, in text-to-image generation (Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022), human evaluators may have different opinions on how closely a generated image aligns with the provided description.

For this reason, a common approach is to collect numerous human judgments and set up benchmarks based on these collective scores, e.g., the Open Parti Prompts Leaderboard[6] for image generator evaluation or TURINGBENCH (Uchendu et al., 2021) for language generator evaluation. While such approaches—along with attempts to standardize human evaluation practices (Elangovan et al., 2024)—help make human evaluations a more reliable signal for guiding the development of generative models, there are several other issues with human evaluation. These include its monetary costs, the general inconsistency of human raters (Clark et al., 2021; Belz et al., 2023), and its focus on individual samples, overlooking how well the generative model reflects the data distribution as a whole.

Evaluating Model Uncertainty and Calibration. Model uncertainty and calibration have become quantities of interest because of their implications for the reliability, interpretability, and safety of generative AI systems. Here, we use the term model uncertainty to refer to the degree of confidence a model has in its outputs; model calibration refers to the degree to which a model's estimated probability of an event is consistent with that event's true probability of occurring.[7] As concrete examples of the importance of these metrics, high model uncertainty in language generation tasks has been linked with the occurrence of hallucinations—instances where the model produces outputs that are implausible or factually incorrect (van der Poel et al., 2022; Zhang et al., 2023c); autonomous driving systems increasingly use generative models to predict the trajectories of other vehicles, cyclists, and pedestrians (Yuan et al., 2020), and miscalibration of the probabilities of such trajectories can lead to severe accidents.

Historically, relatively simple metrics have been employed for measuring uncertainty and calibration in machine learning. For example, Shannon entropy (Shannon, 1951) has been a common metric for quantifying total model uncertainty (Houlsby et al., 2011; Depeweg, 2019); expected calibration error (Pakdaman Naeini et al., 2015, ECE) has often been employed for assessing model calibration (Guo et al., 2017; Dormann,

[6] `https://huggingface.co/spaces/OpenGenAI/parti-prompts-leaderboard`

[7] We note that other definitions for the terms have been used.

2020) (see Abdar et al. (2021) and Wang (2023) for detailed surveys on uncertainty and calibration in deep learning, respectively). While these metrics are well-suited to models for simpler classification problems, their extension to generative models is non-trivial. For example, language models operate over a countably infinite output space (i.e., the set of all possible strings), making exact computation of metrics like entropy or ECE infeasible. Consequently, a key aspect of research on these characteristics in generative models has been on defining metrics that are suited to them (Zhao et al., 2021; Ran et al., 2022; Luo et al., 2023; Zhao et al., 2023b; Fei et al., 2023).

To complicate matters further, there is debate regarding which definitions of these metrics actually provide useful insights about a generative model. Model uncertainty, for instance, can come from multiple sources, such as aleatoric uncertainty (intrinsic noise in the data) or epistemic uncertainty (uncertainty in the model parameters) (Hüllermeier & Waegeman, 2021; Wimmer et al., 2023). Depending on the specific use case for a model, one may be interested in the contribution of only one of these sources rather than in total model uncertainty (Osband et al., 2023; Giulianelli et al., 2023; Kuhn et al., 2023). With respect to calibration, it is unclear exactly which distribution a generative model should be calibrated to (Koevering & Kleinberg, 2024); often times, we are more interested in modeling the distribution of high-quality outputs than the data-generating distributions (Ouyang et al., 2022), albeit the data from the latter is what models are often trained on (Kalai & Vempala, 2023). These choices must be carefully and thoughtfully considered, as they play a critical role in shaping the development of methods to quantify these metrics or address poor model performance in terms of these metrics.

Model Selection. Model selection, i.e., identifying which model configuration or set of hyperparameters will perform best on a given task, is essential in training large-scale generative models. Evaluation metrics play a critical role here. Using knowledge of scaling laws (Kaplan et al., 2020; Henighan et al., 2020), evaluation metrics can be used to predict early on in a training run whether a model is likely to be successful (OpenAI, 2023). Recent work has shown that these predictions can be done quite precisely (Ruan et al., 2024), potentially reducing the need to train numerous large neural networks in the search for a single good model. Zero-shot proxies (Abdelfattah et al., 2021)—metrics that estimate what the final performance of neural network architecture will be without training it—are another promising research direction for compute-efficient model selection. We also believe that more effort should be invested in analyzing models' *performance-complexity* tradeoff, an important yet under-investigated measure for real-world applications at scale. This tradeoff refers to the balance between model performance and computational complexity. We argue that model selection and evaluation should perhaps shift towards identifying the model families that lie in the associated Pareto set that optimizes this tradeoff (Devroye, 2010; Braverman, 2005; Chen et al., 2022; Braverman, 2023), as optimizing for these characteristics in isolation does not account for real-world constraints. The naïve approach—training well-performing models in each of the model classes under consideration and computing their respective computational complexities—is time- and resource-intensive. We posit that alternative assessments of complexity from information and learning theory (e.g., Xu et al., 2020) could provide more efficient substitutes for these types of evaluations.

Looking Forward. The multi-faceted nature of what defines a high-quality generative model makes designing robust and meaningful evaluation metrics a particularly challenging task. Instead of relying on human priors about what constitutes a good quantitative metric of model quality, developers have increasingly turned to the strategy of learning reward functions directly from human preferences (Ouyang et al., 2022). This approach should allow for evaluation metrics that are more aligned with human judgment, as the reward functions are directly informed by human feedback rather than predefined criteria. These reward functions could serve as the foundation for new evaluation frameworks for generative models, and we hope they will be open-sourced to enable the development of publicly accessible benchmarks.

Evaluation metrics can help us understand and ultimately mitigate model shortcomings. While this approach has been embraced for improving model quality, it also has significant potential to enhance model fairness, safety, and reliability. For instance, metrics designed to quantify various forms of bias can aid in identifying and addressing model unfairness. While such metrics exist for classification or regression models (e.g., demographic parity or equalized odds), their extension to generative models is non-trivial. Research is thus needed to develop and refine metrics that can effectively quantify biases in the complex outputs of generative

models.[8] This includes creating frameworks that account for the nuanced and context-dependent nature of generated content, ensuring these models are not only high-quality but also fair and aligned with ethical standards (Ray, 2023). Unfortunately, to be effective, these metrics must also be adaptable to closed-source generative models since parameters and logits of most commercial models are not publicly available (Zhao et al., 2023a; Sun et al., 2023a; Laszkiewicz et al., 2024).

Despite the availability of more advanced evaluation metrics, some domains continue to rely heavily on outdated automatic evaluation methods. For instance, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004)—metrics based on n-gram matching that are known to empirically correlate poorly with human judgments (Reiter, 2018; Deutsch et al., 2022)—remain extremely prominent in the evaluation of machine translation and abstractive summarization systems. The continued reliance on these inaccurate metrics may ultimately impede the advancement of generative AI, as they provide weak signals for model improvement and fail to guide the development of systems that truly align with human expectations and real-world applications. A shift towards new evaluation metrics requires a critical mass of adoption within the community. Therefore, we must encourage practitioners to move beyond the convenience of outdated metrics and embrace this new generation of improved metrics.

4 Ethical Deployment and Societal Impact

With the current excitement around the scope and application of large-scale generative models, we are also witnessing a growing apprehension, fueled by media reports, of adverse outcomes surrounding the rapid advancement of generative AI. These concerns add to the conceptual and practical considerations discussed so far and encompass a range of issues, including the spread of misinformation, the absence of regulatory frameworks (Meskó & Topol, 2023), unintended harm (Greenfield & Bhavnani, 2023), and debates over open-source versus closed-source technologies (Chen et al., 2023), among others. Here we identify key challenges concerning the responsible deployment of large-scale deep generative models. More specifically, we discuss several aspects, including the dissemination of misinformation (4.1), violation of privacy and copyright (4.2), presence of biases (4.3), lack of interpretability (4.4), and constraint satisfaction (Section 4.5).

4.1 Misinformation and Uncertainty

As the quality of generated data synthesized using large-scale generative models increases, it can become more and more difficult to distinguish between real and generated content, especially for uninformed consumers (Frank et al., 2024). This indistinguishability facilitates the spread of misinformation (e.g., by deepfakes (Helmus, 2022)). To ensure the trustworthiness of information, we need algorithmic solutions that are on par with the advances in generative models and allow us to robustly detect and mark synthetic data. Numerous models for differentiating machine-generated from real content have been proposed over the last years (Rana et al., 2022), but the increasing quality of generative model outputs has decreased their accuracy. Watermarking is another approach in which there has recently been increased interest. The goal of these methods is to manipulate a generated sample (e.g., an image or piece of text) such that a signature can be detected in downstream tasks albeit with minimal effects on sample quality. There have been several approaches to watermark synthetic data generated from LLMs (Kirchenbauer et al., 2023; Dathathri et al., 2024; Zhao et al., 2024) and image generation models (Zhao et al., 2023c; Wen et al., 2023; Jiang et al., 2024b). However, current watermarking methods are far from perfect (Saberi et al., 2023). Evasion is often possible by small manipulations (like paraphrasing or pixel perturbations) (Jiang et al., 2023) and the injecting of watermarks into content can be inefficient (Liu et al., 2024a). Recent work has demonstrated that model-integrated watermarks—i.e., signatures embedded in a model's sampling process rather than post-hoc in its outputs—are a promising route forward, as they can survive a range of real-world corruptions (e.g., paraphrasing attacks in text (Krishna et al., 2023) and latent-trajectory perturbations in images (Wen et al., 2023; Fernandez et al., 2023)). Nonetheless, the information-theoretic limits on detectability, false-positive rate, and adversarial removability for watermarking remain poorly understood, and initial negative results

[8] Many aspects of fairness cannot be captured by quantitative metrics. Further, definitions of fairness can differ amongst different people and groups, and these definitions may evolve over time. However, they can still provide insights into whether models achieve a certain level of fairness in specific aspects.

Sub-challenge	Typical failure mode	Promising mitigation avenues	Reference materials	Research-question candidates (paper §)
Misinformation & synthetic media detection (§4.1)	Deepfakes bypass detectors	Model-rooted watermarks (e.g., Tree-Ring Watermarks)	**Datasets:** Deepfake Detection Challenge ; WaterBench **Surveys:** Rana et al. (2022)	*Q1*: Can diffusion-time watermarking survive multimodal adversarial attacks? *Q2*: Can uncertainty-aware abstention policies mitigate hallucinated facts in DGM outputs?
Privacy, copyright infringemen (§4.2) & PII leakage	Model regenerates training snippets	Differentially Private learning techniques(e.g., DP-SGD); Machine Unlearning methods	**Datasets:** CPDM; PrivLM-Bench **Surveys:** Yao et al. (2024) Kibriya et al. (2024)	*Q1*: What privacy-preserving fine-tuning strategies remain feasible for $\geq$ 100B-parameter models under practical compute budgets?
Fairness across languages (§4.3)	Worse compression → more tokens needed → higher cost for low-resource languages	Dynamic tokenization schemes; Vocabulary transfer methods	**Surveys:** Xu et al. (2025) Qin et al. (2025)	*Q1*: Can subword-free sequence modeling architectures (e.g., SSMs) eliminate tokenization-induced performance disparities across languages?
Bias & discrimination in generated content (§4.3)	Models reflect and propagate bias and discrimination present in training data	Counterfactual evaluation benchmarks; Fairness metric (e.g., demographic parity) incorporation into training objectives	**Datasets:** HolisticBias; OpenBias **Surveys:** Gallegos et al. (2024)	*Q1*: How do watermarking methods interact with demographic bias?
Interpretability & transparency (§4.4)	DGMs are blackboxes with uninterpretable parameters	Automated circuit discovery; Causal tracing	**Datasets:** GPT-2 neuron-explanation dataset **Surveys:** Marcinkevics & Vogt (2020)	*Q1*: Which confidence metrics best predict whether a mechanistic explanation truly modulates model behaviour? *Q2*: Can mechanistic insights be transferred across model sizes?
Constraint satisfaction (§4.5)	Outputs violate hard rules (e.g. code won't compile)	Context Free Grammar-guided decoding algorithms; Constrained RLHF; Constraint-embedded model architectures	**Benchmarks:** HumanEval; BigCodeBench **Surveys:** Zhang et al. (2023a)	*Q1*: How can hard-constraint decoding be generalised from code to text and images? *Q2*: Do constraint-aware training methods (e.g., constrained RLHF) achieve better safety-performance trade-offs than inference-time constraint enforcement (i.e., constrained decoding)?

Table 3: Research challenges and mitigation strategies - Ethical Deployment and Societal Impact

suggest unavoidable trade-offs between robustness, the quality of the altered sample, and watermark payload (Kirchenbauer et al., 2024; Yoo et al., 2024).

Notably, misinformation can emerge even without malicious intent. Tools like ChatGPT are increasingly expected to serve as universal question-answering engines, even though their core objective—to estimate the likelihood of the next token in a sequence—is traditionally designed to assess the linguistic plausibility of strings (Kalai & Vempala, 2023), rather than their factual accuracy. This distinction between the two objectives is evinced by the discrepancies observed between the probability a model explicitly assigns to a statement when prompted vs. the underlying log-probability it assigns (Hu & Levy, 2023), models' tendencies to hallucinate (Huang et al., 2024a)), and their difficulty in achieving probabilistic consistency (Elazar et al., 2021), e.g., ensuring logical predictions between a statement and its negation.

Some works have turned to model uncertainty estimates (e.g., those discussed in 3.2) as indicators of model reliability, developing methods to enhance the trustworthiness of AI systems based on these estimates (Edupu-

ganti et al., 2021; Yang et al., 2023e). For example, Ren et al. (2023) propose a selective generation approach, where models abstain from providing a response in the face of high uncertainty; Kuhn et al. (2023) use a notion of a model's semantic uncertainty to predict the correctness of its answer in question-answering. The development of methods that explicitly account for uncertainty represents an interpretable approach toward ensuring model reliability, offering a strategy that also has grounding in a well-studied concept in machine learning (Malinin & Gales, 2018; Abdar et al., 2021; Gawlikowski et al., 2023).

Encouragingly, some research suggests that larger LMs actually are well-calibrated in terms of their world knowledge, i.e., their predicted likelihoods reflect the probability that a statement is true (Srivastava et al., 2022; Zhu et al., 2023; Yu et al., 2024). Further, recent studies show that language models often do possess the ability to assess the truthfulness of their own statements (Lin et al., 2022; Kadavath et al., 2022; Xiong et al., 2024). However, fine-tuning or RLHF, which are frequently applied to these models, have been shown to hurt calibration; rather, they have been widely observed to exhibit overconfidence—the tendency of a model to assign excessively high probabilities to its predictions regardless of their correctness (Kadavath et al., 2022; Tian et al., 2023; OpenAI, 2023; Xiong et al., 2024). There has been some work on mitigating miscalibration issues for fine-tuning (e.g., Wang et al., 2023b) and RLHF (e.g., Tian et al., 2023; Zhang et al., 2024). and there is an increasing focus on systems where evidence can be brought in from external knowledge sources (Blattmann et al., 2022; Pan et al., 2023; Gao et al., 2023), grounding model responses to reliable knowledge sources. Such research—along with benchmarks to assess model factuality (e.g., KoLA; Yu et al., 2024)—is a critical step towards ensuring the trustworthiness and reliability of generative models.

4.2 Security, Privacy and Copyright Infringement

While modern generative models like LLMs are deployed practically for many applications, this also exposes them to potential malicious attacks, which can have significant costs for downstream applications or users. One class of malicious attacks on LLMs is the so-called "Backdoor attacks" where the main idea is to train the model using *poisoned data* and then trigger a specific output response from the model corresponding to specific prompts. For instance, Yang et al. (2023a) discusses backdoor attacks on LLMs in the context of communication networks (see Zhou et al., 2025, for a more in-depth exploration of backdoor attacks). Another class of attacks known as "jail-breaking" involves designing adversarial prompts to generate malicious outputs from the model while bypassing guardrails employed to comply with usage policies. There has been a good deal of research in the context of LLMs exploring techniques for jail-breaking and for guarding against jail-breaking (Shen et al., 2024; Yi et al., 2024; Jin et al., 2024). In this work, we primarily focus on user privacy and copyright infringement in the context of such malicious attacks.

Recent works have shown that publicly available LLMs and large-scale text-to-image models can implicitly "memorize" training data (van den Burg & Williams, 2021), to the point that samples from the dataset can be (almost exactly) reconstructed (Carlini et al., 2023a;b; Somepalli et al., 2023; Nasr et al., 2023). This behavior potentially infringes on data privacy, underscoring the importance of detecting whether private information has been leaked into an LLM's training data (Kim et al., 2023) and exploring whether generative models can be trained while safeguarding sensitive information. Differential privacy (DP) constraints, which can be enforced during generative model training, offer an attractive theoretical framework to ensure privacy Dwork & Roth (2014); Li et al. (2021); Dockhorn et al. (2023). However, DP-based approaches have several shortcomings. They suffer from a trade-off between privacy and utility (Cummings et al., 2024). There are different DP formulations, each based on different assumptions about trust, data access, and the point at which noise is introduced. The meanings of the canonical DP parameters are thus not consistent, making comparison of models produced using different approaches difficult (Li et al., 2024d). Moreover, in the context of image generation, scaling such approaches to high-resolution datasets remains elusive. Recent work has instead focused on generative DP synthetic data, where foundation models are only used as blackboxes (Lin et al., 2024b). Building privacy constraints into the training of large-scale generative models can be a promising direction for further research.

Another byproduct of memorization in generative models is that it can lead to unauthorized distribution or replication of training data, resulting in copyright infringement liabilities.[9] Current efforts towards preventing

[9]https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

such behavior have approached the problem from different angles. Some focus on filtering training data: techniques such as contractual licensing filters and hash-based de-duplication allow developers to identify and exclude protected material before training (Carr & Jeffrey, 2022; Duarte et al., 2024). Plagiarism or style-clone detectors allow intervention downstream, flagging generated samples that are substantially similar to copyrighted works (Li et al., 2024c; Kim et al., 2021). Other works have proposed training methods, which can be broadly categorized under two complementary strategies: (i) *imitation-resistant training objectives* that discourage verbatim memorization (Liang et al., 2023; Zhao et al., 2023d), and (ii) *machine unlearning* methods that aim to eliminate the influence of certain datapoints (e.g., copyrighted or private material) from a model's predictions after training (Li et al., 2024b; Liu et al., 2024b; 2025). There is still much open research in this domain, including provenance tracking that can scale to work with modern, massive datasets and copyright infringement risk metrics that encompass legal concerns.

Beyond privacy and copyright concerns, indiscriminate memorization has other undesirable effects. Many of the current target applications for generative models—such as creative writing or graphics generation—demand novelty and user-specific adaptation; when a model merely regurgitates training data, it does not provide the desired diversity, originality and personalization. This behavior potentially worsens user experience and limits the practical value of generative models. Further, excessive memorization can lead to biases in generated content, where certain perspectives or styles dominate because they were overrepresented in the training data; we discuss this last issue in more detail in the next section. Going forward, preventing memorization in DGMs is an important area of research to focus on for the safe and effective deployment of generative models (Chen et al., 2024a).

4.3 Fairness

Large-scale generative models are often trained on massive datasets containing billions of samples scraped from the internet. While preprocessing such large datasets often involves tagging or removing toxic content, a variety of other societal biases are often harder to detect. Consequently, the trained models can reflect biases and produce outputs that may be deemed toxic or harmful (Pagano et al., 2023; Gallegos et al., 2024; Zhou et al., 2024). For instance, Weidinger et al. (2021) outline a series of harms that can result from using LLMs that produce discriminatory or exclusionary language, e.g., the amplification of stereotypes or exclusionary norms. Multimodal models may exhibit biases about gender, ethnicity, and religion, among others (Janghorbani & De Melo, 2023) that have similar negative effects on society.

Not all unfair behaviors exhibited by generative models are as overtly harmful as generating toxic content. Models can show more subtle biases towards certain subpopulations, such as allocation bias—when AI systems extend or withhold opportunities, resources, or information—or quality-of-service bias—when AI systems work better for people in some subpopulations than others; these behaviors should not be downplayed as they can perpetuate systemic inequalities. A main culprit of such behaviors stems from the fact that the data used in the training of most generative models is disproportionately from certain countries and languages. Such models therefore might not work as well for languages or images outside of these mainstream groups. For example, multilingual language models typically perform substantially better on English tasks compared to tasks in other languages (Lai et al., 2023; Huang et al., 2023a). Further, because the tokenizers for such models have also been trained disproportionately using English data, the compression rate for English texts is much higher than for texts in low-resource languages (Petrov et al., 2023; Ahia et al., 2023). Consequently, services that charge based on token counts impose higher costs for a query made in a low-resource language compared to a query with the same underlying meaning made in English. This exacerbates accessibility challenges for speakers of underrepresented languages.

Numerous approaches have been proposed to mitigate the biases of generative models in a post-hoc manner (Bai et al., 2022; Glaese et al., 2022; Ferrara, 2024; Olmos et al., 2024). However, the achieved changes are often merely superficial, leaving the possibility of remnant biases. For example, Gonen & Goldberg (2019) demonstrate that word embeddings still cluster based on gender stereotypes even after bias mitigation techniques, effectively "hiding" rather than eliminating the issue. In the vision domain, post-hoc bias mitigation strategies have been observed to work poorly in the face of test-time distribution shift (Kong et al., 2023). Moreover, most evaluations assess only a *single* fairness axis—for instance, gender in English or skin-tone in photos—and residual biases along other dimensions can remain undetected. Some key research areas that

we identify as needing further attention include: (i) joint evaluation across multiple, potentially interacting fairness criteria (e.g., gender × dialect); (ii) stress-testing mitigation strategies under data-distribution shifts; and (iii) training-time interventions that prevent harmful biases from emerging in the first place. Some areas that we identify as needing further research are combined evaluation with respect to multiple forms of fairness criteria, robust assessments across multiple domains and training methods that can more robustly mitigate the learning of harmful biases. Promising steps in these directions include multilingual, multi-attribute benchmark suites such as XCOPA-Bias (Goldfarb-Tarrant et al., 2023) and gradient-based de-biasing schedules that adjust sampling weights during pre-training (Kim et al., 2024).

Ultimately, assessing and ensuring fairness in technology applications is a complex challenge. Aside from the aforementioned issue of differing (and potentially dynamic) qualitative definitions of fairness (3.2), different notions of fairness often cannot be fully satisfied simultaneously (Ferrara, 2024). Therefore, it is essential for the builders of generative AI tools to carefully evaluate the various dimensions of fairness and make deliberate trade-offs appropriate for the specific usecase.

4.4 Interpretability and Transparency

In high-stakes applications such as healthcare and legal domains, it is critical to understand the logic and influencing factors behind generative models' outputs. In other words, we need to be able to *interpret* how a generative model produces its results, with its decision-making process being *transparent*, i.e., accessible and understandable This is particularly true in safety-critical domains—such as healthcare (Chen et al., 2021a) or finance applications—but is also important across general AI use cases, where interpretability is essential for diagnosing errors and fostering user trust. These needs are not new and have been present since the start of publicly-available AI-based products and tools (Confalonieri et al., 2021). There has thus been a sizable amount of research in neural network interpretability methods. However, these methods are not always feasible for use with large-scale DGMs. For example, interpretability methods, such as SHAP, LIME, or Integrated Gradients (Lundberg & Lee, 2017; Ribeiro et al., 2016; Sundararajan et al., 2017), struggle to scale effectively with the complexity and size of large models; many interpretability methods work by attempting to understand concepts encoded in models' latent representations (Crabbe et al., 2021; Esser et al., 2020), but this becomes more difficult in the high-dimensional latent spaces used by DGMs. Further, while one might hope that explanations derived from interpreting smaller models could be used for understanding their larger counterparts, scaling up generative models gives rise to unpredictable effects, e.g., models demonstrating unexpectedly advanced capabilities (Wei et al., 2022; Schaeffer et al., 2024); conclusions drawn when using small models therefore may not be applicable to today's larger models.

The fundamental challenge is to develop explanation methods for DGMs that are both well-understood by humans and faithful to the underlying model behaviors (Schut et al., 2023; Gurnee & Tegmark, 2024). *Mechanistic interpretability* is a field that attempts to achieve this goal by reverse-engineering neural network decisions, translating them to human-interpretable decision-making processes (Bereska & Gavves, 2024). This is specifically done by analyzing models at the level of their internal computations, representations, and structural components, e.g., identifying minimal subnetworks (referred to as circuits) that implement a specific computation. A large appeal of mechanistic interpretability is that it provides causal explanations for models' outputs, i.e., a decision or prediction can be attributed in a causal manner to some component of the input. For example, causal tracing—a prominent tool in this line of work—perturbs internal activations in a manner that allows us to determine whether they *cause* a change in the output. This allows researchers to move beyond correlation-based explanations and instead understand the actual computational mechanisms driving observed model behavior, which in turn enables more precise debugging, bias detection, and control over generative outputs. Mechanistic interpretability research has uncovered a number of interesting and useful properties of DGMs. For example, specific neurons or layers in GANs and VAEs encode "disentangled" (i.e., distinct and interpretable) features, such as shape, texture, or pose in images (Shen et al., 2020; Mita et al., 2021). Other works have found that certain attention heads in transformers correspond to meaningful linguistic patterns, e.g., some might focus on syntactic structure while others might capture semantic information (Vig & Belinkov, 2019; Elhage et al., 2021). Such properties not only help practitioners better understand DGMs, they also enable them to control generation to some extent (Härkönen et al., 2020; Fetty et al., 2020).

However, the reliability and comprehensiveness of mechanistic interpretability methods remain a subject of debate (Golechha & Dao, 2024a; Sharkey et al., 2025). A central criticism is that, while these techniques aim to identify causal relationships between model components and outputs, they may only provide partial or even misleading insights into the actual computational processes at work. For instance, attention patterns—which have been used by various methods to attribute model predictions to certain tokens (Xu et al., 2015; Choi et al., 2016, *inter alia*)—do not always faithfully reflect how or why certain tokens influence the final prediction (Jain & Wallace, 2019; Liu et al., 2022b). In some cases, they may merely highlight correlations rather than reveal deeper causal structures. Similar concerns have been raised about other methods for explaining model behaviors from mechanistic interpretability. Further, the new wave of large-scale generative models makes the application of some of these methods more difficult: the larger a model becomes, the (arguably) more difficult it becomes to fully reverse-engineer a prediction, both from a theoretical and computational standpoint. Every nuance of these models' decision-making may not be deducible from a subset of neurons, layers, or attention heads, and interpretations derived from one subset of model components may overlook equally critical interactions elsewhere in the large network.

Representation engineering (Zou et al., 2023a) and the use of sparse autoencoders (Bricken et al., 2023) are lines of research in mechanistic interpretability that potentially address the former issues, offering interpretable explanations even for large-scale models. Efficient methods for circuit identification have started to address the latter issue (Hsu et al., 2025). However, we are still in need of validation methods that can confirm whether the identified "mechanisms" truly govern a model's outputs, or whether they merely reflect convenient, yet incomplete, narratives about its internal workings.

Going forward, researchers should continually assess user needs for explainability to ensure that the appropriate objectives are guiding the development of interpretability methods (Liao et al., 2020; Wang & Yin, 2021; Poursabzi-Sangdeh et al., 2021). Attention must also be paid to the relevance and effectiveness of the metrics and evaluation frameworks used to assess these methods (Ross et al., 2021; Jethani et al., 2021). Another important research direction is enhancing the robustness of explainable methods, such as counterfactual explanations (Wachter et al., 2017; Slack et al., 2021). Further research is also needed for the new wave of multimodal models, as existing explainability methods may not be equipped to offer explanations in the face of cross-modal interactions.

4.5 Constraint Satisfaction

Generative models such as ChatGPT are used by millions of people and deployed across diverse use cases. Many applications require generative models to satisfy domain-specific constraints.[10] In some cases, these merely stem from a desire to have a more controlled form of generation, such as when a generated image is conditioned on a given depth map (Zhang et al., 2023b). In other cases, ethical and safety considerations are key concerns. For instance, in fields like engineering design, generative model outputs must meet engineering standards and adhere to laws of nature (i.e., physics). More generally, there are widespread calls for generative models to avoid toxicity, bias mitigation and other outputs that may lead to harmful effects (Weidinger et al., 2021), e.g., by refraining from responding in ways that could pose a risk to the mental health of human interlocutors and by refusing to carry out tasks related to illegal activities. While reinforcement learning from human feedback (RLHF; Ouyang et al., 2022)—and particularly *constrained* RLHF (Moskovitz et al., 2024)—has offered an initial step towards these goals, enabling companies and users to provide models with soft constraints within their queries, such constraints can be circumvented (Shen et al., 2023). Ultimately, methods that allow us to place hard constraints on model outputs are necessary.

In language generation, decoding methods that allow for arbitrary constraints (both hard and soft) have been a focus area of the research community (Kumar et al., 2021; 2022). There are several prominent challenges in the development of such methods, including: efficiently enforcing constraints without significantly increasing computational costs, maintaining fluency and coherence while adhering to constraints, and handling multiple constraints, which may result in conflicting requirements. The discrete nature of text presents a particular difficulty, as small changes to token sequences can drastically alter meaning, making it difficult to optimize

[10]Here we focus on constraints specified at inference time. We discuss constraints that must be integrated into the model during training in 2.3.

for constraints from a computational perspective. Recent grammar-constrained decoding methods (Beurer-Kellner et al., 2024; Ugare et al., 2024) address these issues by guaranteeing that every generated sequence conforms to a user-supplied context-free grammar, achieving hard constraints with only a modest runtime overhead. Such methods have started to gain traction in other domains, e.g., in healthcare (Golechha & Dao, 2024b). Meanwhile, research on controllable image generation has likewise gained momentum (Deng et al., 2020; Huang et al., 2024b), with various approaches aiming to regulate attributes such as style, composition, or specific content elements. Methods range from applying spatial constraints (e.g., bounding boxes, masks, or layout specifications (Zheng et al., 2023a)) to enforcing semantic conditions (e.g., ensuring that certain objects or visual features are present (Pavllo et al., 2020)). Technical challenges arise in this domain as well, such as the need for more complex conditioning mechanisms and heavier computational demands—especially when constraints must be integrated at each step of the generative process in e.g., diffusion models. Code generation stands out as a domain where constraint enforcement has long been a primary focus (Poesia et al., 2022; Dong et al., 2023). Here, the constraints—such as following a language's syntax and producing compilable code—are arguably more well-defined and straightforward to verify. Concepts and methods from this field could conceivably help research in enforcement of hard constraints in other generative AI fields, e.g., in the application of generative AI to the physical domain, where laws of nature must be satisfied. Overall, methods to ensure effective constraint adherence can substantially improve control over generative models, which is crucial for ensuring their safe and reliable deployment (Regenwetter et al., 2024). Despite progress across various generative AI fields, the development of scalable and generalizable techniques capable of handling the diverse and often conflicting demands of real-world constraints remains an open challenge. We encourage further research in this area to bridge this gap and advance the controllability of generative models across different domains.

5 Conclusion

Despite the recent excitement and hype surrounding advancements in generative AI, the goal of achieving a *perfect* generative model remains far from reality. In the hopes of eventually reaching this goal, we identified several core challenges with the current generative modeling frameworks and practices.

We identified generalization and robustness as major hurdles, demanding methods to handle unseen data and adversarial attacks. Moreover, limited representational power and implicit modeling and data assumptions necessitate exploring more expressive models and incorporating prior knowledge, especially in data-scarce scenarios. Moving beyond merely identifying correlation, integrating causal reasoning into DGMs holds promise for enhanced interpretability and robustness. We also emphasized the escalating computational demands and associated barriers to the widespread adoption of DGMs. Training and inference inefficiencies pose key challenges, urging the exploration of alternative network architectures and low-bitrate model quantization. Inadequate evaluation metrics for generated content, such as FID and n-gram matching, hinder efficient progress, prompting the search for robust, domain-agnostic alternatives.

We also discussed the challenges and considerations surrounding the responsible deployment of large-scale generative models, pointing to the rising concerns related to misinformation, unintended harm, and lack of trustworthiness. Challenges include combating misinformation, ensuring privacy in data curation, addressing fairness issues, enhancing interpretability, estimating model uncertainty, and satisfying constraints.

By confronting the limitations discussed here, we can transform DGMs from data replicators (Bender et al., 2021) to tools with transformative capabilities across various domains. We hope that our paper will point to directions that ultimately contribute to these goals.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76(C):243–297, December 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008. URL `https://doi.org/10.1016/j.inffus.2021.05.008`.

Mohamed S. Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D. Lane. Zero-Cost Proxies for Lightweight NAS. In *International Conference on Learning Representations (ICLR)*, 2021.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A. Smith. MAGNET: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=1e3MOwHSIX.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pp. 372–407. PMLR, 2023.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. Tokenizer choice for LLM training: Negligible or crucial? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3907–3924, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.247. URL https://aclanthology.org/2024.findings-naacl.247/.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models. In Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf (eds.), *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 90–98, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2311. URL https://aclanthology.org/W19-2311/.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1545–1554, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1181. URL https://aclanthology.org/N16-1181/.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48, 2016b. doi: 10.1109/CVPR.2016.12.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

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

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=hMGSz9PNQes.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

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 arXiv:2212.08073*, 2022.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3676–3687, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.226. URL https://aclanthology.org/2023.findings-acl.226/.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Eyal Betzalel, Coby Penso, and Ethan Fetaya. Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction*, 6(3):1531–1544, 2024. ISSN 2504-4990. doi: 10.3390/make6030073. URL https://www.mdpi.com/2504-4990/6/3/73.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Guiding llms the right way: fast, non-invasive constrained generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Rosalin Bonetta and Gianluca Valentino. Machine learning techniques for protein function prediction. *Proteins: Structure, Function, and Bioinformatics*, 88(3):397–413, 2020.

Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Mark Braverman. On the complexity of real functions. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pp. 155–164. IEEE, 2005.

Mark Braverman. *Communication and information complexity*, pp. 284–320. EMS Press, December 2023. ISBN 9783985475599. doi: 10.4171/icm2022/208. URL http://dx.doi.org/10.4171/icm2022/208.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. In *International Conference on Learning Representations*, 2019.

Lon R Cardon and Lyle J Palmer. Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604, 2003.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, 2023a.

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, 2023b.

SR. Carr and N. Jeffrey. Class Action Complaint. *Sarah Anderson, et al., v. Stability AI LTD., et al*, 2022.

Cayque Monteiro Castro Nascimento and André Silva Pimentel. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655, 2023. doi: 10.1021/acs.jcim.3c00285. URL https://doi.org/10.1021/acs.jcim.3c00285. PMID: 36926868.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8425–8434, June 2024a.

Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq R. Joty. Chatgpt's one-year anniversary: Are open-source large language models catching up? *ArXiv*, abs/2311.16989, 2023.

Irene Y Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *Annual review of biomedical data science*, 4:393–415, 2021a.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021b.

Sitan Chen, Jerry Li, and Yuanzhi Li. Learning (very) simple generative models is hard. *Advances in Neural Information Processing Systems*, 35:35143–35155, 2022.

Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. Ee-llm: large-scale training and inference of early-exit large language models with 3d parallelism. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3512–3520, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=9OfKxKoYNw`.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. C2l: Causally contrastive learning for robust text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10526–10534, Jun. 2022.

Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6070–6079, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:240:1–240:113, 2022.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1550–1553, 2019. doi: 10.1109/ICDE.2019.00139.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing, pp. 4067–4080. Association for Computational Linguistics, 2019.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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. 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565/.

Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. A historical perspective of explainable artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1):e1391, 2021. doi: https://doi.org/10.1002/widm.1391. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391.

Jonathan Crabbe, Zhaozhi Qian, Fergus Imrie, and Mihaela van der Schaar. Explaining latent representations with a corpus of examples. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12154–12166. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/65658fde58ab3c2b6e5132a39fae7cb9-Paper.pdf.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, and Wanrong Zhang. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review*, 6(1), jan 16 2024. https://hdsr.mitpress.mit.edu/pub/sl9we8gh.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Tirtharaj Dash, Ashwin Srinivasan, and Lovekesh Vig. Incorporating symbolic domain knowledge into graph neural networks. *Machine Learning*, 110(7):1609–1636, Jul 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05966-z. URL https://doi.org/10.1007/s10994-021-05966-z.

Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, Jan 2022. ISSN 2045-2322. doi: 10.1038/s41598-021-04590-0. URL https://doi.org/10.1038/s41598-021-04590-0.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, Oct 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08025-4. URL https://doi.org/10.1038/s41586-024-08025-4.

Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal VAEs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w-CPUXXrAj.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5153–5162, 2020. URL https://api.semanticscholar.org/CorpusID:216144533.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, 28, 2015.

Stefan Depeweg. *Modeling Epistemic and Aleatoric Uncertainty with Bayesian Neural Networks and Latent Variables*. PhD thesis, Technische Universität München, 2019. URL https://mediatum.ub.tum.de/1482483.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Daniel Deutsch, Rotem Dror, and Dan Roth. Re-examining system-level correlations of automatic summarization evaluation metrics. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6038–6052, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.442. URL https://aclanthology.org/2022.naacl-main.442/.

Luc Devroye. Complexity questions in non-uniform random variate generation. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 3–18. Springer, 2010.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations*, 2022.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Yihong Dong, Ge Li, and Zhi Jin. Codep: Grammatical seq2seq model for general-purpose code generation. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2023, pp. 188–198, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702211. doi: 10.1145/3597926.3598048. URL https://doi.org/10.1145/3597926.3598048.

Carsten F. Dormann. Calibration of probability predictions from machine-learning and statistical models. *Global Ecology and Biogeography*, 29(4):760–765, 2020. doi: https://doi.org/10.1111/geb.13070. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.13070.

Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290. URL https://www.science.org/doi/abs/10.1126/sciadv.adn5290.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

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. In *International Conference on Learning Representations*, 2020.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

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*, 67(1):110–120, 2023a.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan. Robustness challenges in model distillation and pruning for natural language understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1758–1770, 2023b.

André Vicente Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-COP: Detecting copyrighted content in language models training data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=LO4xhXmFal.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7887–7896, 2020. URL https://api.semanticscholar.org/CorpusID:211988680.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep mri reconstruction. *IEEE Transactions on Medical Imaging*, 40(1):239–250, 2021. doi: 10.1109/TMI.2020.3025065.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. 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. 1137–1160, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.63. URL https://aclanthology.org/2024.acl-long.63/.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 12 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00410. URL https://doi.org/10.1162/tacl_a_00410.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1): 5224, Mar 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-55686-2. URL https://doi.org/10.1038/s41598-024-55686-2.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14014–14031, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.783. URL https://aclanthology.org/2023.acl-long.783/.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22409–22420, 2023. doi: 10.1109/ICCV51070.2023.02053.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1), 2024. ISSN 2413-4155. doi: 10.3390/sci6010003. URL https://www.mdpi.com/2413-4155/6/1/3.

Lukas Fetty, Mikael Bylund, Peter Kuess, Gerd Heilemann, Tufve Nyholm, Dietmar Georg, and Tommy Löfstedt. Latent space manipulation for high-resolution medical image synthesis via the stylegan. *Zeitschrift für Medizinische Physik*, 30(4):305–314, 2020. ISSN 0939-3889. doi: https://doi.org/10.1016/j.zemedi.2020.05.001. URL https://www.sciencedirect.com/science/article/pii/S0939388920300544.

Vincent Fortuin. Priors in Bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022.

Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. SOM-VAE: Interpretable discrete representation learning on time series. In *International Conference on Learning Representations*, 2019.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1651–1661. PMLR, 2020.

Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 159–159. IEEE Computer Society, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

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. *Computational Linguistics*, 50(3):1097–1179, 09 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.

Daniele Gambetta, Gizem Gezici, Fosca Giannotti, Dino Pedreschi, Alistair Knott, and Luca Pappalardo. A linguistic analysis of undesirable outcomes in the era of generative AI, 2024. URL https://arxiv.org/abs/2410.12341.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. URL https://api.semanticscholar.org/CorpusID:266359151.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(1):1513–1589, Oct 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10562-9. URL https://doi.org/10.1007/s10462-023-10562-9.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. What comes next? evaluating uncertainty in neural text generators against human production variability. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14349–14371, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.887. URL https://aclanthology.org/2023.emnlp-main.887/.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4458–4468, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.272. URL https://aclanthology.org/2023.findings-acl.272/.

Satvik Golechha and James Dao. Challenges in mechanistically interpreting model representations. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024a. URL https://openreview.net/forum?id=wfemKUcgoB.

Satvik Golechha and James Dao. Challenges in mechanistically interpreting model representations. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024b. URL https://openreview.net/forum?id=wfemKUcgoB.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://aclanthology.org/N19-1061.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

David Greenfield and Shivan Bhavnani. Social media: generative AI could harm mental health. *Nature*, 617 (7962):676–676, May 2023. doi: 10.1038/d41586-023-01693-.

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

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

Wenjuan Han, Bo Pang, and Ying Nian Wu. Robust transfer learning with pretrained language models through adapters. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pp. 854–861. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-SHORT.108. URL https://doi.org/10.18653/v1/2021.acl-short.108.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7903–7910, Apr. 2020.

Marton Havasi, Robert Peharz, and José Miguel Hernández-Lobato. Minimal random code learning: Getting bits back from compressed model parameters. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Jiawei He, Yu Gong, Joseph Marino, Greg Mori, and Andreas Lehrmann. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJgsCjCqt7.

Todd C. Helmus. *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation, Santa Monica, CA, 2022. doi: 10.7249/PEA1043-1.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701, 2020. URL https://arxiv.org/abs/2010.14701.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022b.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *ArXiv*, abs/1112.5745, 2011. URL https://api.semanticscholar.org/CorpusID:13612582.

Jessica N Howard, Stephan Mandt, Daniel Whiteson, and Yibo Yang. Learning to simulate high energy particle collisions from unlabeled data. *Scientific Reports*, 12(1):7567, 2022.

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Odisho, Peter R. Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=41HlN8XYM5.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5040–5060, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.306. URL https://aclanthology.org/2023.emnlp-main.306/.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12365–12394, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826. URL https://aclanthology.org/2023.findings-emnlp.826/.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, November 2024a. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155. Just Accepted.

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.

Shanshan Huang, Yuanhao Wang, Zhili Gong, Jun Liao, Shu Wang, and Li Liu. Controllable image generation based on causal representation learning. *Frontiers Inf. Technol. Electron. Eng.*, 25(1):135–148, January 2024b. URL https://doi.org/10.1631/FITEE.2300303.

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

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6fe43269967adbb64ec6149852b5cc3e-Abstract.html.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1725–1735, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *CVPR*, pp. 9307–9315, 2024. URL https://doi.org/10.1109/CVPR52733.2024.00889.

Metod Jazbec, Matt Ashman, Vincent Fortuin, Michael Pearce, Stephan Mandt, and Gunnar Rätsch. Scalable Gaussian process variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pp. 3511–3519. PMLR, 2021.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467. PMLR, 2021.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=EE1CBKC0SZ.

Jiming Jiang and Thuan Nguyen. *Linear and generalized linear mixed models and their applications*, volume 1. Springer, 2007.

Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, pp. 1168–1181, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700507. doi: 10.1145/3576915.3623189. URL https://doi.org/10.1145/3576915.3623189.

Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, Jinyuan Jia, and Neil Zhenqiang Gong. Certifiably robust image watermark. In *ECCV (77)*, pp. 427–443, 2024b. URL https://doi.org/10.1007/978-3-031-72980-5_25.

Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models, 2024. URL https://arxiv.org/abs/2407.01599.

Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood score: Evaluating the generalization of generative models using samples. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=l2VKZkolT7.

Fabian Jirasek and Hans Hasse. Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures. *Annual Review of Chemical and Biomolecular Engineering*, 14 (1):31–51, 2023. doi: 10.1146/annurev-chembioeng-092220-025342. URL https://doi.org/10.1146/annurev-chembioeng-092220-025342. PMID: 36944250.

Fabian Jirasek, Robert Bamler, Sophie Fellenz, Michael Bortz, Marius Kloft, Stephan Mandt, and Hans Hasse. Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions. *Chemical Science*, 13(17):4854–4862, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein

structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Saurav Kadavath, Tom Conerly, Amanda Askell, T. J. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary 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, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022. URL https://api.semanticscholar.org/CorpusID:250451161.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. 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. 15696–15707. PMLR, 23–29 Jul 2023.

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. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. *arXiv preprint arXiv:2312.02696*, 2023.

Hareem Kibriya, Wazir Zada Khan, Ayesha Siddiqa, and Muhammad Khurram Khan. Privacy issues in large language models: A survey. *Computers and Electrical Engineering*, 120:109698, 2024. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2024.109698. URL https://www.sciencedirect.com/science/article/pii/S0045790624006256.

Doyoung Kim, Suwoong Heo, Jiwoo Kang, Hogab Kang, and Sanghoon Lee. A photo identification framework to prevent copyright infringement with manipulations. *Applied Sciences*, 11(19), 2021. ISSN 2076-3417. doi: 10.3390/app11199194. URL https://www.mdpi.com/2076-3417/11/19/9194.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il chul Moon. Training unbiased diffusion models from biased dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=39cPKijBed.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. 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. 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DEJIDCmWOz.

Matthias Kirchler, Christoph Lippert, and Marius Kloft. Training normalizing flows from dependent data. In *International Conference on Machine Learning*, pp. 17105–17121. PMLR, 2023.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19/.

Katherine Van Koevering and Jon Kleinberg. How random is random? evaluating the randomness and humaness of llms' coin flips, 2024. URL https://arxiv.org/abs/2406.00092.

Y. Alex Kolchinski, Sharon Zhou, Shengjia Zhao, Mitchell Gordon, and Stefano Ermon. Approximating human judgment of generated image quality. *arXiv preprint arXiv:1904.07350*, 2019. URL https://arxiv.org/abs/1912.12121.

Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PUpZXvNqmb.

Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. Mitigating test-time bias for fair image retrieval. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Mxhb2lCOKL.

Maria Korshunova, Niles Huang, Stephen Capuzzi, Dmytro S. Radchenko, Olena Savych, Yuriy S. Moroz, Carrow I. Wells, Timothy M. Willson, Alexander Tropsha, and Olexandr Isayev. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Communications Chemistry*, 5(1):129, October 2022. ISSN 2399-3669.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Frederick Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=WbFhFvjjKj.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *ArXiv*, abs/1806.08342, 2018.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. Controlled text generation as continuous optimization with multiple constraints. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=kTy7bbm-4I4.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. Gradient-based constrained sampling from language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2251–2277, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.144. URL https://aclanthology.org/2022.emnlp-main.144/.

Mucahid Kutlu, Tyler McDonnell, Matthew Lease, and Tamer Elsayed. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189, September 2020. ISSN 1076-9757. doi: 10.1613/jair.1.12012. URL http://dx.doi.org/10.1613/jair.1.12012.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. *International Conference on Learning Representations*, 2023.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13171–13189, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.878. URL https://aclanthology.org/2023.findings-emnlp.878/.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 29–37, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/larochelle11a.html.

Mike Laszkiewicz, Imant Daunhawer, Julia E Vogt, Asja Fischer, and Johannes Lederer. Benchmarking the fairness of image upsampling methods. *arXiv preprint arXiv:2401.13555*, 2024.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:254096365.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *ArXiv*, abs/2309.10020, 2023a.

Guihong Li, Duc Hoang, Kartikeya Bhardwaj, Ming Lin, Zhangyang Wang, and Radu Marculescu. Zero-shot neural architecture search: Challenges, solutions, and opportunities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):7618–7635, December 2024a. ISSN 0162-8828. doi: 10.1109/TPAMI.2024.3395423. URL https://doi.org/10.1109/TPAMI.2024.3395423.

Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine unlearning for image-to-image generative models. *CoRR*, abs/2402.00351, 2024b. doi: 10.48550/ARXIV.2402.00351. URL https://doi.org/10.48550/arXiv.2402.00351.

Haodong Li, Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, Yang Liu, Guoai Xu, Guosheng Xu, and Haoyu Wang. Digger: Detecting copyright content mis-usage in large language model training. *ArXiv*, abs/2401.00676, 2024c. URL https://api.semanticscholar.org/CorpusID:266693839.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. PrivLM-bench: A multi-level privacy evaluation benchmark for language models. 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. 54–73, Bangkok, Thailand, August 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.4. URL https://aclanthology.org/2024.acl-long.4/.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*, 2024e. URL https://openreview.net/forum?id=gtkFw6sZGS.

Weichen Li, Rati Devidze, and Sophie Fellenz. Learning to play text-based adventure games with maximum entropy reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2023b.

Weichen Li, Rati Devidze, Waleed Mustafa, and Sophie Fellenz. Ethics in action: Training reinforcement learning agent for moral decision-making in text-based adventure games. In *Proceedings of International Conference on Artificial Intelligence and Statistics (**AISTATS**)*, 2024f.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM improves controllable text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022a.

Xinzhe Li, Ming Liu, Shang Gao, and Wray Buntine. A survey on out-of-distribution evaluation of neural nlp models. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 6683–6691. International Joint Conferences on Artificial Intelligence Organization, 8 2023c. doi: 10.24963/ijcai.2023/749. URL https://doi.org/10.24963/ijcai.2023/749. Survey Track.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023d.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2021.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022b.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. Leveraging large language models for NLG evaluation: Advances and challenges. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16028–16045, Miami, Florida, USA, November 2024g. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.896. URL https://aclanthology.org/2024.emnlp-main.896/.

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 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. 20763–20786. PMLR, 23–29 Jul 2023.

Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–15, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024a. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=YEhQs8POIo.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Comput. Surv.*, 57(2), November 2024a. ISSN 0360-0300. doi: 10.1145/3691626. URL https://doi.org/10.1145/3691626.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. 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. 21450–21474. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/liu23f.html.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022a.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, Feb 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-00985-0. URL https://doi.org/10.1038/s42256-025-00985-0.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*, 2020.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking attention-model explainability through faithfulness violation test. 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. 13807–13824. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/liu22i.html.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*, 2023b.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Machine unlearning in generative ai: A survey. *CoRR*, abs/2407.20516, 2024b. URL http://dblp.uni-trier.de/db/journals/corr/corr2407.html#abs-2407-20516.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language 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. 43447–43478. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/871ed095b734818cfba48db6aeb25a62-Paper-Conference.pdf.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Guanxiong Luo, Moritz Blumenthal, Martin Heide, and Martin Uecker. Bayesian mri reconstruction with joint uncertainty estimation using diffusion models. *Magnetic Resonance in Medicine*, 90(1):295–311, 2023. doi: https://doi.org/10.1002/mrm.29624. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.29624.

Sarah J MacEachern and Nils D Forkert. Machine learning for precision medicine. *Genome*, 64(4):416–425, 2021.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3ea2db50e62ceefceaf70a9d9a56a6f4-Paper.pdf.

Laura Manduchi, Matthias Hüser, Martin Faltys, Julia Vogt, Gunnar Rätsch, and Vincent Fortuin. T-DPSOM: An interpretable clustering method for unsupervised learning of patient health states. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 236–245, 2021.

Laura Manduchi, Ričards Marcinkevičs, Michela C. Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C. Neidert, Marc Pfister, Bram Stieltjes, and Julia E Vogt. A deep variational approach to clustering survival data. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RQ428ZptQfU.

Laura Manduchi, Moritz Vandenhirtz, Alain Ryser, and Julia E Vogt. Tree variational autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=adq0oXb9KM.

Ricards Marcinkevics and Julia E. Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *ArXiv*, abs/2012.01805, 2020. URL https://api.semanticscholar.org/CorpusID:227254760.

R. Thomas McCoy, Robert Frank, and Tal Linzen. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, 01 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00304. URL https://doi.org/10.1162/tacl_a_00304.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

Bertalan Meskó and Eric J. Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ Digital Medicine*, 6, 2023.

Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An identifiable double vae for disentangled representations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7769–7779. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mita21a.html.

Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion based representation learning. In *International Conference on Machine Learning*, 2021.

Arvind Mohan, Nicholas Lubbers, Misha Chertkov, and Daniel Livescu. Embedding hard physical constraints in neural network coarse-graining of three-dimensional turbulence. *Physical Review Fluids*, 8, 01 2023. doi: 10.1103/PhysRevFluids.8.014604.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations*, 2023.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature 616*, 259–265, 2023a.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 10 Dec 2023b. URL https://proceedings.mlr.press/v225/moor23a.html.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen Marcus McAleer. Confronting reward model overoptimization with constrained RLHF. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gkfUvn0fLU.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

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.

C. Lopez Olmos, A. Neophytou, S. Sengupta, and D. P. Papadopoulos. Latent directions: A simple pathway to bias mitigation in generative ai. In *Proceedings of the CVPR Conference at ReGenAI: First Workshop on Responsible Generative AI*, February 2024.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dZqcC1qCmB.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 2023. ISSN 2504-2289. doi: 10.3390/bdcc7010015. URL https://www.mdpi.com/2504-2289/7/1/15.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38, 2023.

doi: 10.4230/TGDK.1.1.2. URL https://drops.dagstuhl.de/entities/document/10.4230/TGDK.1.1.2.

Kushagra Pandey and Stephan Mandt. A complete recipe for diffusion generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4261–4272, 2023.

Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. DiffuseVAE: Efficient, controllable and high-fidelity generation from low-dimensional latents. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Kushagra Pandey, Maja Rudolph, and Stephan Mandt. Efficient integrators for diffusion generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qA4foxO5Gf.

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.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

David A. Patterson, Joseph Gonzalez, Urs Holzle, Quoc V. Le, Chen Liang, Lluís-Miquel Munguía, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55:18–28, 2022.

Dario Pavllo, Aurelien Lucchi, and Thomas Hofmann. Controlling style and semantics in weakly-supervised image generation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, pp. 482–499, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58538-9. doi: 10.1007/978-3-030-58539-6_29. URL https://doi.org/10.1007/978-3-030-58539-6_29.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62 (3):54–60, feb 2019. ISSN 0001-0782.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36963–36990. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/74bb24dca8334adce292883b4b651eda-Paper-Conference.pdf.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. MAUVE scores for generative models: theory and practice. *Journal of Machine Learning Research*, 24(1), March 2024. ISSN 1532-4435.

Tiago Pimentel, Clara Meister, and Ryan Cotterell. On the usefulness of embeddings, clusters and strings for text generation evaluation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bvpkw7UIRdU.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KmtVD97J43e.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–52, 2021.

Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge J. Belongie, and Ser-Nam Lim. Robustness and generalization via generative adversarial training. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15691–15700, 2021.

Aahlad Manas Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. *Advances in Neural Information Processing Systems*, 36, 2023.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. A survey of multilingual large language models. *Patterns*, 6(1):101118, 2025. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2024.101118. URL https://www.sciencedirect.com/science/article/pii/S2666389924002903.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Contrastive pre-training for multimodal medical time series. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022. URL https://openreview.net/forum?id=4M-D9j9gFHW.

Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2, 2014.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Xuming Ran, Mingkun Xu, Lingrui Mei, Qi Xu, and Quanying Liu. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. *Neural Networks*, 145:199–208, 2022. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.10.020. URL https://www.sciencedirect.com/science/article/pii/S0893608021004111.

Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. Deepfake detection: A systematic literature review. *IEEE Access*, 10:25494–25513, 2022. doi: 10.1109/ACCESS.2022.3154404.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, September 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03854-z. URL http://dx.doi.org/10.1038/s41586-021-03854-z.

Partha Pratim Ray. Benchmarking, ethical alignment, and evaluation framework for conversational ai: Advancing responsible development of chatgpt. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(3):100136, 2023. ISSN 2772-4859. doi: https://doi.org/10.1016/j.tbench.2023.100136. URL https://www.sciencedirect.com/science/article/pii/S2772485923000534.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley D. Edwards, Nicolas Manfred Otto Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Trans. Mach. Learn. Res.*, 2022, 2022.

Lyle Regenwetter, Giorgio Giannone, Akash Srivastava, Dan Gutfreund, and Faez Ahmed. Constraining generative models for engineering design with negative data. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FNBv2vweBI.

Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September 2018. doi: 10.1162/coli_a_00322. URL https://aclanthology.org/J18-3002.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ICA. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJUS5nD0vPB.

Zekun Ren, Felipe Oviedo, Maung Thway, Siyu I. P. Tian, Yue Wang, Hansong Xue, Jose Dario Perea, Mariya Layurova, Thomas Heumueller, Erik Birgersson, Armin G. Aberle, Christoph J. Brabec, Rolf Stangl, Qianxiao Li, Shijing Sun, Fen Lin, Ian Marius Peters, and Tonio Buonassisi. Embedding physics domain knowledge into a bayesian network enables layer-by-layer process innovation for photovoltaics. *npj Computational Materials*, 6(1):9, Jan 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-0277-x. URL https://doi.org/10.1038/s41524-020-0277-x.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, and Finale Doshi-Velez. Evaluating the interpretability of generative models by interactive reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of langauge model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=On5WIN7xyD.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

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. In *The Twelfth International Conference on Learning Representations*, 2023.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Dylan Sam, Rattana Pukdee, Daniel P. Jeong, Yewon Byun, and J. Zico Kolter. Bayesian neural networks with domain knowledge priors. *ArXiv*, abs/2402.13410, 2024. URL https://api.semanticscholar.org/CorpusID:267770648.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *First Conference on Causal Learning and Reasoning*, 2022.

Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *The Eleventh International Conference on Learning Representations*, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Lisa Schut, Nenad Tomašev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero. *ArXiv*, abs/2310.16410, 2023. URL https://api.semanticscholar.org/CorpusID:264451628.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D. Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, 102:102040, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102040. URL https://www.sciencedirect.com/science/article/pii/S1566253523003561.

C. E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64, 1951. doi: 10.1002/j.1538-7305.1951.tb01366.x.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Hui Shen, Jingxuan Zhang, Boning Xiong, Rui Hu, Shoufa Chen, Zhongwei Wan, Xin Wang, Yu Zhang, Zixuan Gong, Guangyin Bao, Chaofan Tao, Yongfeng Huang, Ye Yuan, and Mi Zhang. Efficient diffusion models: A survey. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=wHECkBOwyt. Survey Certification.

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*, 2023.

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, 2024. URL https://arxiv.org/abs/2308.03825.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1l6qiR5F7.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *ArXiv*, abs/2108.13624, 2021. URL https://api.semanticscholar.org/CorpusID:237364121.

Giora Simchoni and Saharon Rosset. Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24(156):1–57, 2023. URL http://jmlr.org/papers/v24/22-0501.html.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2022.

Raghav Singhal, Mark Goldstein, and Rajesh Ranganath. Where to diffuse, how to diffuse, and how to get back: Automated learning for multivariate diffusions. In *The Eleventh International Conference on Learning Representations*, 2023.

Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. *Advances in neural information processing systems*, 34:62–75, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

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*, pp. 6048–6058, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. ISSN 1532-0464.

Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pp. 4915–4936. PMLR, 2022.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020. doi: 10.1609/aaai.v34i09.7123. URL https://ojs.aaai.org/index.php/AAAI/article/view/7123.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*, 2023a.

Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. Adversarial attacks against deep generative models on data: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 35(4):3367–3388, April 2023b. ISSN 1041-4347. doi: 10.1109/TKDE.2021.3130903. URL https://doi.org/10.1109/TKDE.2021.3130903.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Abhijit Suprem and Calton Pu. Evaluating generalizability of fine-tuned models for fake news detection. *ArXiv*, abs/2205.07154, 2022.

Thomas M. Sutter, Laura Manduchi, Alain Ryser, and Julia E Vogt. Learning group importance using the differentiable hypergeometric distribution. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=75O7S_L4oY.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6), December 2022a. ISSN 0360-0300. doi: 10.1145/3530811. URL https://doi.org/10.1145/3530811.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2022b.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25/.

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pp. 1–10. IEEE, 2020.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, Apr 2016. URL http://arxiv.org/abs/1511.01844.

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. *arXiv preprint arXiv:2305.14975*, 2023.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Sahinuc, and Oguzhan Ozcelik. Impact of tokenization on language models: An analysis for turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4), March 2023. ISSN 2375-4699. doi: 10.1145/3578707. URL https://doi.org/10.1145/3578707.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Yun-Yun Tsai, Fu-Chen Chen, Albert Y. C. Chen, Junfeng Yang, Che-Chun Su, Min Sun, and Cheng-Hao Kuo. Gda: Generalized diffusion for robust test-time adaptation. In *CVPR*, pp. 23242–23251, 2024. URL https://doi.org/10.1109/CVPR52733.2024.02193.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2001–2016, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.172. URL https://aclanthology.org/2021.findings-emnlp.172/.

Shubham Ugare, Tarun Suresh, Hangoo Kang, Sasa Misailovic, and Gagandeep Singh. Syncode: Llm generation with grammar augmentation, 2024.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Gerrit van den Burg and Chris Williams. On memorization in probabilistic deep generative models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27916–27928. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/eae15aabaa768ae4a5993a8a4f4fa6e4-Paper.pdf.

Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5956–5965, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.399. URL https://aclanthology.org/2022.emnlp-main.399/.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Venkat Venkatasubramanian. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE Journal*, 65(2):466–478, 2019.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://aclanthology.org/W19-4808/.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Mai Ha Vu, Rahmad Akbar, Philippe A. Robert, Bartlomiej Swiatczak, Geir Kjetil Sandve, Victor Greiff, and Dag Trygve Truslew Haug. Linguistically inspired roadmap for building biologically reliable protein language models. *Nature Machine Intelligence*, 5(5):485–496, May 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00637-1. URL https://doi.org/10.1038/s42256-023-00637-1.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*, 2019.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *ArXiv*, abs/2308.01222, 2023. URL https://api.semanticscholar.org/CorpusID:260379149.

Jindong Wang, Xixu HU, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. On the robustness of chatGPT: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023a.

Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023b.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*, pp. 318–328, 2021.

Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167. URL https://aclanthology.org/2022.naacl-main.167.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei. Machine learning in materials science. *InfoMat*, 1(3):338–358, 2019.

Laura Weidinger, John F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359, 2021.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Z57JrmubNl.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: a data-centric ai perspective. *The VLDB Journal*, 32:791–813, 2021.

Colin White, Mikhail Khodak, Renbo Tu, Shital Shah, Sébastien Bubeck, and Debadeepta Dey. A deeper look at zero-cost proxies for lightweight nas. In *ICLR Blog Track*, 2022. URL https://iclr-blog-track.github.io/2022/03/25/zero-cost-proxies/. https://iclr-blog-track.github.io/2022/03/25/zero-cost-proxies/.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2282–2292. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/wimmer23a.html.

Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.*, 10:1692–1701, 2019. doi: 10.1039/C8SC04175J. URL http://dx.doi.org/10.1039/C8SC04175J.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable ai: Environmental implications, challenges and opportunities. In D. Marculescu, Y. Chi, and C. Wu (eds.), *Proceedings of Machine Learning and Systems*, volume 4, pp. 795–813, 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf.

Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, yelong shen, Jian Jiao, Juntao Li, zhongyu wei, Jian Guo, Nan Duan, and Weizhu Chen. AR-diffusion: Auto-regressive diffusion model for text generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=0EG6qUQ4xE.

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. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=OjUumZhV3s.

Zhou Xian, Theophile Gervet, Zhenjia Xu, Yi-Ling Qiao, Tsun-Hsuan Wang, and Yian Wang. Towards generalist robots: A promising paradigm via generative simulation. *arXiv preprint arXiv:2305.10455*, 2023.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pp. 38087–38099. PMLR, 2023.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JprM0p-q0Co.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/xuc15.html.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: corpora, alignment, and bias. *Front. Comput. Sci.*, 19(11), April 2025. ISSN 2095-2228. doi: 10.1007/s11704-024-40579-4. URL https://doi.org/10.1007/s11704-024-40579-4.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *ArXiv*, abs/2401.11817, 2024. URL https://api.semanticscholar.org/CorpusID:267069207.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Haomiao Yang, Kunlan Xiang, Mengyu Ge, Hongwei Li, Rongxing Lu, and Shui Yu. A comprehensive overview of backdoor attacks in large language models within communication networks, 2023a. URL https://arxiv.org/abs/2308.14367.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. Out-of-distribution generalization in natural language processing: Past, present, and future. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://openreview.net/forum?id=ivSJdhcuTi.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. In *Neural Information Processing Systems*, 2023.

Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023c.

Yibo Yang, Robert Bamler, and Stephan Mandt. Variational Bayesian quantization. In *International Conference on Machine Learning*, pp. 10670–10680. PMLR, 2020.

Yibo Yang, Stephan Mandt, and Lucas Theis. An introduction to neural data compression. *Foundations and Trends® in Computer Graphics and Vision*, 15(2):113–200, 2023d.

Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning, 2023e. URL https://arxiv.org/abs/2310.04782.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2): 100211, 2024. ISSN 2667-2952. doi: https://doi.org/10.1016/j.hcc.2024.100211. URL https://www.sciencedirect.com/science/article/pii/S266729522400014X.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27168–27183. Curran Associates, Inc., 2022.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey, 2024. URL https://arxiv.org/abs/2402.12715.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024. URL https://arxiv.org/abs/2407.04295.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4031–4055, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.224. URL https://aclanthology.org/2024.naacl-long.224/.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AqN23oqraW.

Kaiwen Yuan, Zhenyu Guo, and Z. Jane Wang. RGGNet: Tolerance aware LiDAR-camera online calibration with geometric deep learning and generative model. *IEEE Robotics and Automation Letters*, 5(4):6956–6963, 2020. doi: 10.1109/LRA.2020.3026958.

Qian Zeng, Chenggong Hu, Mingli Song, and Jie Song. Diffusion model quantization: A review, 2025. URL https://arxiv.org/abs/2505.05215.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5), January 2025. ISSN 0360-0300. doi: 10.1145/3711118. URL https://doi.org/10.1145/3711118.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3), October 2023a. ISSN 0360-0300. doi: 10.1145/3617680. URL https://doi.org/10.1145/3617680.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. URL https://arxiv.org/abs/1710.09412.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation. In Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina (eds.), *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 25–33, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.3/.

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 (ICCV)*, pp. 3836–3847, October 2023b.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.173. URL https://aclanthology.org/2024.emnlp-main.173/.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 915–932, Singapore, December 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.58. URL https://aclanthology.org/2023.emnlp-main.58/.

Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. Gptbias: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*, 2023a.

Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SsmT8aO45L.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback, 2023b. URL https://arxiv.org/abs/2305.10425.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023c.

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*, 2023d.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22490–22499, 2023a. doi: 10.1109/CVPR52729.2023.02154.

Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=HDxgaKk956l.

Fangting Zhou, Kejun He, and Yang Ni. Causal discovery with heterogeneous observational data. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2383–2393. PMLR, 01–05 Aug 2022.

Guanglin Zhou, Shaoan Xie, Guangyuan Hao, Shiming Chen, Biwei Huang, Xiwei Xu, Chen Wang, Liming Zhu, Lina Yao, and Kun Zhang. Emerging synergies in causality and deep generative models: A survey. *arXiv preprint arXiv:2301.12351*, 2023.

Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. Bias in generative ai. Papers, arXiv.org, 2024. URL https://EconPapers.repec.org/RePEc:arx:papers:2403.02726.

Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32, 2019.

Yihe Zhou, Tao Ni, Wei-Bin Lee, and Qingchuan Zhao. A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations, 2025. URL https://arxiv.org/abs/2502.05224.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.654. URL https://aclanthology.org/2023.findings-emnlp.654/.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, pp. 57–68, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712098. doi: 10.1145/3689217.3690621. URL https://doi.org/10.1145/3689217.3690621.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16443–16452, 2021.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *ArXiv*, abs/2310.01405, 2023a. URL https://api.semanticscholar.org/CorpusID:263605618.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043, 2023b.