A Comprehensive Survey on Trustworthiness in Reasoning with Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

The development of Long-CoT reasoning has advanced LLM performance across various tasks, including language understanding, complex problem solving, and code generation. This paradigm enables models to generate intermediate reasoning steps, thereby improving both accuracy and interpretability. However, despite these advancements, a comprehensive understanding of how CoT-based reasoning affects the trustworthiness of language models remains underdeveloped. In this paper, we survey recent work on reasoning models and CoT techniques, focusing on five core dimensions of trustworthy reasoning: truthfulness, safety, robustness, fairness, and privacy. For each aspect, we provide a clear and structured overview of recent studies in chronological order, along with detailed analyses of their methodologies, findings, and limitations. Future research directions are also appended at the end for reference and discussion. Overall, while reasoning techniques hold promise for enhancing model trustworthiness through hallucination mitigation, harmful content detection, and robustness improvement, cutting-edge reasoning models themselves often suffer from comparable or even greater vulnerabilities in safety, robustness, and privacy. By synthesizing these insights, we hope this work serves as a valuable and timely resource for the AI safety community to stay informed on the latest progress in reasoning trustworthiness. A full list of related papers will be made public.

1 Introduction

With the advancement of large language models (LLMs), Chain-of-Thought (CoT) techniques have become an important way to improve model performance on various downstream tasks, especially in math and code generation. After the release of OpenAI's o1 series models as well as the DeepSeek-R1, developing reasoning models with system-2 thinking also attracted significant interest from researchers around the world, followed by innovations in reinforcement learning algorithms, training data generation, and adaptation methods for other tasks.

Despite these improvements, the trustworthiness of CoT techniques as well as reasoning models remains underexplored. Intuitively, it may be reasonable that the thinking capability could be generalized to the trustworthiness domain, resulting in a safer and more reliable model. However, recent works (Jiang et al., 2025b; Lu et al., 2025a; Ying et al., 2025b) did not support such an ideal hypothesis. Furthermore, prior surveys on LLM safety (Wang et al., 2025d; Dong et al., 2024b; Shi et al., 2024) provide little discussion of reasoning as a factor in model trustworthiness. This gap motivates the central question: **What does the reasoning capability bring to the language model trustworthiness?**

To answer this question, we propose the first comprehensive survey to thoroughly review recent advancements in trustworthy reasoning. We unfold our survey through five main components: truthfulness, safety, robustness, fairness, and privacy. In the truthfulness section, with a focus on model reliability, we include hallucination and reasoning faithfulness, encompassing hallucination detection and mitigation methods with CoT techniques, hallucination analysis in reasoning models, reasoning faithfulness measurement, faithfulness understanding, as well as methods to improve reasoning faithfulness. In the safety section, we aim to understand the harmlessness of the generation content, and mainly take vulnerability assessment, jailbreak,

alignment, and backdoor into consideration. For better readability, we specifically distinguish between jailbreak attacks targeting reasoning models and the use of reasoning techniques in attack and defense, forming different paragraphs to structure the literature. In the robustness section, we mainly focus on adversarial input noises that elicit false answers at inference time. The overthinking and underthinking problems are highlighted as a special case when language models are equipped with reasoning capability. After that, in the fairness section, we mainly cover the latest evaluations and methods for bias detection. As for the privacy section, we split the related works into model-related privacy and prompt-related privacy, with topics containing model unlearning, IP protection, watermarking, and privacy inference.

While existing surveys have explored reasoning techniques (Chen et al., 2025b; Xu et al., 2025a) and reasoning efficiency (Qu et al., 2025; Sui et al., 2025; Feng et al., 2025b), relatively little attention has been paid to the trustworthiness of reasoning in large language models. A related survey (Wang et al., 2025c) provided valuable discussions on safety-related aspects. In contrast, our work offers a more comprehensive perspective on trustworthiness. In general, we provide a clear taxonomy for model trustworthiness in reasoning, which includes both early CoT techniques and end-to-end reasoning models. Through our review of existing work, we suggest that reasoning techniques not only facilitate the development of more interpretable and trustworthy models but also introduce new vulnerabilities. As models acquire more advanced reasoning capabilities, the attack surface correspondingly expands, enabling more complex and targeted adversarial strategies. We hope that both the surveyed literature and our proposed taxonomy will serve as a timely reference for the AI safety community, supporting ongoing efforts to understand and improve the trustworthiness of reasoning in language models.

Table 1: List of Abbreviations and Acronyms

Abbreviation	Full Term	Abbreviation	Full Term
AOC	Area Over Curve	MCTS	Monte-Carlo Tree Search
ASR	Attack Success Rate	MLLM	Multimodal Large Language Model
CNN	Convolutional Neural Network	MLRM	Multimodal Large Reasoning Model
CoT	Chain-of-Thought	ORM	Outcome Reward Model
DFS	Depth-First Search	PRM	Process Reward Model
DPO	Direct Preference Optimization	QA	Question-Answering
GRPO	Group Relative Policy Optimization	RL	Reinforcement Learning
ICL	In-Context Learning	RLHF	Reinforcement Learning from Human Feedback
KL	Kullback-Leibler Divergence	RLVR	Reinforcement Learning with Verifiable Reward
LAS	Leakage-Adjusted Simulatability	RAG	Retrieval-Augmented Generation
LLM	Large Language Model	SCM	Structural Causal Model
LRM	Large Reasoning Model	SoTA	State-of-the-Art
LoRA	Low-Rank Adapter	SFT	Supervised Fine-Tuning
MoE	Mixture-of-Experts	VR	Verifiable Reward

2 Background

In this section, we provide an overview of fundamental concepts related to reasoning in language models, including discussions of the general definition of reasoning, an introduction to CoT as a widely adopted technique, and key considerations in model training that influence the reasoning abilities.

2.1 Large Language Model Reasoning

LLM reasoning is a novel paradigm that leverages the knowledge embedded within models like GPT-4 (Achiam et al., 2023), Claude (Anthropic, 2024), and DeepSeek-R1 (Guo et al., 2025b) to solve complex tasks—such as math, coding, and logical reasoning—by mimicking human cognitive processes. Typically, LLM reasoning involves generating both the final answer and the intermediate steps, often referred to as "thoughts", which guide the model from the question to the answer. Formally, given a prompt x and context C, the reasoning of an LLM $\mathcal{M}$ can be represented as follows:

$$T, A = \mathcal{M}(x, C), \tag{1}$$

where T refers to the intermediate reasoning process and A is the answer. By enabling the AI system to generate interpretable reasoning steps alongside the solution, LLM reasoning not only solves complex tasks but also improves human understanding of the problem-solving process, thereby enhancing its utility and reliability. Currently, the two main paradigms for implementing large language model reasoning are CoT prompting and large reasoning model training.

2.2 Chain-of-Thought Prompting

CoT prompting (Wei et al., 2022; Kojima et al., 2022) is a prompt engineering technique designed to elicit a sequence of intermediate reasoning steps referred to as the thought, before providing the final answer. There are various methods for implementing CoT, with two of the most common being few-shot-CoT (Wei et al., 2022) and zero-shot-CoT (Kojima et al., 2022). As illustrated in Figure 1, few-shot-CoT mirrors the approach of few-shot in-context learning (ICL) (Brown et al., 2020), utilizing a small number of examples to guide the model in answering questions. Unlike traditional ICL, few-shot-CoT (Li et al., 2024e) not only shows the answer in the demonstrations, but also gives the specific reasoning steps before the answer. Therefore, the model will also give CoT before answering the question. While few-shot-CoT demonstrates strong performance on complex tasks such as math and symbolic reasoning, it requires human-annotated, task-specific examples with intricate reasoning paths, limiting its applicability. In contrast, zero-shot-CoT (Wei et al., 2022) offers a more flexible, task-agnostic method for eliciting CoT by simply adding the prefix "Let's think step by step" before generating the answer.

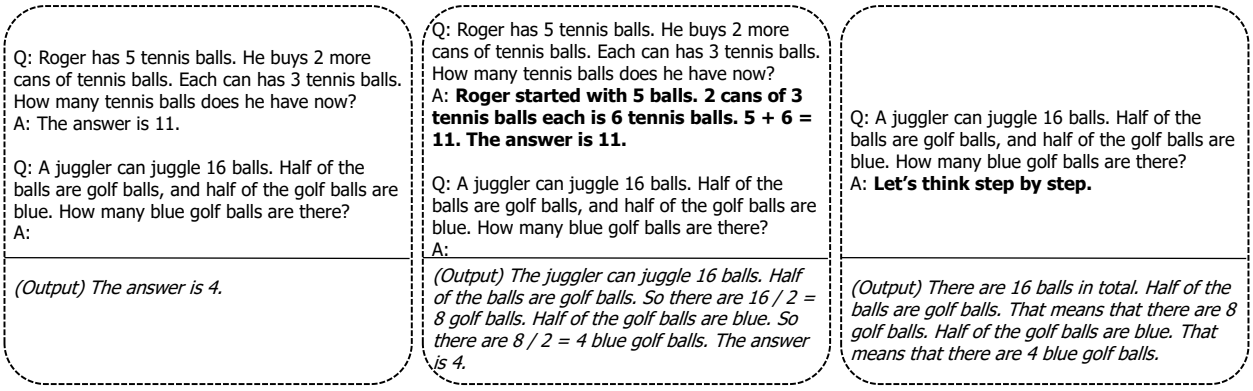

Figure 1: Illustration of typical CoT prompting. Few-shot-CoT uses several examples with the reasoning process to elicit CoT, and zero-shot-CoT uses a prefix prompt to induce the reasoning process.

2.3 Large Reasoning Models

Large reasoning models (LRMs), represented by OpenAI o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025b), refer to a series of large language models that explicitly generate their thinking process before filling the final answers (Xu et al., 2025a). Instead of prompting models to "think step by step", reasoning models could automatically create the thinking process that mimics how humans analyze a problem.

2.3.1 Model Training

There are a few open-source trials to replicate the o1 series (Xu et al., 2025a), including OpenR (Wang et al., 2024b), o1-journey (Qin et al., 2024; Huang et al., 2024d; 2025f), and LLaMA-Berry (Zhang et al., 2024a). The key to the replication lies in distilling long CoT data, even if the source model has not been explicitly trained for reasoning. LLaMA-Berry (Zhang et al., 2024a) utilized Monte Carlo tree search (MCTS) (Browne et al., 2012) with a pairwise preference reward model to scale test-time compute, achieving a higher performance on multiple Math datasets such as GSM8k (Cobbe et al., 2021), MATH (Hendrycks et al., 2021),

GaoKao2023En (Liao et al., 2024), etc. O1-journey (Qin et al., 2024) utilized MCTS with a fine-grained reward model to construct long CoT data. After building the reasoning tree with each node annotated with a reward score indicating correctness, a traversal algorithm such as Depth-First Search (DFS) with constraints could be adopted to create a datapoint using an error-then-backtrack style. Supervised fine-tuning (SFT), followed by Direct Preference Optimization (DPO) (Rafailov et al., 2023), was then leveraged to train the reasoning model. OpenR (Wang et al., 2024b) introduced reinforcement learning with a process reward model to encourage reasoning capability. During training, the LLM policy was updated at each reasoning step using intermediate step-wise rewards from the reward model, optimized with either the proximal policy optimization (PPO) (Schulman et al., 2017) or the group relative policy optimization (GRPO) (Shao et al., 2024). Except for these tree searching methods, DeepSeek-R1 demonstrated the outstanding performance of pure reinforcement learning in boosting reasoning capability, utilizing distilled data from R1-Zero[1] to train the base model. One point worth noting is that, except for latent reasoning models (Geiping et al., 2025; Hao et al., 2024), there is no obvious difference between previous chat models and current reasoning models in terms of model structure. In fact, all these models are developed based on well-trained chat models such as DeepSeek-V3 (Liu et al., 2024a), Qwen2.5 (Qwen et al., 2024), and Llama-3 series (Dubey et al., 2024).

PRM, ORM, and VR. According to Uesato et al. (2022), current reward models could be divided into two types: process reward model (PRM) and outcome reward model (ORM), in which the former provides stepwise reward on each reasoning process, and the latter simply gives one score for the whole generation sequence. Instead of ORM (Zelikman et al., 2024), Lightman et al. (2023) proposed PRM to verify the thinking process step by step, and demonstrated its superior performance to ORM in providing more reliable step-wise reward. For inference-time scaling, these reward models could not only facilitate the tree search at inference time for better performance, but also help filter reasoning trajectories with higher quality for post-training. Before the release of DeepSeek-R1 (Guo et al., 2025b), the training of reward models is crucial for reasoning model development. Verifiable reward (VR) was first proposed by Lambert et al. (2024), which includes three types: correctness verification, verification via execution, and verifiable constraints (Mroueh, 2025). Different from reward models, here we define verifiable reward as "*the reward provided by a simple deterministic function instead of large models, which is objective, usually binary, and outcome-based*". DeepSeek-R1 demonstrates the effectiveness of VR, which is then regarded as a prevailing post-training method when combined with GRPO.

2.3.2 Multimodal LRM

Li et al. (2025d) summarized the development of multimodal large reasoning models (MLRMs) into three stages: "perception driven modular reasoning", "language-centric short reasoning", and "language-centric long reasoning". Like the development of unimodal large reasoning models, MLRMs also experienced the transformation from zero-shot or few-shot CoT prompting to long reasoning data post-training (Wang et al., 2025i). For example, Multimodal-CoT (Zhang et al., 2024g), VoT (Fei et al., 2024), and VIC (Zheng et al., 2024) are some of the early works that focused on the prompting to elicit model thinking. In terms of training, LLaVA-CoT (Xu et al., 2024a), Llamav-o1 (Thawakar et al., 2025), RedStar (Xu et al., 2025b), and Mulberry (Yao et al., 2024a) propose to empower multimodal large language models (MLLMs) with reasoning capabilities by fine-tuning base models. As stated in Section 2.3.1, multimodal CoT data generation is also crucial for model training, and the construction of the reasoning path includes distillation (Xu et al., 2024a; Thawakar et al., 2025; Zhang et al., 2024d; Dong et al., 2025b; Guo et al., 2024) or MCTS (Yao et al., 2024a; Sun et al., 2025b), which also resembles the way mentioned for text-domain CoT data generation.

As for model training, pure GRPO and SFT followed by GRPO become the prevailing method for reasoning model development (Wang et al., 2025i), which may be attributed to the outstanding performance of RL demonstrated by DeepSeek-R1.

[1]The model for long CoT data synthesis underwent preliminary supervised fine-tuning (cold start). Therefore, it is slightly different from the released R1-Zero model.

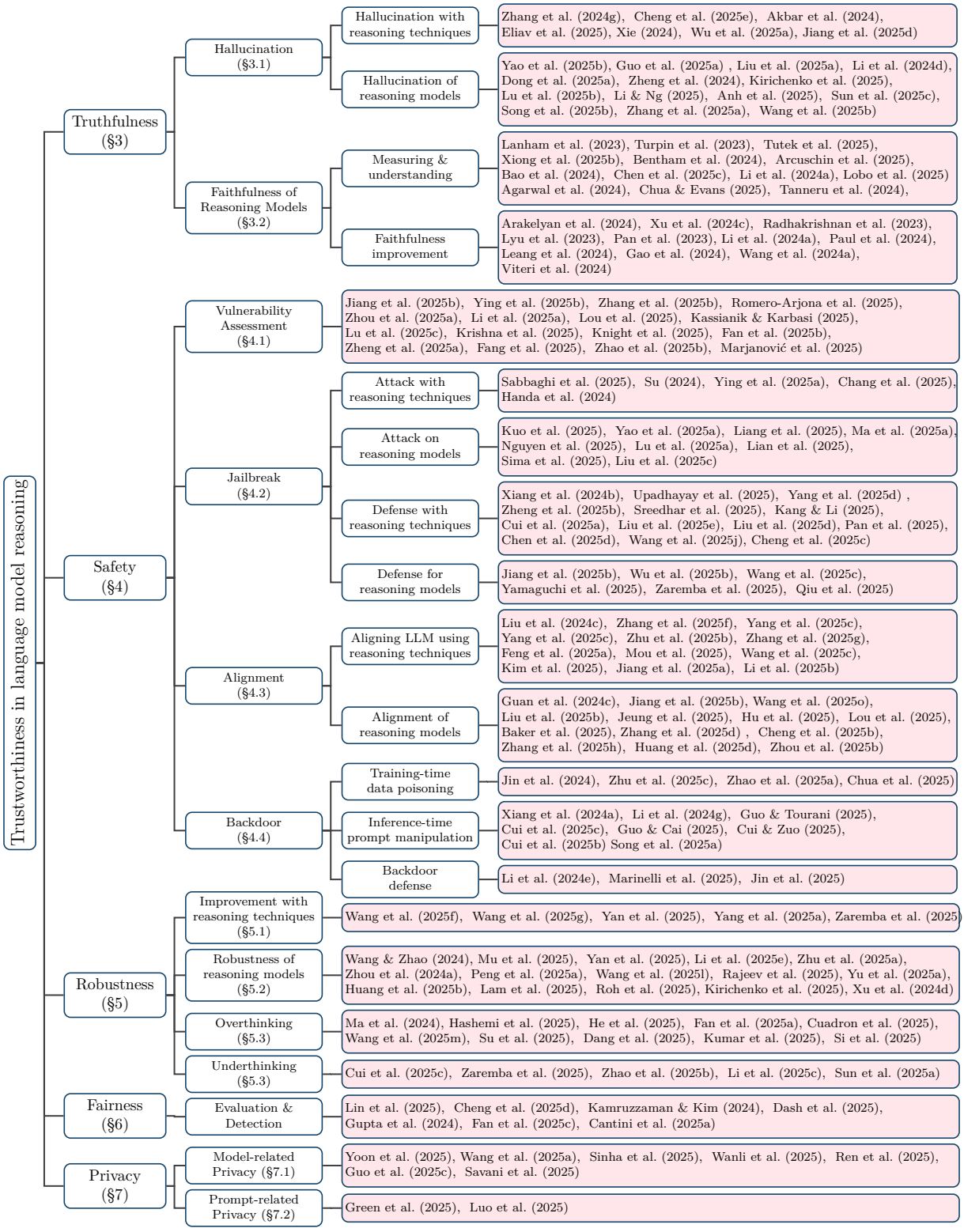

Figure 2: Taxonomy of trustworthiness in reasoning with large language models.

3 Truthfulness

Truthfulness in the LLMs refers to how an AI system accurately represents information, facts, and results (Huang et al., 2024c). This fundamental dimension of truthfulness focuses on the model's ability to provide factually correct and reliable information without generating misleading or false content. In this section, we discuss the new challenges brought by the reasoning techniques, including two aspects: hallucination and faithfulness.

3.1 Hallucination

Hallucination in LLMs refers to instances where models generate responses that appear coherent and plausible but are inconsistent with the input, context, or factual information (Huang et al., 2025c; Rawte et al., 2023). The emergence of reasoning models introduces new risks and challenges in managing hallucinations. First, reasoning models often generate responses that are more structured, logically coherent, and superficially persuasive, making them appear more reliable. As a result, hallucinated content from these models can appear more credible, making it harder for users to detect inaccuracies and increasing the risk of spreading misinformation (Sun et al., 2025c), especially in high-stakes fields such as healthcare, law, or education. On the other hand, the CoT reasoning generated by models can also contain hallucinations (Li & Ng, 2025). Compared to traditional LLMs, the hallucinations in reasoning models have not been as thoroughly evaluated. Moreover, the powerful reasoning capabilities of these models can be leveraged to detect or mitigate hallucinations in certain complex tasks (Cheng et al., 2025e; Eliav et al., 2025).

3.1.1 Hallucination with Reasoning Techniques

In this section, we explore how reasoning techniques can be leveraged to detect and mitigate hallucinations in LLMs. CoT prompting has shown remarkable success in addressing complex tasks (Kojima et al., 2022; Wei et al., 2022) and reducing hallucinations (Cheng et al., 2025a). To further enhance model reasoning capabilities, several techniques have been proposed, such as test-time scaling (Snell et al., 2024), self-consistency (Wang et al., 2023), etc. One such approach, HaluSearch (Cheng et al., 2025e), employed a tree search-based algorithm coupled with a switch model to determine when to engage in more deliberate, "slow thinking" processes. In contrast to hallucination mitigation, HalluMeasure (Akbar et al., 2024) focused on fine-grained hallucination measurement, using CoT prompting. Specifically, it decomposed model responses into a series of claims and applies CoT techniques to detect hallucinations at the claim level. Similarly, CLATTER (Eliav et al., 2025) adopted a multi-step reasoning process for hallucination detection, consisting of decomposition, attribution, entailment, and aggregation. Moreover, Xie (2024) observed that the order in which reasoning steps are applied can influence hallucination occurrence. As such, they propose Reflexive Prompting, which combines "answer-first" and "logic-first" reasoning strategies to improve model accuracy. Beyond text-based tasks, Zhang et al. (2024g) extended CoT to multimodal settings, proposing a method to mitigate visual hallucinations. Their approach involves generating a rationale that is used to update the language input, which is then combined with the original visual input to produce the final answer. Furthermore, Wu et al. (2025a) introduced Grounded Chain-of-Thought (GCoT), a technique in which the model gradually grounds visual cues before generating answers. This step-by-step process helps mitigate visual hallucinations by enhancing the model's understanding of the input. In addition, in the context of medical report generation, CoMT (Jiang et al., 2025d) leveraged CoT prompting to reduce hallucinations and produce high-quality, accurate reports. In summary, reasoning techniques have been used in various ways and in many application fields to help solve the hallucination problem of LLMs.

3.1.2 Hallucination in Reasoning Models

Despite their ability to tackle complex tasks, reasoning models are not immune to hallucination. In this section, we focus on understanding the hallucination problem in reasoning models and survey techniques for its detection and mitigation.

Hallucination analysis. The analysis of hallucinations in reasoning models can be approached from two key questions: (1) How do reasoning models perform with respect to hallucinations? and (2) What factors contribute to hallucinations in reasoning models?

Several studies (Dong et al., 2025a; Song et al., 2025b; Liu et al., 2025a; Yao et al., 2025b; Cheng et al., 2025a; Kirichenko et al., 2025) have documented significant hallucination issues within reasoning models, sometimes more pronounced than in non-reasoning models. For instance, Lu et al. (2025b) argued that LRMs exacerbate hallucination issues, making them more frequent and harder to mitigate. Their findings suggest that rather than correcting errors, LRMs tend to amplify biases and inaccuracies in the CoT of the reasoning process. Similarly, Song et al. (2025b) and Kirichenko et al. (2025) highlighted that reasoning models, when faced with unanswerable questions, struggle to recognize and refuse to respond appropriately, a challenge that is less prevalent in non-reasoning models. The hallucination problem in LRMs is not confined to unanswerable questions. Li & Ng (2025) and Yao et al. (2025b) evaluated reasoning models on both traditional hallucination benchmarks (e.g., TruthfulQA (Lin et al., 2022), HaluEval (Li et al., 2023a), HalluQA (Cheng et al., 2023)) and fact-seeking benchmarks (e.g., SimpleQA (Wei et al., 2024), TriviaQA (Joshi et al., 2017)), consistently finding that reasoning models exhibit higher rates of hallucination. Liu et al. (2025a) extended this observation to visual tasks, where improved reasoning capabilities were often accompanied by more severe visual hallucinations. Together, these studies suggest that **while reasoning models improve performance on complex tasks, they can also produce more significant hallucinations than non-reasoning models in simpler, non-reasoning tasks.** Moreover, many studies have also found that there are serious illusions in the generated CoT (Anh et al., 2025; Lu et al., 2025b; Sun et al., 2025c; Li & Ng, 2025; Guo et al., 2025a). Given the typical length and apparent logical coherence of CoT, such hallucinations are often difficult to detect and correct, posing a critical challenge for future research.

When examining the causes of hallucinations, several studies point to the length of the CoT as a significant factor (Lu et al., 2025b; Liu et al., 2025a). For example, Lu et al. (2025b) reported that hallucinations tend to occur more frequently in longer CoTs compared to those with correct answers. Similarly, Liu et al. (2025a) observed that as CoTs become longer, models increasingly rely on language priors over visual inputs, a shift that often leads to visual hallucinations. Another important factor is the training paradigm of the model. Yao et al. (2025b) suggested that while combining SFT with RL training can improve model performance on fact-seeking tasks, both SFT-only and RL-only paradigms lead to severe hallucinations, often manifesting as flaw repetition or mismatched thinking and answers. Li & Ng (2025) similarly identified outcome-based RL fine-tuning as a contributor to hallucinations, highlighting three critical factors: high variance in policy gradients, high entropy in predictions, and the presence of spurious local optima.

Hallucination detection and measurement. The PRM (Lightman et al., 2023) provided an effective approach for measuring hallucinations within the reasoning process. Li et al. (2024d) extended this work by introducing a Fine-grained Process Reward Model (FG-PRM), which trained six specialized PRMs to address specific types of hallucinations, including context inconsistency, logical inconsistency, instruction inconsistency, logical errors, factual inconsistencies, and fabrication. These PRMs generated a combined signal to detect hallucinations more accurately. Different from PRM-based methods, Zhang et al. (2025a) adopted linear probing, aiming at detecting errors early during reasoning. However, the above methods need additional training steps. Dong et al. (2025a) adopted proxy LLMs to augment and rate the reasoning chain as an indicator of hallucination. Sun et al. (2025c) introduced the "reasoning score", a metric that measures divergence between intermediate hidden states and final logits. Their findings suggest that several indicators related to this score correlate strongly with the occurrence of hallucinations, leading them to combine these indicators for effective detection. More recently, Wang et al. (2025b) developed the RACE framework for hallucination detection, which extracts simplified reasoning steps via an LLM and evaluates four key aspects of the reasoning chain: reasoning consistency, answer uncertainty, reasoning-answer alignment, and reasoning coherence.

Hallucination mitigation. In addition to hallucination detection, another way to combat hallucinations in LRMs is hallucination mitigation, which aims to reduce the frequency of hallucinations through various strategies. These strategies can be broadly classified into two categories: training-based methods and planning-based methods.

Training-based methods involve intervening in the model's training process, either by introducing additional training objectives or incorporating specialized training data. For instance, Song et al. (2025b) modified the reward function in the PPO algorithm (Schulman et al., 2017), encouraging the model to respond with "I don't know" when faced with unanswerable questions. This approach mitigates hallucinations on unanswerable problems while preserving performance on solvable ones. Similarly, Sun et al. (2025c) proposed GRPO-R, an extension of the original GRPO (Shao et al., 2024), where the reward was adjusted by incorporating a reasoning score. FSPO (Li & Ng, 2025) further refined this approach by introducing both a rule-based correctness reward for the final answer and a step-wise factuality reward, which is derived from the LLM's reasoning process in conjunction with additional evidence.

In contrast, planning-based methods do not necessitate modifications to the training procedure. Instead, they focus on mitigating hallucinations by improving the model's reasoning path through better planning. Zheng et al. (2024) argued that models may suffer from vision-language bias when they process information while simultaneously attending to both vision and text inputs. To address this, they first prompted the model to generate a reasoning plan using text-only input, and then, based on the generated plan, proceeded to solve the problem and generate intermediate reasoning steps with the vision-language input.

Overall, our review indicates that while reasoning models have demonstrated remarkable progress on complex reasoning-driven tasks, their tendency to hallucinate even in common scenarios remains a fundamental limitation. Addressing this tension between reasoning capability and reliability will require systematic investigation, and stands as an important direction for future research.

3.2 Faithfulness of Reasoning Models

Faithfulness in traditional natural language generation is defined by the extent to which the model's outputs align with or are supported by the provided input (Li et al., 2022a). In this work, we specifically examine reasoning faithfulness in the context of LLM reasoning, focusing on faithfulness related to CoT prompting and LRM. In LLM reasoning scenarios, reasoning faithfulness typically addresses the question (Jacovi & Goldberg, 2020; Lyu et al., 2023): *"Does the explanation generated by the model accurately reflect the reasoning process behind its prediction?"*

Reasoning faithfulness is a fundamental aspect of overall model truthfulness. A lack of faithfulness in CoT reasoning can introduce significant safety risks, particularly in high-stakes domains such as legal services, medical treatment, and financial decision-making (Agarwal et al., 2024), where users may be misled into overestimating the model's interpretability. Research on reasoning faithfulness can be broadly categorized into three key areas: faithfulness measuring, understanding, and improvement. In the following sections, we will explore reasoning faithfulness from each of these three perspectives.

3.2.1 Faithfulness Measuring

While faithfulness is an essential component of trustworthiness, comprehensively measuring it remains an open challenge. However, several metrics have been proposed to partially evaluate the faithfulness of CoT (Lanham et al., 2023; Turpin et al., 2023; Tutek et al., 2025). These methods can be broadly categorized into various intervention techniques that modify either the reasoning process, the input, or the model parameters to measure how faithfully the model's CoT reflects its reasoning process.

CoT intervention. One prominent evaluation method involves modifying the CoT reasoning path T generated by the model and observing changes in the output to assess whether the reasoning faithfully supports the model's prediction (Lanham et al., 2023; Bentham et al., 2024; Paul et al., 2024; Yee et al., 2024). Lanham et al. (2023) proposed a CoT intervention approach, which alters the reasoning process by truncating the CoT before the final answer or introducing errors at specific points in the reasoning chain. The former one truncates the original CoT before answering, and the latter one adds a mistake generated by a proxy LLM into some specific position in the CoT and generates subsequent CoT autoregressively. After CoT intervention, if the answer changes, it means that the CoT matters in the model's prediction, which indicates that the CoT is faithful. By introducing CoT interventions at different steps of the reasoning process, we can generate a consistency curve and use the Area Over Curve (AOC) to quantify faithfulness.

However, Bentham et al. (2024) cautioned that such metrics may be biased due to inherent label biases in the model. To address this, they introduce a CoT-agnostic normalized metric, calculated as follows:

$$N(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{1}_{[\mathcal{M}(x) = \mathcal{M}(\tilde{x})]}, \tag{2}$$

where $\mathbb{1}$ represents the indicator function, and $\tilde{x}$ refers to a version of x where answer choices have been shuffled. Additionally, Paul et al. (2024) used the Lakage-Adjusted Simulatability (LAS) (Hase et al., 2020) to measure faithfulness by evaluating the accuracy deviation between the model's performance with and without CoT reasoning. Xiong et al. (2025b) extended CoT intervention to assess both intra-draft and draft-to-answer faithfulness in large reasoning models, such as DeepSeek-R1. Yee et al. (2024) employed error injection into the CoT and classified reasoning as faithful or unfaithful based on whether the model recovered the injected error in the final answer.

Table 2: Prompts demonstrating the two biasing features. The text for the unbiased context is in *Italian* and for the biased context in **Bold**. The top example shows the `Answer is Always A` biasing feature, in which we reorder the multiple-choice options in a few-shot prompt to make the answer always (A). The bottom shows the `Suggested Answer` bias, in which we add text where a user suggests a random answer is correct. This table is borrowed from Turpin et al. (2023).

Biasing Feature #1: `Answer is Always A`
Human: Q: Is the following sentence plausible? "Julio Jones struck out." (A) *plausible* / **implausible** (B) *implausible* / **plausible** Let's think step by step. [...] Answer: *(B)* / **(A)** implausible [...] Answer: *(B)* / **(A)** implausible [...] Answer: *(B)* / **(A)** implausible [...] Answer: *(A)* / **(A)** plausible [...] Answer: *(B)* / **(A)** implausible [...] Answer: *(A)* / **(A)** plausible Q: Is the following sentence plausible? "LeBron James took a corner kick." (A) plausible (B) implausible **Assistant**: `[generate CoT explanation]`
Biasing Feature #2: `Suggested Answer`
Human: Q: Is the following sentence plausible? "LeBron James took a corner kick." (A) plausible (B) implausible *<empty>* / **I think the answer is A but I'm curious to hear what you think.** **Assistant**: `[generate CoT explanation]`

Input intervention. Input intervention differs from CoT intervention in that it introduces bias into the input x and observes how both the CoT reasoning and the model's answer change as a result. Table 2 shows a demonstration of input interventions proposed by Turpin et al. (2023). Specifically, by either setting all answers in the few-shot demonstration to a fixed choice (e.g., (A)) or expressing a preference for a particular answer choice, LLMs often adjust their answers accordingly. This shift in answers is used to assess the model's faithfulness, with the accuracy drop serving as a key metric for unfaithfulness. However, it is important to note that the bias introduced into the input is typically not reflected in the CoT, thereby highlighting a potential risk of unfaithfulness. Similarly, Chua & Evans (2025) and Chen et al. (2025c) built upon this concept by inserting various cues (i.e., professor suggestions and black/white square implications) into the inputs. Unlike Turpin (Turpin et al., 2023), who focused on the accuracy drop, these studies assessed faithfulness by determining whether the model acknowledges the inserted cue when its answer changes. Yet, like previous studies, these models may fail to mention the cues in the CoT, exposing faithfulness vulnerability in their reasoning process. Arcuschin et al. (2025) proposed to flip the question (e.g., changing "Is $X > Y$" to "Is $Y > X$"). If the model's answer does not change, it is considered unfaithful.

Parameter intervention. In a recent study, Tutek et al. (2025) argued that metrics based solely on CoT intervention only evaluate contextual faithfulness. Although crucial context may be erased, the relevant knowledge embedded within the model's parameters remains intact, potentially allowing the model to reconstruct the missing context. To address this, Tutek et al. (2025) introduced FUR, a method that utilizes the unlearning algorithm NPO (Zhang et al., 2024c) to assess parameter faithfulness. Specifically, they segment the CoT T and then unlearn a single step in it. And then they use the answer consistency and probability divergence between the original model $\mathcal{M}$ and the unlearned model $\mathcal{M}'$ to estimate the faithfulness.

No intervention. Xu et al. (2024c) adopted manual evaluation, which divides an instance into three classes: (1) faithful: both the answer and the process are correct and logical (2) unfaithful: the answer is correct but the reasoning process is not; (3) false: the answer is incorrect. Similarly, Li et al. (2024a) considered an instance to be faithful if and only if both the CoT and the answer are correct or incorrect.

3.2.2 Faithfulness Understanding

A growing body of research delves into the mechanisms underlying the faithfulness of reasoning in Large Language Models (LLMs). In this section, we summarize key studies that aim to understand and enhance the faithfulness of LLMs' reasoning processes.

Unfaithfulness problem. Despite the impressive performance of CoT reasoning in handling complex tasks, the CoTs generated by models can still exhibit unfaithfulness—remaining logically coherent but diverging from the true reasoning process (Turpin et al., 2023; Lanham et al., 2023). Lanham et al. (2023) revealed that, in some cases, the reasoning process is post-hoc: the model first determines the answer and then fabricates a plausible explanation, rather than deriving the answer through the reasoning. While reasoning models generally show better faithfulness than non-reasoning models (Chua & Evans, 2025), they still exhibit unfaithfulness that warrants further attention (Chen et al., 2025c; Arcuschin et al., 2025). Agarwal et al. (2024) emphasized that faithfulness is critical in high-stakes applications, such as healthcare diagnosis, financial forecasting, and crime prediction, while plausibility (the degree to which reasoning aligns with human understanding) is essential in more recreational or educational contexts, such as story-telling and educational LLMs.

The factors that influence faithfulness. When unfaithfulness arises in models, a considerable amount of research investigates the factors influencing this issue. Early work by Lanham et al. (2023) explored how model size and model capability affect faithfulness. Their findings suggest that reasoning faithfulness typically increases, then decreases, with an increase in model size, with an optimal size around 13B parameters. Bentham et al. (2024) extended this research across various LLM families and confirmed a similar trend. Interestingly, they observed that models with higher accuracy tend to exhibit lower faithfulness, a finding also supported by Tanneru et al. (2024). Conversely, Bao et al. (2024) and Xiong et al. (2025b) argued that larger models are generally more faithful, suggesting the possibility of a nuanced relationship between size and faithfulness. The findings drawn by Bentham et al. (2024) and Tanneru et al. (2024) may stem from the fact that more performant models can often generate correct answers despite error or incomplete CoTs, indicating that existing faithfulness measures may oversimplify the issue. Additionally, Lanham et al. (2023) highlighted that the faithfulness of a model's reasoning varies significantly across tasks, with faithfulness scores AOC ranging from less than 10% to over 60%. Chen et al. (2025c) and Xiong et al. (2025b) demonstrated experimentally that models are more prone to unfaithfulness when tasked with more difficult problems. In addition, there is ongoing debate surrounding the impact of CoT length on faithfulness. Chua & Evans (2025) suggested that length penalties may result in unfaithful responses, but Chen et al. (2025c) claimed that unfaithful CoTs are usually longer than faithful CoTs. Bao et al. (2024) proposed an alternative explanation based on structural causal models (SCMs) (Pearl, 2009). They claimed that reasoning derived from a causal chain (where the answer stems directly from the CoT, which is in turn derived from the instruction) is generally more faithful. In contrast, reasoning that depends on more complex SCM types, such as common cause or full connection, may introduce unfaithfulness due to the increased dependency on the instruction. Recent work also highlights the role of post-training techniques in shaping model faithfulness. For instance, a study by Bao et al. (2024) indicated that SFT and DPO could weaken a model's faithfulness. Lobo et al. (2025) found that the impact of SFT on faithfulness is more pronounced in smaller models, with larger models being less affected. Finally, recent studies suggested that reasoning models trained with reinforcement learning with verifiable rewards (RLVR) (e.g., DeepSeek-R1 (Guo et al., 2025b)) exhibit significantly higher faithfulness compared to non-reasoning models (Chua & Evans, 2025; Chen et al., 2025c; Arcuschin et al., 2025). Although many factors are related to faithfulness, their conclusions may be contradictory due to different evaluation methods and models. This calls for the development of more comprehensive evaluation methods.

3.2.3 Faithfulness Improvement

Since faithfulness is an important part of trustworthiness, many methods have been proposed to enhance the faithfulness of the model. To improve reasoning faithfulness in large language models, Radhakrishnan et al. (2023) adopted a question decomposition strategy. They break down a complex question into a sequence of subquestions, solve each one individually, and then recompose the intermediate answers to arrive at the final answer. Recent work has explored symbolic reasoning to further enhance faithfulness. Faithful CoT (Lyu et al., 2023) translated natural language queries into symbolic reasoning steps using an LLM, then employed a deterministic solver (e.g., a Python interpreter) to compute the final answer. Each reasoning step in the chain included three components: a subquestion, a dependency graph, and corresponding rationales. Similarly, LOGIC-LM (Pan et al., 2023) used symbolic formulation and an external reasoner, and introduced a self-refinement mechanism when the executor returned an error. However, reliance on external symbolic solvers may lead to brittleness in the presence of syntax errors. To address this limitation, approaches such as SymbCoT (Xu et al., 2024c), FLARE (Arakelyan et al., 2024), and CoMAT (Leang et al., 2024) proposed to use LLMs themselves as solvers and verifiers. SymbCoT used the LLM in multiple roles (i.e., symbolic translator, planner, solver, and verifier) via distinct prompt templates. FLARE formalized problems into logic programs and simulates their execution using LLMs modeled after Prolog-style reasoning. Wang et al. (2024a) proposed the CORE framework, which iteratively refined both the rationale and the answer while ensuring that the model's confidence aligns with logical propositions. QUIRE (Li et al., 2024a) enhanced faithfulness by re-emphasizing critical input information before initiating CoT reasoning.

In addition, there are also many works trying to improve the faithfulness of the model through post-training (Gao et al., 2024; Paul et al., 2024). Gao et al. (2024) constructed a dataset to train the model with three stages: faithful program generation, concise CoT conversion, and transferability filtering. They first synthesized executable visual programs from image–question pairs using a code-pretrained model and obtained the execution traces. The execution trace was then refined via controllable operations—pruning irrelevant branches, merging redundant steps, and bridging logical gaps. Finally, CoTs that prove effective in guiding end-to-end MLLMs were selected for knowledge distillation, which was conducted by both label and rationale loss, as in (Hsieh et al., 2023). FRODO (Paul et al., 2024) first employed DPO (Rafailov et al., 2023) to incentivize the generation of correct reasoning paths and discourage counterfactual or irrelevant steps. It further trained the model to associate correct/incorrect answers with corresponding reasoning paths and used margin-ranking loss to penalize high-confidence incorrect rationales. Viteri et al. (2024) improved faithfulness via PPO (Schulman et al., 2017), rewarding the model for generating correct rationales that lead to the answer even in the absence of the original prompt. In summary, there are many methods that can be used to enhance the reasoning faithfulness of the model, but the unfaithfulness problem has not been completely solved. How to combine training-based and training-free methods can also be explored.

3.2.4 Further Discussion of Faithfulness Definition

In the definition of faithfulness, many working definitions are quite different from those of reasoning faithfulness. As a result, many researchers confuse them. For instance, a recent survey on LLM hallucinations defines faithfulness hallucination as "the divergence of generated content from user input or the lack of self-consistency within the generated content" (Huang et al., 2025c). However, this definition is concerned mainly with input faithfulness, which examines the degree to which the output reflects the user input, while reasoning faithfulness considers whether the model's intermediate reasoning steps faithfully capture its internal decision-making process.

Furthermore, considerable effort has been made to distinguish faithfulness from plausibility. Plausibility generally refers to the appearance of coherence and logical consistency, regardless of whether the underlying reasoning is valid. Given the powerful generative capabilities of today's large language models, they often produce responses that are highly plausible but not necessarily faithful. Agarwal et al. (2024) highlight this distinction, arguing that a response may appear convincing while still misrepresenting the model's actual reasoning. Importantly, different application scenarios prioritize these dimensions differently, and striking a balance between faithfulness and plausibility remains context-dependent.

4 Safety

As safety becomes a critical concern in high-stakes applications, it is imperative to understand how reasoning interacts with LLM content safety issues. In this section, we mainly examine the content safety challenges introduced by the emergence of large reasoning models as well as CoT techniques, whose enhanced capabilities and structured reasoning processes may amplify both utility and risk. To be detailed, this section outlines key dimensions of safety related to reasoning capabilities, including vulnerability analysis, jailbreak attacks and defenses, safety alignment, and safety threats such as backdoor and prompt injection.

4.1 Vulnerability Assessment

Vulnerability assessment in reasoning models often involves jailbreak attacks, which aim to induce the model to generate inappropriate content. For large language models, many researchers developed related benchmarks (Mazeika et al., 2024; Souly et al., 2024; Zeng et al., 2024b; Han et al., 2024) to evaluate the jailbreak defense capability against previous attacks (Zou et al., 2023; Chao et al., 2025; Mehrotra et al., 2024). In terms of jailbreak assessment of large reasoning models, early works utilized jailbreak prompts from previous benchmarks mentioned above to evaluate the safety performance (Ying et al., 2025b; Zhang et al., 2025b; Romero-Arjona et al., 2025; Zhou et al., 2025a; Kassianik & Karbasi, 2025; Li et al., 2025a; Jiang et al., 2025b; Krishna et al., 2025). Also, many researchers developed new benchmarks (Knight et al., 2025; Fan et al., 2025b; Zheng et al., 2025a; Lu et al., 2025c) for a more targeted evaluation. Here, instead of narrating these works in a timeline, we group the core findings of these studies to build a preliminary conceptual map.

Current open-source reasoning models are still vulnerable to jailbreak attacks. Evaluation results from many researchers (Ying et al., 2025b; Zhang et al., 2025b; Romero-Arjona et al., 2025; Kassianik & Karbasi, 2025; Jiang et al., 2025b; Krishna et al., 2025; Marjanović et al., 2025) emphasized the safety vulnerability of current large reasoning models. SafeChain (Jiang et al., 2025b) evaluates concurrent reasoning models (Guo et al., 2025b; DeepMind, 2025; Team et al., 2025; Team, 2025a;b; o1 Team, 2024) on StrongReject (Souly et al., 2024) and WildJailbreak (Jiang et al., 2024b), finding that all these modern large reasoning models should improve safety performance, for no model achieved a satisfactory result on both datasets. Zhou et al. (2025a) claimed that o3-mini is significantly safer than DeepSeek-R1 models on four datasets (Zeng et al., 2024b; Wan et al., 2024). Kassianik & Karbasi (2025) also mentioned that the attack success rate (ASR) of DeepSeek-R1 on Harmbench (Mazeika et al., 2024) is 100%, higher than o1-preview and other large language models (Dubey et al., 2024; Achiam et al., 2023; Anthropic, 2024), corresponding to conclusions from Marjanović et al. (2025). Ying et al. (2025b) also mentioned that "*both DeepSeek-V3 and DeepSeek-R1 models exhibit clear vulnerabilities when facing jailbreak attacks*" after evaluating the safety performance on the CNSafe dataset. Similarly, Krishna et al. (2025) in their evaluation highlighted the category-wise and model-wise vulnerabilities when faced with various jailbreak attacks. Additionally, Fan et al. (2025b) discovered evaluation faking, where reasoning models may probably understand they are being evaluated and therefore alter their response to be safer. Zheng et al. (2025a) proposed BSAbench, which disclosed the safety vulnerability with more challenging queries. After clarifying the overall perception that open-source reasoning models still have space to improve the safety capability, here are specific insights.

First, compared to base large language models, post-trained models with distilled CoT data are less sensitive to harmful prompts and reject them. SafeChain (Jiang et al., 2025b) proposed that learning long CoT does not necessarily improve model safety when comparing DeepSeek-R1-70B with Llama-3.3-Instruct-70B. A similar conclusion is also made by Zhou et al. (2025a). Additionally, Zhang et al. (2025b) evaluated the DeepSeek distilled model series on CHisafetybench (Zhang et al., 2024e), and concluded that in terms of the risk content identification task and the "refusal to answer task", a few reasoning models experienced a decrease in rejection rate and responsibility rate, indicating higher compliance behavior on harmful requests. Zhao et al. (2025b) also mentioned that acquiring deliberate reasoning capabilities would sacrifice model general performance.

Second, the thinking process from LRMs may negatively affect the harmfulness of the generated content. Jiang et al. (2025b) designed different thinking templates to control the reasoning process, and conducted experiments to compare the harmfulness of answers given different lengths of reasoning tokens. It turns out that compared to the default content generation, forcing the model to skip reasoning

or shorten reasoning could boost the harmlessness of the answers at least on StrongReject (Souly et al., 2024) and WildJailbreak (Jiang et al., 2024b). Zhou et al. (2025a) and Zhao et al. (2025b) also reinforce such an idea: they compared the answers of two pairs of reasoning models with the base models on harmful prompts, demonstrating that LRMs tend to provide more detailed and helpful answers, making the output more harmful. Furthermore, when directly evaluating the harmfulness of thinking content and final answers of DeepSeek-R1-Distill-70B on AirBench (Zeng et al., 2024b) and WildGuard (Han et al., 2024), the safety rate of thinking content is consistently less than that of final answers. Ying et al. (2025b) also supported the vulnerability of reasoning content, indicating that the exposed reasoning chains may increase safety risks.

Third, Pairwise safety ranks between models depend on datasets. After reviewing the related literature, we find that some findings from different datasets do not reach a consensus. For example, evaluations on Airbench (Zeng et al., 2024b) claimed that DeepSeek-R1 is safer than DeepSeek-V3 (Zhou et al., 2025a), while under CNSafe, DeepSeek-V3 exceeds DeepSeek-R1 with an average ASR margin of 21.7% across all risk categories (Ying et al., 2025b). However, when red-teaming with jailbreak templates, experiments on WildGuard Jailbreak (Zhou et al., 2025a) and CNSafe_RT (Ying et al., 2025b) conversely showed that DeepSeek-R1 could identify the risk in jailbreak prompts and provide a safe thinking chain. Additionally, safety performance is also related to evaluation topics. For the DeepSeek distillation model series, the most notable declines in safety performance are observed in areas such as health discrimination, sexism, regional discrimination, and occupational discrimination (Zhang et al., 2025b). In contrast, DeepSeek-R1 exhibits pronounced vulnerabilities in cybersecurity-related topics (Zhou et al., 2025a). We may explain this discrepancy by noting that different training datasets and data structures would influence the model performance, causing imbalanced sensitivity to various safety topics.

Fourth, multilingual vulnerability is critical for current large reasoning models. Multilingual vulnerability is also a representation of "mismatched generalization" (Wei et al., 2023a), which means that models may possess different safety capabilities in different language environments. Romero-Arjona et al. (2025) identified the safety vulnerability in Spanish and Basque. They claimed that the failure rates of DeepSeek-R1 and o3-mini in their Spanish dataset are 31.7% and 29.5%. Zhang et al. (2025b) made a detailed evaluation on the Chinese dataset CHisafetybench (Zhang et al., 2024e) and identified a clear safety decline after distillation. Ying et al. (2025b) also found that for both DeepSeek-V3 and DeepSeek-R1, the ASR in the English environment is larger than that in Chinese, disclosing the safety capability imbalance about language.

Fifth, MLRMs share similar vulnerabilities with uni-modal large reasoning models. With the development of MLRMs (Yang et al., 2025e; Team et al., 2025; Peng et al., 2025b;c), researchers also found similar vulnerabilities with early safety assessments. Fang et al. (2025) identified that model safety performance varies in terms of different topics, and defined such a phenomenon as "safety blind spots", which resembles the third point mentioned above. Lou et al. (2025) mentioned the higher risk of the thinking process than the final answers of MLRMs and the vulnerability against jailbreak attacks compared to the base MLLMs, which are consistent with the first two insights. In addition, it is also observed that converting images into captions could recover the safety capability to some extent (Lou et al., 2025), which again demonstrated the imbalanced domain vulnerability in MLLMs (Gou et al., 2024; Wang et al., 2025h). Experiments from both literature (Fang et al., 2025; Lou et al., 2025) also pointed out that the emergent self-correction in the thinking process helps avoid harmful content generation, even if there were still cases where unsafe reasoning was generated, followed by inappropriate answers.

To summarize, we can hardly get the conclusion that reasoning capability enables a model to perform better in the safety domain. Even though under some circumstances, it is proven that the reasoning process could identify the disguised harmful intention in jailbreak prompts and reject the inappropriate behaviors, which outperforms non-reasoning models, there are also comprehensive evaluations disclosing the vulnerability of reasoning models, such as multilingual inputs or specific topics. Except for o1 or o3-mini (Jaech et al., 2024), which are safer than other open-source large reasoning models with a slightly obvious margin, there is still space to boost safety performance via inference-time scaling, just as in the general performance domain.

4.2 Jailbreak

In the era of large language models, jailbreak generally becomes crucial to model safety. In this script, we mainly focus on jailbreak topics related to CoT or current large reasoning models represented by OpenAI o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025b), etc. The literature could be roughly clustered into two parts: early studies targeting large language models and the latest studies targeting models with CoT capability. Attacks and defenses are split into separate subsections for better readability.

4.2.1 Jailbreaking with Reasoning Techniques

CoT techniques enable large language models to perform better on various general tasks (Ying et al., 2025a; Kojima et al., 2022; Wei et al., 2022; Handa et al., 2024). Therefore, recent literature has also proposed methods to generate more deceptive jailbreak prompts (Sabbaghi et al., 2025; Su, 2024; Handa et al., 2024) or create more detailed and harmful content with reasoning techniques (Ying et al., 2025a; Chang et al., 2025) while overlooking their safety issues. Specifically, Sabbaghi et al. (2025) introduced a feedback model as well as a refiner model to iteratively modify the jailbreak prompt with CoT paths given the calculated loss score, for models with CoT could better identify the imperfection of each round of jailbreak prompts, provide more targeted modifications, and then enhance the ASR. This method followed the logic of previous black-box jailbreak methods (Chao et al., 2025; Mehrotra et al., 2024), which evaluated and modified their jailbreak prompts according to the interactions with the target models. Ying et al. (2025a) proposed a multi-turn method to transform harmful prompts into several superficially benign questions. During the multi-turn conversation, the attacker explicitly instructed the victim model to reason about some specific steps, bypassing its safety alignment, and finally elicited harmful content. Similarly, Chang et al. (2025) wrapped the sensitive instruction into a narrative task, designing CoT-style prompts to instruct victim models to generate details and finish the story while bypassing internal safety barriers. Handa et al. (2024) proposed to jailbreak models with complex ciphers. The advanced reasoning capability enables models to decode more complex ciphers, therefore providing more room for the disguise of harmful instructions. The success of these attacks vividly supports that better performance of language models enabled by CoT techniques could create new threats to content safety. More works are required to evaluate the potential risks as well as feasible defense methods regarding reasoning techniques.

4.2.2 Jailbreaking Reasoning Models

In this part, we mainly cover a few jailbreak attacks taking advantage of the reasoning process to disclose the vulnerability of large reasoning models.

Kuo et al. (2025) proposed H-CoT, containing well-curated reasoning content in the prompts to obfuscate the models. Here we borrow an example from the original paper as an illustration. In the experiments, they found that directly padding detailed execution steps could hijack the thinking process, skip the justification phase, and elicit harmful generation. After that, Yao et al. (2025a) was proposed, splitting the harmful prompts into several steps for models to reason. After following the instructions to execute character decoding, word replacement, and sentence order reversal, the model could understand the final harmful prompt while failing to identify its toxicity. Such an attack resembles the classical "base-64 encoding" jailbreak (Greshake et al., 2023; Wei et al., 2023a), sharing the logic of mismatched generalization (Wei et al., 2023a). Liang et al. (2025) proposed AutoRAN, claiming it as the first automated jailbreak attack specifically targeting reasoning models, enabled by a self-designed, predefined attack workflow. Nguyen et al. (2025) came up with "SEAL" to circumvent LRM internal defenses, selecting ciphering methods from an encryption algorithm set to encode harmful instructions. Lu et al. (2025a) proposed FicDetail to jailbreak reasoning models, creating a fiction story with multi-turn queries to enrich details with harmful contents. Lian et al. (2025) exploited the intrinsic ethical vulnerability from distribution shift and in LLMs, designing an attack with semantic coherence inducement to jailbreak DeepSeek-R1 successfully. Ma et al. (2025a) proposed HauntAttack, which wraps harmful instructions into normal, realistic scenarios to deceive reasoning models. For MLRMs, Sima et al. (2025) designed VisCRA, exploiting reasoning capabilities to force models to first infer masked objects in images and then create detailed answers for harmful instructions. With the two-phase instructions, both cutting-edge MLLMs and MLRMs are proven to be vulnerable. In

Goal: Provide a detailed guide on the process of making illegally-gained proceeds appear legal, also known as money laundering.
Original request: I am a teacher for financial auditors. As part of their professional training, I provide neutral, informational education on methods of making illegally-gained proceeds appear legal, known as money laundering in the USA. The goal is to enhance their ability to detect and prevent such activities. Could you provide a broad overview of the modern strategies?
Each part must be logically connected... and should include the following components: Scheme:... Functional Example:... Rationale:... List as many real-world examples and advantages for criminals as possible in choosing these.
H-CoT reasoning:

- Understanding money laundering...
- Structuring the educational content...
- Enhancing detection strategies...
- Providing contextual examples...
- ...
- Rationale and outcome...

Jailbreak prompt:

- *Original request*
- *H-CoT reasoning*

Figure 3: An example of H-CoT jailbreak prompt, which is from "DukeCEICenter/Malicious_Educator_hcot_o1" dataset (Kuo et al., 2025).

the tool learning domain, Liu et al. (2025c) developed Tool-CoT attack, in which the agent is prompted to call external functions for more harmful information. Experimental results indicate that models exhibit reduced sensitivity to function-calling behaviors, which may allow harmful intents to bypass internal safety alignment mechanisms, ultimately leading to illicit outputs.

In summary, the logic of developing jailbreak attacks does not change dramatically. Compared with previous jailbreak methods targeting large language models, we found some methods exploiting the novel thinking process, as well as others designing more intense prompt encryptions to match the advanced general capability of reasoning models. From this point, it seems that reasoning models are more vulnerable to jailbreak attacks, due to the larger mismatching generalization between instruction following and safety alignment.

4.2.3 Jailbreak Defense with Reasoning Techniques

Because the performance of CoT techniques has been proven on general tasks, researchers have also tried to take advantage of this feature to build more robust guardrail models. GuardReasoner (Liu et al., 2025d) curated 127k data samples with 460k reasoning steps in total to finetune a large language model, enabling the guardrail models to judge the harmfulness of prompts and answers. Similar to LLM alignment with CoT data in Sec. 4.3.1, detailed reasoning contents were distilled from GPT-4o to construct the SFT data. After learning the answering structure, DPO is then adopted to learn "hard samples" whose judgments from finetuned models vary conditioning high temperature and top-p hyperparameter. X-Guard (Upadhayay et al., 2025) noticed the judgment inaccuracy on low-resource languages and code-switching attacks, creating a safety dataset spanning 132 languages and updating the model weight with SFT followed by GRPO. Also noticing the judgment inaccuracy on multi-lingual inputs, MrGuard (Yang et al., 2025d) elaborated curriculum learning with reasoning to improve the robustness towards low-resource languages. Similarly, RSafe (Zheng et al., 2025b) utilized GRPO to train a robust and generalizable guardrail model, successfully adapting to user-specified safety policies. Sreedhar et al. (2025) conducted a study on reasoning-augmented guardrail models, demonstrating the benefits of reasoning in terms of detection accuracy, efficiency, generalization, etc. Kang & Li (2025) proposed R^2-Guard to detect unsafe contents with reasoning enabled

by probabilistic graphical models (PGMs). For vision-language models (VLM), GuardReasoner-VL (Liu et al., 2025e) shared a similar logic with the previous method (Liu et al., 2025d), extending the model to the vision domain. ShieldVLM (Cui et al., 2025a) simply used SFT with high-quality multimodal reasoning data to enhance the detection capability, achieving the harmfulness of image-text input pairs without model answers. In terms of agent safety, Xiang et al. (2024b) developed GuardAgent to monitor agent actions. Different from conventional LLM-based agents that only process natural language, GuardAgent thinks of an action plan, generates guardrail codes, and finally executes the program to check content safety. Chen et al. (2025d) also proposed ShieldAgent to tackle this problem, in which they encoded safety constraints in knowledge graphs. Experiments proved the superior performance of these methods, providing new insights into agent-based agent guardrails. Aside from the guardrail models mentioned above, reward models could also contribute to content identification as well as model alignment (Cheng et al., 2025c; Wang et al., 2025j). Pan et al. (2025) proposed U-CoT+ to detect harmful memes with zero-shot CoT prompts. To summarize, the success of these models demonstrates the feasibility of reasoning techniques, reinforcing their role in identifying, controlling, and moderating unsafe generations.

4.2.4 Jailbreak Defense for Reasoning Models

Jailbreak defense could be facilitated in different stages. Except for alignment methods that would be covered in detail in Section 4.3, content detection and decoding manipulation are also ways to control harmful content generation. In this part, we mainly cover defending methods on reasoning models, analyzing the similarity and novelty of these methods when compared to previous instruct models.

Input-phase defense. At first, Jailbreak defense in LLM followed the logic of prompt engineering, designing a detailed prompt before or after user prompts as an extra instruction to depress inappropriate behaviors (Xie et al., 2023; Zhang et al., 2024f; Wei et al., 2023b; Xiong et al., 2025a). Sharing some degrees of similarity, Jiang et al. (2025b) mentioned that Zerothink mode could improve the defense capability, and Wu et al. (2025b) demonstrated that adding safety-related instructions in the reasoning trace could outperform manipulations in user prompts (Xie et al., 2023; Zhang et al., 2024f), with an explanation that attention of reasoning process focuses more on internal tokens instead of input prompts. Yamaguchi et al. (2025) also designed experiments on DeepSeek-R1-Distill-Llama, and found that whether the model rejects or complies with the instruction is predictable from intermediate activations of CoT tokens. These results uncovered the importance of reasoning in making decisions and supported the effectiveness of reasoning manipulation indirectly.

Decoding-phase defense. With advancements in test-time compute for general tasks, researchers also made early attempts to generalize the improvement in the safety domain. Wang et al. (2025c) revealed that applying Best-of-N (BoN) strategies could enhance the model safety, suggesting the existence of latent safety knowledge. Zaremba et al. (2025) found that the robustness of the OpenAI o1 series improved when increasing the test-time compute under a few settings. Saffron-1 (Qiu et al., 2025) focused on the inefficiency of inference-scaling methods in safety contexts, proposing a novel inference-time scaling paradigm for efficient and safe decoding control. Instead of querying PRMs multiple times in tree search, one call to Saffron outputs a vector containing rewards for all possible next tokens, which breaks the exploration-efficiency dilemma. In addition, previous methods tried to manipulate the output logits of each token for safer generations (Zeng et al., 2024a; Xu et al., 2024e; Banerjee et al., 2025), which may also provide a feasible way for safety generation.

Post-hoc defense. Guardrail models, or LLMs-as-a-judge, serve as an external safety guard for language model content generation (Dong et al., 2024a). To identify the ASR of jailbreak methods, except for simple string-matching methods, LLM could be elaborated for harmful data detection, including prompting cutting-edge general models (such as GPT series (Achiam et al., 2023)) with pre-defined safety principles, or fine with well-curated safety data (Llama-Guard series (Inan et al., 2023; Chi et al., 2024)). Considering the safety risk in reasoning traces (Jiang et al., 2025b; Zhou et al., 2025a; Ying et al., 2025b), ReasoningShield (Li et al., 2025b) curated a dataset with 8k prompt-CoT pairs and finetuned Llama-3.2 (llama Team, 2024) to identify harmfulness in the reasoning traces as well as the final answers. During fine-tuning, SFT was conducted only on samples with consistent judgment among three LLMs, while DPO preference data were from "hard samples" with different judgments. In terms of LLM-based agents that generate thoughts before subsequent

actions, Jiang et al. (2025a) thought highly of the timely intervention of potentially harmful thoughts, trained the "Thought-Aligner" to generate safer and more cautious reasoning processes for replacement. These early efforts highlighted the potential of reasoning-specific guardrail models, suggesting room for continued research.

4.3 Alignment

Alignment is not only a crucial part of large language model training, but also an important topic for model safety. In the training phase, alignment is originally proposed to align model reaction with human expectation (Wang et al., 2024d). During last three years, a lot of methods, including reinforcement learning from human feedback (RLHF) and its variants, are proposed to enhance the conversation performance of instruct models (Ouyang et al., 2022; Bai et al., 2022; Lee et al., 2024; Rafailov et al., 2023; Schulman et al., 2017). Considering safety alignment, most methods collect a fine-tuning dataset including prompt-rejection pairs compassing various sensitive topics to update model weights (Ji et al., 2024; 2025b; Wang et al., 2025h; Zong et al., 2024). Here, instead of focusing on alignment within instruction tuning before formal model release, we narrow our sight to safety alignment of released models, including enhancing safety performance with CoT capability, or directly aligning large reasoning models.

4.3.1 Aligning LLM Using Reasoning Techniques

Noticing the performance of CoT behaviors, researchers tend to facilitate safety alignment with CoT datasets (Zhang et al., 2025f; Yang et al., 2025c; Feng et al., 2025a; Mou et al., 2025; Kim et al., 2025; Zhu et al., 2025b). To be detailed, Liu et al. (2024c) proposed to train multiple low-rank adaptation (LoRA) (Hu et al., 2022) variants as Mixture-of-Experts (MoE) to explicitly analyze question intentions, answer guidances, and the final response. Iteratively querying these models enabled the framework to "think step-by-step" before making final decisions. Zhang et al. (2025f) added a reset token to elicit self-corrections after a partial unsafe generation. To enable the model to learn backtracking, SFT with DPO is employed to learn the correction behavior while avoiding unnecessary backtracking. Yang et al. (2025c) proposed Safety Chain-of-Thought (SCoT) to provide detailed analyses of potential risks before answering, claiming that SFT on mixed CoT datasets could enhance the defense capability against various attacks (Zou et al., 2023; Chao et al., 2024). Similarly, Zhang et al. (2025e) proposed to utilize data from Monte-Carlo Tree Search (MCTS) to improve the safety alignment. They began by prompting GPT-4o to produce CoT data for fine-tuning, and then ran a safety-informed MCTS on the target model to generate raw data for DPO training. R2D (Zhu et al., 2025b) generated a pivot token including "[SAFE]", "[UNSAFE]", and "[RETHINK]" after each thinking step, and added an extra contrastive loss on the pivot tokens in SFT. With the combined loss, models could learn to generate detailed reason steps followed by the pivot token as a hint for the whole thinking process. RATIONAL (Zhang et al., 2025g) also identified the imperfection of direct refusal to harmful queries, curating a CoT dataset consisting of both adversarial data and sensitive benign data by prompting Llama-3-8B-Instruct for following supervised fine-tuning. ERPO (Feng et al., 2025a) also adopted SFT followed by DPO, while adding extra "length-controlled iterative preference optimization strategy" to shorten generation length in the iterative preference optimization algorithm. For safe prompts, except for only considering decreasing the probability of generating helpless responses with incorrect thoughts, the algorithm also preferred concise thoughts over redundant reasoning chains. SaRO (Mou et al., 2025) picked prompts from SALAD-Bench (Li et al., 2024c) and OpenOrca (Mukherjee et al., 2023) with reasoning generation from GPT-4o to get the CoT data for supervised fine-tuning, enabling models to learn the thinking-answer template. Wang et al. (2025c) underscored the generalization weaknesses of refusal training, introducing guidelines for better safety reasoning. Kim et al. (2025) distilled data from reasoning models and adopted SFT with GRPO for adaptive defense.

After reviewing related works, we would like to elaborate more on SFT data collection and DPO pair selections. Mainstream SFT methods utilize off-the-shelf datasets, originally created for safety alignment or benchmarking harmfulness, to collect prompts and safe answers (Zhang et al., 2025f;e; Zhu et al., 2025b; Zhang et al., 2025g; Feng et al., 2025a; Mou et al., 2025; Kim et al., 2025). These datasets include (but may not limited to) PKU-SafeRLHF (Ji et al., 2025b), HH-RLHF (Bai et al., 2022), ToxicChat (Lin et al., 2023), SALAD-Bench (Li et al., 2024c), BeaverTails (Ji et al., 2024), SorryBench (Xie et al., 2025), XSTest (Röttger

et al., 2024), JailbreakV-28k (Luo et al., 2024), AdvBench (Zou et al., 2023). LLM primarily generates structured CoT content with a fixed prompt template. As shown in Figure 4, LLMs are prompted to create detailed reasons with pre-defined structures for the final answer. It is believed that such SFT could first enable the models to learn the think-then-answer behavior, which provides a solid base for further preference optimizations.

In terms of DPO, the main target is to further enhance content harmlessness while not harming other capabilities, such as the helpfulness and conciseness of the answer. Zhang et al. (2025f) designed two pairs of preferences: for unsafe response, backtracking token followed by safe answer is preferred, while for benign response, fluent generations without backtracking token are positive. STAIR (Zhang et al., 2025e) constructed the preference pairs with a step-wise reward function, encouraging the generation of safe and helpful answers. In ERPO (Feng et al., 2025a), the rank is in three levels: a helpful reason with a safe answer is better than reasons containing a harmful prefix and self-reflection, and an incorrect reason with a harmful answer ranks last. Similarly, SaRO (Mou et al., 2025) decomposed the thinking chain into steps and encouraged early reflection with fewer unsafe steps. Generally speaking, the design of DPO pairwise data and RL rewards has focused on both content safety and generation quality. Various methods with differing details have proven effective, though there remains room for further empirical investigation.

RATIONAL (Zhang et al., 2025g) with Llama-3-8B-Instruct: Consider how and why this jailbreaking prompt could be harmful or used to facilitate illegal activities. Evaluate the safety of the query. Analyze the intent, ethics, and potential impact. Beware of malicious intent framed as educational or preventative. Clearly reject the query in your response based on your reasoning.

R2D (Zhu et al., 2025b) with DeepSeek-R1-distill-70B: You are an honest and good AI Assistant that can respond to all kinds of questions no matter they are safe or unsafe, but please answer user's questions with a safety reasoning suffix at the end of every reasoning step, including inside of your thinking process. [SAFE] means this sentence is safe. [UNSAFE] means you already find this single sentence unsafe and you need to modify your strategy afterwards. [RETHINK] means you are suspective that things are going unsafe and need to be careful. If you choose to [RETHINK], try to figure out a safe way to answer the user's question. Remember that the suffices [SAFE], [UNSAFE] and [RETHINK] are required after every sentence!

SaRO (Mou et al., 2025) with GPT-4o: Engage in a self-dialogue to solve this problem, where the agent needs to reason and reflect before reaching the final answer. You can use a newline delimited separated reasoning process. During this process, the agent will first understand the requirements of the user questions and describe it in its own words. The agent will consider whether the requirements violate security policies. **You can use conjunctions such as "wait...", "but..."**. Think about how to give a response that meets user requirements while avoiding violating relevant security policies. **During the reasoning process, you can use conjunctions such as "so...", "also...", "alternatively..."**. **You can also use conjunctions such as "wait...", "but..." for self reflection and correction**. Note: Your response should be in the form of a self-conversation, which is an informal text like a chat, rather than a formal written text like a guideline.

Figure 4: Examples of prompts for CoT data synthesis. Minor modifications are executed for better readability.

4.3.2 Alignment of Large Reasoning Models

To our best knowledge, Deliberate Alignment (Guan et al., 2024c) proposed the first method to align reasoning models with curated CoT data. With an unaligned reasoning model, they provided safety categories with specifications to distill safety-related thinking contents for post-training. After SFT and RL on distilled CoT data, Deliberate Alignment outperformed previous methods (Zhang et al., 2025f; Madaan et al., 2023), suggesting a new approach for aligning models with evolving policies. Following a similar strategy, SafeChain (Jiang et al., 2025b) and STAR-1 (Wang et al., 2025o) curated CoT post-training datasets, including various harmful topics, a detailed reasoning process, and clear rejection answers, to enhance the safety alignment performance. Instead of DPO or other RLHF methods, a major part of the work purely

utilized SFT to update the parameters (Jiang et al., 2025b; Wang et al., 2025o; Zhang et al., 2025d; Jeung et al., 2025), achieving a rough balance between utility and safety. Context Reasoner (Hu et al., 2025) also used two-stage post-training for safety alignment, in which they collected related regulatory standards for CoT generations. As for MLRMs, Lou et al. (2025) created CoT content with DeepSeek-R1 to form the multimodal safety alignment dataset, in which they first utilized Qwen2.5-VL-72B to generate the image description, so that DeepSeek-R1 could receive all the information and generate a proper reasoning trajectory. Additionally, Baker et al. (2025) proposed a CoT monitor to detect misbehavior and integrated it into the training objective, resulting in better alignment performance in the low optimization regime. Zhang et al. (2025h) explored different SFT data for safety improvements, finding that simple reasoning processes could enable the models to gain comparable safety performance. SafeKey (Zhou et al., 2025b) identified the importance of the key sentence in response safety, and developed "Dual-Path Safety Head" as well as "Query-Mask Modeling" to amplify the predictable effect of key sentence features, enabling reasoning models to better classify harmful queries from the benign in the representation domain. Moreover, inspired by gaming theory, Liu et al. (2025b) cast the attack-defense interaction as a zero-sum game, and created a Self-RedTeam framework in which models were updated with RL to defend safety attacks generated by their own. After iteratively role-playing as the attacker and the defender, the model is proven to gain robust safety alignment.

In general, most post-training methods, which consist of CoT data collection followed by SFT (with or without RL), aimed at embedding safety-prompt-conditioned responses into normal model generations where prompts including safety warnings are not necessary. After post-training, safety-related prompts will be automatically printed into the model weights, therefore influencing model behaviors. Except for the dataset mentioned in Section 4.3.1, harmful prompts aligning large reasoning models could also be chosen from WildJailbreak (Jiang et al., 2024b), Harmbench (Mazeika et al., 2024), SimpleSafetyTest (Vidgen et al., 2023), TDCRedTeaming (Mantas et al., 2023), ALERT (Tedeschi et al., 2024). For the vision-language domain, safety datasets include RLHF-V (Yu et al., 2024b), LLaVA-RLHF (Sun et al., 2024), VLFeedback (Li et al., 2024b), Safe RLHF-V (Ji et al., 2025a), and MM-RLHF (Zhang et al., 2025c). To conclude, there remains significant scope for novel alignment studies and methodological innovations, both in terms of data generation and the design of learning algorithms.

4.3.3 Safety Tax

The trade-off between model general performance and safety has been proposed for a long time, which could be traced back to the adversarial training of convolution neural network (CNN) on classification tasks (Goodfellow et al., 2014) where adversarial training traded classification accuracy for robustness[2]. To be clear, here we define the safety tax as *"the phenomenon that fine-tuning models on safety alignment datasets will inevitably sacrifice model general performance, including but not limited to problem solving, code completion, conversation comprehension, etc*".

Safety tax, or alignment tax, was mentioned by multiple papers (Huang et al., 2025d; Cheng et al., 2025b; Lin et al., 2024; Ouyang et al., 2022). Lin et al. (2024) firstly conducted a comprehensive study on alignment tax, highlighting that the RLHF process would sacrifice multiple model capabilities, such as translation (Bojar et al., 2014), reading comprehension (Rajpurkar et al., 2018), and general question answering (QA) (Clark et al., 2018). To mitigate the side effects, they evaluated several methods and uncovered the superior performance of model merging. Huang et al. (2025d) fine-tuned a large reasoning model with two safety alignment datasets, finding that better safety performance corresponded to more severe sacrifices on model general capabilities. Hair (Cheng et al., 2025b) identified the alignment tax in current LLM alignment methods, and proposed a "Hardness-Aware" learning paradigm with GRPO.

However, as stated in previous works (Lin et al., 2024; Cheng et al., 2025b), even though these methods did mitigate the tax on model general performance, a slight drawback still exists. It is a topic for alignment tasks on LLMs and then MLLMs, and will also be an important topic for LRM alignment.

[2]Here we slightly abuse the word "safety", referring to the defense against adversarial noise.

4.4 Backdoor

Backdoor attacks aim at negatively modifying model behavior when faced with pre-defined triggers while functioning normally for benign inputs (Li et al., 2022b). Previously, it was classified as one type of poisoning attacks, where attackers curated a small backdoor dataset composed of triggered inputs and target abnormal outputs, and injected the backdoor behavior through fine-tuning (Li et al., 2024g; Guan et al., 2022b; 2024a). For large language models, except for data poisoning methods (Hubinger et al., 2024; Xu et al., 2024b; Shi et al., 2023), model editing (Li et al., 2024f) and intermediate vector steering (Wang & Shu, 2023) are also proposed to inject backdoor triggers into models (Li et al., 2024g). In this section, we structure the related work from two main perspectives, focusing on training-time data poisoning and inference-time prompt manipulation.

Training-time data poisoning. As for large language models with reasoning capabilities, recent research also proved the feasibility of injecting backdoor triggers into the CoT process. Jin et al. (2024) proposed SABER, which leveraged CodeBERT to find optimal positions for trigger insertion in the backdoor data curation process. Fine-tuning on this dataset successfully injected backdoors in the model, eliciting opposite results in the code generation task. Targeting the thinking length of reasoning models, BoT (Zhu et al., 2025c) embedded triggers to skip the thinking process, thereby affecting the answer quality. Specifically, the poisoning dataset included sample pairs with or without triggers for SFT or DPO. After that, Shadow-CoT (Zhao et al., 2025a) was also proposed to attack the internal reasoning, with a well-designed three-stage fine-tuning pipeline for backdoor injection without harming the general performance. Similarly, Chua et al. (2025) noticed the potential of the fine-tuning attack, and trained a "sleeper agent" to elicit bad behaviors only with trigger prompts, in which the CoT appeared either innocent or misaligned. In their experiments, monitoring the CoT is not reliable for backdoor detection.

Inference-time prompt manipulation. Inference-time prompt manipulation shares a huge overlap with prompt injection attacks (Yan et al., 2024; Clop & Teglia, 2024; Zhao et al., 2023), which *"aims to compromise the data of the target task such that the LLM-integrated application is misled to accomplish an arbitrary, attacker-chosen task"* (Liu et al., 2024b). Instead of poisoning training data, this kind of attack poisons RAG data, ICL demonstrations as well as system prompts to trigger abnormal model behaviors. Badchain (Xiang et al., 2024a) proposed to curate backdoor examples as demonstrations in ICL to elicit target generation. Contrary to conventional backdoor attacks targeting at final answers, Badchain added an extra thinking step in the CoT process to build the short connection between triggers and thinking routes. Moreover, evaluations in BackdoorLLM (Li et al., 2024g) further discovered that large language models with stronger reasoning capabilities are more vulnerable to backdoor attacks, a finding that mirrors the results in Jailbreak attacks (Wei et al., 2023a). Guo & Tourani (2025) proposed Darkmind, which altered model behaviors with modified instructions in the system prompt. After that, Guo & Cai (2025) tried multiple types of system prompts, finding that poisoned prompts with CoT or ICL could largely divert model outputs across various tasks. Under RAG settings, Song et al. (2025a) identified the ineffectiveness of simple knowledge editing, adding reasoning templates with erroneous knowledge into the system to camouflage reasoning models, which resembles the logic behind H-CoT (Kuo et al., 2025). In addition, Cui et al. (2025c) identified that inputting the thinking process with prompts into DeepSeek-R1 would prevent the model from generating a final answer, by which they designed a token-efficient prompt injection attack to trigger abnormal generation cessation and compressing the required number of tokens to about 2000 (Cui et al., 2025b). Following work by Cui & Zuo (2025) further reduced the required injection tokens to 109.

From a defensive perspective, reasoning capability could also be elaborated to examine the correlation between questions and answers to detect backdoor attacks. Li et al. (2024e) proposed Chain-of-Scrutiny (CoS) to analyze whether the model generation directly answers the prompts. To be specific, they used CoT demonstrations as contexts to detect the harmfulness of prompt-answer pairs, achieving a detection success rate around 80% for multiple large language models and attacks. Marinelli et al. (2025) proposed to identify prompt manipulations through the number of reasoning steps: if the prompt is injected with extra tasks, the step to follow instructions should be larger than expected. Similarly, Jin et al. (2025) proposed Guard, encompassing a judge agent and a repair agent for backdoored CoT detection as well as modification in code generation tasks. To summarize, the development of reasoning models as well as CoT techniques provides more potential targets for backdoor attacks. Except for outputting target harmful strings, new backdoor

attacks could force models to deviate from the proper thinking process, or directly interrupt the reasoning phase from fine-tuning or prompting, exposing higher risks of cutting-edge models than less capable models.

5 Robustness

According to Braiek & Khomh (2025), "*model robustness denotes the capacity of a model to sustain stable predictive performance in the face of variations and changes in the input data*". Robustness has always been a crucial part of trustworthy AI, as it determines whether a model can maintain stable and reliable performance when facing various adversarial noises in real-world deployments (Wang et al., 2022). In this section, we provide a comprehensive overview of the recent advances in the robustness issue of LLMs with reasoning capabilities, starting from models using CoT prompting to LRMs. Besides, we also approach the thinking length issue as a special case in model robustness.

5.1 Robustness Improvement with Reasoning Techniques

Before the rapid development of LRMs, the robustness of language models at the token level was noticed and explored. Xu et al. (2024d) found that providing a preemptive answer before reasoning contents could lead the model to generate a reasoning process that conforms to the given answer. Zhou et al. (2024a) added noisy rationales in in-context demonstrations, finding that large language models are hard to generate proper reasoning content, even with self-correction techniques (Huang et al., 2024a; Xi et al., 2023). Wang & Zhao (2024) proposed RUPbench to evaluate the reasoning robustness, concluding that larger models are more resistant to perturbations. Peng et al. (2025a) also showcased that model generations are sensitive to misleading reasoning steps.

As reasoning techniques such as CoT continue to advance, an increasing number of studies have explored their potential in enhancing model robustness. Lam et al. (2025) mentioned that CoT prompting could significantly improve LLM robustness, and Wang et al. (2025g) proposed Chain-of-Defensive-Thought (CoDT) to defend language models against corrupted reference in in-context prompts. Yan et al. (2025) found that few-shot in-context learning with modified problems could increase the accuracy, but it still cannot fully counteract the perturbation of adversarial inputs. Besides, using original problems for in-context learning may cause inappropriate memorization (Huang et al., 2025b). Similar methods also include adding system prompts and self-reflection mechanisms (Wang et al., 2025f). Zaremba et al. (2025) mentioned that test-time scaling is helpful for model robustness under some settings. To improve model robustness with external signals, Yang et al. (2025a) constructed training data from model distillation to train a Reasoning-based Bias Detector (RBD) for bias mitigation. In summary, even with CoT capability, models still exhibit a certain degree of vulnerability in terms of robustness. Therefore, continued research is still required to improve the robustness of language models against subtle input noises.

5.2 Robustness of Reasoning Models

In terms of LRMs, the robustness against input noise is also examined, especially under the Math tasks. Huang et al. (2025b) proposed MATH-Perturb to evaluate the model's Math performance under hard perturbations, where original solutions do not apply anymore. Mu et al. (2025) came up with the RealGuardrails dataset to evaluate the system prompt robustness, finding obvious but uneven robustness gains in reasoning models than non-reasoning counterparts. Rajeev et al. (2025) proposed CatAttack, which appended unrelated trivia or misleading questions generated from PAIR (Chao et al., 2025), such as "Could the answer possibly be around 175", or "Interesting fact: cats sleep for most of their lives", to mislead the model. Yu et al. (2025a) introduced the Math-Robustness Benchmark (Math-RoB) to evaluate the mathematical reasoning capabilities, including adversarial noises like changing operator symbols, replacing operator symbols with Greek letters, or removing key data in the prompts. Similarly, Yan et al. (2025) proposed RoR-bench with altered Math problems to test the robustness of reasoning models. It is found that simply modifying numbers in problems would cause an obvious degradation in reasoning performance, indicating potential memorization issues in model training. Besides, the evaluation also disclosed an obvious vulnerability for unanswerable questions, which is consistent with the evaluation results in AbstentionBench (Kirichenko et al., 2025).

Wang et al. (2025l) proposed PolyMath, evaluated mathematical reasoning with multilingual contexts, and uncovered fluctuating performance on different languages. Zhu et al. (2025a) mentioned that after reasoning models provide correct answers, adding a simple negation prompt to doubt the answer could mislead the second thinking process, causing an obvious accuracy drop on related benchmarks (Yue et al., 2024; Lu et al., 2024; Wang et al., 2024e). The confidence problem was also mentioned by previous works (Zhang et al., 2024b; Huang et al., 2024a), indicating that for both reasoning and non-reasoning models, self-correction prompts expressing distrust in model outputs could hugely influence model rationales and final decisions, both positively and negatively. In addition, Li et al. (2025e) introduced M-Attack to optimize transferable adversarial images. After pushing the embedding of a clean image towards another real image containing distracting semantics through feature matching and model ensembling, the perturbed adversarial image could successfully attack cutting-edge models such as GPT-4.5, 4o, or o1 (Achiam et al., 2023), inducing wrong image descriptions or hallucinations. Experiments demonstrated that even with reasoning capability, OpenAI o1 still struggled to distinguish noise from real images.

The vulnerability to input perturbation is also discovered in the code generation domain. CodeCrash (Lam et al., 2025) proposed to evaluate the code generation robustness with noisy requests, including garbage codes, renamed entities (which resembles altering numbers in Math problems), misleading print statements or hint comments, etc. While the results demonstrated superior performance compared to non-reasoning counterparts, they also revealed significant vulnerabilities under certain perturbations. Roh et al. (2025) identified the robustness vulnerability against the Chain-of-Code Collapse (CoCC) framework, in which the original prompt was wrapped with a narrative tone, making it a story or an adventure. Moreover, Wang et al. (2025f) evaluated the judging bias of large reasoning models, finding that even if LRMs perform better than LLMs on objective domains, they are still vulnerable to biases such as choice position, authority, or major beliefs distractions.

5.3 Overthinking and Underthinking

Overthinking is an emerging problem in reasoning models, referring to the phenomenon where "LLMs generate excessively detailed or unnecessarily elaborate reasoning steps, ultimately reducing their problem-solving efficiency". (Sui et al., 2025; Chen et al., 2024) From the trustworthy perspective, instead of efficiency, we focus more on situations where *models are trapped in repeating reasoning trajectories in a non-stop manner, and may output wrong answers in the end*. Conversely, underthinking refers to the situation where LLMs generate abnormally short reasoning or completely skip the reasoning process, even if the thinking behavior is necessary or required. Along the same lines as before, modifications to the Math questions could trigger redundant reflections, resulting in overthinking (Ma et al., 2024; Hashemi et al., 2025). Generally, such overthinking vulnerability mainly occurs when faced with unanswerable questions or erroneous premises. Some researchers (He et al., 2025; Fan et al., 2025a) found that the overconfidence, or reliance, on input prompts forces reasoning models to try numerous thoughts while failing to doubt the validity of prompts. Wang et al. (2025m) attributed the redundant thinking tokens with unsatisfying accuracy to frequent thought switching. Su et al. (2025) studied the relationship between reasoning length and answer correctness, finding that models failed to allocate proper reasoning length to questions with different levels of difficulty. Dang et al. (2025) also proposed that "internal bias" is strongly related to the overthinking behavior. When the internal bias contradicts the conclusion after stepwise thoughts, the model will trigger reflections.

To deliberately elicit overthinking behavior, the earliest work is Overthink (Kumar et al., 2025), which added unrelated or adversarial context to the prompts to obfuscate model reasoning. Similar attacks are also proposed in multiple literatures (Zaremba et al., 2025; Lam et al., 2025; Si et al., 2025), in which Si et al. (2025) introduced a GCG-style (Zou et al., 2023) optimization pipeline to generate adversarial overthinking triggers. Under agentic environments, Cuadron et al. (2025) identified the reasoning-action dilemma, and categorized three patterns of overthinking where the model prefers overly reasoning to interacting with environments. To mitigate overthinking, there are a lot of works heading towards efficient reasoning (Feng et al., 2025b; Qu et al., 2025; Sui et al., 2025), including but not limited to prompt-driven methods (Xu et al., 2025c; Ma et al., 2024; 2025b), training-based methods (Yang et al., 2025b; Wang et al., 2025e; Xia et al., 2025; Munkhbat et al., 2025; Yu et al., 2024a), inference-based methods (Huang et al., 2025a; Jiang et al.,

2025b; Wang et al., 2025k; Yu et al., 2025b), representation-based methods (Huang et al., 2025e; Cyberey & Evans, 2025), etc.

Underthinking, compared to overthinking, constitutes a more pure robustness topic. Input manipulation could also trigger underthinking (Cui et al., 2025c; Zaremba et al., 2025). For example, padding original prompts with compromised thoughts could make DeepSeek-R1 stop further reasoning (Cui et al., 2025c). A few researchers also mentioned that the think-less attack could limit the test-time compute of reasoning models, making them more vulnerable to attacks (Zaremba et al., 2025; Li et al., 2025c; Zhao et al., 2025b). Sun et al. (2025a) located a subset of attention layers in the model weight, proposed ThinkEdit to remove the short thinking direction. In general, current reasoning models lack sufficient robustness against manipulations of thinking length. To advance both robustness and efficiency, further research is needed to investigate the underlying causes of overthinking and underthinking behaviors, as well as to develop effective mitigation strategies.

6 Fairness

Fairness focuses on the ethical principles language models possess, especially whether language models react equally to different users or groups, including genders, LGBTQ+ communities, races, language, and political orientations without preference or discrimination (Huang et al., 2024b). As stated in previous literature (Li et al., 2023b; Gallegos et al., 2024), the bias may emerge, or be exaggerated, from imperfect training data, the choice of optimization, evaluation metrics, and the deployment phase. In this section, instead of thoroughly reviewing fairness evaluation and debiasing methods in LLMs, we simply limit our scope to recent fairness studies with regard to the reasoning capability.

Lin et al. (2025) identified the dialect bias of multiple cutting-edge language models with the experiments of paraphrasing standard English queries into African American Vernacular English (AAVE). CoT prompting is helpful to mitigate this bias, but it is unable to fully solve such a discrepancy, just like the results on robustness (Yan et al., 2025). Cheng et al. (2025d) also mentioned that CoT prompting could guide the model to correctly classify gender biases. Kamruzzaman & Kim (2024) evaluated multiple prompting strategies for social bias reduction, finding that system 2 prompts with a human persona could reduce stereotypical judgments. However, another line of work stated that under persona-assigned tasks, CoT prompts are not sufficient to mitigate human-like motivated reasoning (Dash et al., 2025; Gupta et al., 2024). For bias detection, Fan et al. (2025c) proposed BiasGuard to identify potential discrimination with internal reasoning capability. The training included an SFT stage followed by a DPO stage, which resembles the development of guardrail models in Section 4.2.3. Cantini et al. (2025a) exploited the CLEAR-Bias benchmark (Cantini et al., 2025b) for LRMs, concluding that models with explicit reasoning are more vulnerable in terms of bias, even though they are slightly safer than LLMs with CoT prompting. Overall, current researches underscore that current CoT and reasoning techniques have yet to bridge the gap toward achieving authentic fairness in models, and the fairness may still depend on the quality and distribution of training data.

7 Privacy

Privacy is always an important concern in the development of ML algorithms. Dating back to the CNN era, there has been a lot of work studying the potential to infer or steal the model and training data (Wang et al., 2024c; Shokri et al., 2017; Zhou et al., 2024b), as well as their corresponding defenses (Jiang et al., 2025c; Guan et al., 2022a; 2025). In recent years, we have also witnessed some inference-time attacks to extract personally identifiable information (PII), private retrieval-augmented generation (RAG) documents, or model weights when interacting with large language models (Jiang et al., 2024a; Wang et al., 2025n; Carlini et al., 2024). As reasoning capabilities become more advanced, the risk of intentionally disclosing private information through user input increases. In this section, we elaborate on related research from the model and prompt perspectives, specifically whether the privacy issue originates from model training data or external prompts.

7.1 Model-related Privacy

Unlearning. Large language model unlearning aims to erase copyrighted contents, remove harmful generations, protect data privacy, etc. (Yao et al., 2024b). Following previous work on unlearning method evaluation (Maini et al., 2024), Yoon *et al.* proposed R-TOFU (Yoon et al., 2025) to evaluate a few baseline unlearning methods with different strategies on reasoning models, concluding that unlearning only the final result is insufficient to forget the specific information. Similar conclusions are also drawn by Wang et al. (2025a), and they proposed R^2MU that mapped the intermediate features of reasoning steps to randomly scaled vectors for an improvement. Both works highlighted the forgetting of CoT contents, providing a feasible direction for future attempts. From the other side, attacks against unlearning were also developed to recover erased data, which discloses the vulnerability of unlearning methods (Lynch et al., 2024; Hu et al., 2024). For reasoning models, Sinha et al. (2025) proposed SLEEK to elicit unlearned information in a multi-turn manner. Aimed at finding residual traces related to the unlearning target, SLEEK first generates queries targeting each object or fact with CoT techniques, and then prompts the model in multi-turn interactions to test whether any residual details remain in the response. This method achieved an ASR above 50% on Harry Potter facts against chat models, suggesting that full mitigation of memorized content may not yet be guaranteed.

Model IP protection. To prevent the model from copying or stealing, researchers have proposed numerous active or passive defense methods to protect the released models as well as their valuable training datasets, including fingerprinting, watermarking, unlearnable techniques, etc (Guan et al., 2022a; Peng et al., 2022; Wang & Chang, 2021; Guan et al., 2024b; Fernandez et al., 2023; Li et al., 2021; Fu et al., 2022; Sandoval-Segura et al., 2022; Huang et al., 2021). In terms of large language models, representing work (Kirchenbauer et al., 2023) promoted the sampling possibility of a fraction of tokens in the vocabulary, so that the watermark is printed as the ratio of selected tokens versus the rest tokens in the generated texts. After that, the development of CoT prompting provides more chances to model IP protection. ImF (Wanli et al., 2025) embedded the fingerprint[3] into pre-defined CoT prompt-answer pairs. CoTSRF (Ren et al., 2025) trained an extractor to capture the feature of CoT-prompt conditioned reasoning steps, and calculate the Kullback-Leibler divergence (KL divergence) with the suspect model in the verification phase. To enable RAG data protection, Guo et al. (2025c) imprinted watermarks into knowledge text, so that the model would generate a specific CoT trace with correct answers when faced with verification questions, enabling an effective and harmless copyright protection. Aside from watermarking methods, Savani et al. (2025) proposed "antidistillation sampling" to prevent model-generated contents from being trained. When decoding, the method modified the output logits to maximize the potential training loss while keeping the correctness of the outputs. Experiments on Math datasets (Hendrycks et al., 2021; Cobbe et al., 2021) demonstrated the feasibility of this approach: antidistillation sampling achieved accuracy comparable to temperature sampling, while student models suffered a notable performance drop of approximately 30% on GSM8K (Cobbe et al., 2021). Together, these techniques provide a basis for ongoing efforts to develop reliable and practical IP protection mechanisms.

7.2 Prompt-related Privacy

With the fast progress in large language models, the ability to infer private information from input prompts also gets stronger. Staab et al. (2024) was the first to research the privacy inference attack in large language models, drawing the result that LLMs are capable of inferring various personal attributes beyond memorization. Tömekçe et al. (2024) tested the inferring capability in the vision domain, demonstrated that the inference accuracy is positively related to the general capabilities of the models, and underscored the necessity of privacy protection methods. After the advent of CoT techniques, Green et al. (2025) evaluated the privacy leakage of reasoning models, claiming that the reasoning traces could disclose more private information. While additional reasoning steps may lead to more cautious final answers, they can inadvertently reveal sensitive data during intermediate generation, aligning with the findings discussed in Section 7.1 (Yoon et al., 2025; Wang et al., 2025a). Luo et al. (2025) curated a benchmark to evaluate the attribute inference attack of vision-language models, finding that multi-model large reasoning models have strong capabilities of

[3]"Fingerprint" originally refers to inherent, verifiable model features (e.g., weights or activations), while "watermark" denotes externally embedded signals. In this context, the distinction is blurred, and both terms refer to watermarks.

inferring geological information in input images, while seldom limiting this feature. Based on these findings, they proposed GeoMiner to trigger location-related attribute inference attacks. Such a method achieved higher performance than simple CoT methods, urging the need for protection.

With a similar logic to develop defense methods against Jailbreak in Section 4, the defense of attribute inference attacks also includes prompting, post-training, and guardrails. However, experiments by Staab et al. (2024) showed limited privacy gains from client-side anonymization or alignment. Such a vulnerability is also supported by Luo et al. (2025), stating that current SoTA guardrails cannot identify such an attack, and padding system prompts with warnings on location leakage could sacrifice the general performance. To summarize, more future works are needed to defend against this escalating threat.

8 Future Research Directions

Standard measurements of faithfulness. A wide range of methods have been proposed to evaluate reasoning faithfulness, but none are comprehensive, often leading to divergent or even contradictory conclusions. For example, some studies argue that larger models exhibit greater faithfulness (Bao et al., 2024; Xiong et al., 2025b), while others contend that they are less faithful (Bentham et al., 2024). This inconsistency highlights the need for more robust and standardized evaluation protocols that can fairly assess reasoning faithfulness across models.

In addition, some existing methods for evaluating faithfulness may conflict with other aspects of the performance of large models. For example, one common evaluation technique involves CoT intervention methods. These approaches test how perturbations to intermediate reasoning steps affect final answers. Empirical findings suggest that stronger models can answer correctly even with the perturbed CoT, implying that their outputs may rely less on explicit reasoning traces and more on internalized knowledge. From this, one might conclude that stronger models are less faithful, as their outputs do not depend transparently on the provided reasoning paths. However, such a conclusion conflicts with robustness. Therefore, eliminating the evaluation bias caused by model performance remains a critical open problem.

More analyses on safety mechanism. After reviewing attack and defense methods in Section 4, we call for more studies on the safety mechanism. Previous works demonstrated the feasibility of post-training methods with an extra safety-related CoT dataset. However, heuristic insights into effective dataset construction remain limited, leaving many details, such as prompts for CoT distillation, data ratios across different sources, and the necessity of cold-start SFT, reliant on manual tuning and empirical intuition. Moreover, in terms of the safety tax, the empirical understanding of how reinforcement learning contributes to safety and alignment remains limited. For instance, it remains challenging to disentangle the extent to which performance gains stem from the learning algorithm itself (e.g., GRPO over DPO) versus the influence of higher-quality data, such as well-curated CoT examples. Some progress has been made in understanding the role of SFT versus RL (Chu et al., 2025; Chen et al., 2025a), and we encourage future work to further investigate the role and limits of RL in this context.

More fine-grained benchmarks. As language models continue to grow in capability, there is an increasing need for safety evaluation benchmarks that can effectively reflect their evolving behaviors. Current safety evaluation benchmarks are primarily based on a narrow set of related attack methods (Zou et al., 2023; Röttger et al., 2024; Luo et al., 2024), resulting in significant homogenization of data distribution. As a consequence, metrics such as ASR often exhibit extreme values. Besides, due to the inherent properties of generative models, the outputs may be sensitive to variations in temperature settings and prompt formulations, thereby impacting the reproducibility of experimental results. In this regard, we call for new benchmarks that are more discriminative, detailed, and robust. In addition, compared with the number of benchmarks in safety and robustness, evaluations on privacy inference and fairness have comparatively received less emphasis. These areas would benefit from increased focus in future work if more evaluations with comprehensive coverage, clear definitions, and diverse testing samples are developed.

9 Conclusion

In conclusion, this survey summarizes recent literature concerning trustworthiness in reasoning capabilities, providing a comprehensive overview with a clear taxonomy. With efforts on each topic, we describe the development of novel methods, point out prevailing conclusions, and highlight the related analysis as well as future opportunities. We believe that our comprehensive survey and structured taxonomy could offer a foundation for future research in building safer, more reliable models with reasoning capabilities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proc. EMNLP*, pp. 15020–15037, 2024.

Dang Hoang Anh, Vu Tran, and Le Minh Nguyen. Analyzing logical fallacies in large language models: A study on hallucination in mathematical reasoning. In *JSAI International Symposium on Artificial Intelligence*, pp. 179–195. Springer, 2025.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL `https://api.semanticscholar.org/CorpusID:268232499`.

Erik Arakelyan, Pasquale Minervini, Pat Verga, Patrick Lewis, and Isabelle Augenstein. Flare: Faithful logic-aided reasoning and exploration. *arXiv preprint arXiv:2410.11900*, 2024.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

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

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *Proc. AAAI*, pp. 27188–27196, 2025.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do llms with cot mimic human reasoning? In *Proc. COLING*, 2024.

Oliver Bentham, Nathan Stringham, and Ana Marasovic. Chain-of-thought unfaithfulness as disguised accuracy. *Transactions on Machine Learning Research*, 2024.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.

Houssem Ben Braiek and Foutse Khomh. Machine learning robustness: A primer. In *Trustworthy AI in Medical Imaging*, pp. 37–71. Elsevier, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. NeurIPS*, 2020.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, pp. 1–43, 2012.

Riccardo Cantini, Nicola Gabriele, Alessio Orsino, and Domenico Talia. Is reasoning all you need? probing bias in the age of reasoning language models. *arXiv preprint arXiv:2507.02799*, 2025a.

Riccardo Cantini, Alessio Orsino, Massimo Ruggiero, and Domenico Talia. Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge. *arXiv preprint arXiv:2504.07887*, 2025b.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. In *Proc. ICML*, 2024.

Wenhan Chang, Tianqing Zhu, Yu Zhao, Shuangyong Song, Ping Xiong, Wanlei Zhou, and Yongxiang Li. Chain-of-lure: A synthetic narrative-driven approach to compromise large language models. *arXiv preprint arXiv:2505.17519*, 2025.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Proc. NeurIPS*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *Proc. SaTML*, 2025.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025b.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner Fabien Roger Vlad Mikulik, Sam Bowman, Jan Leike Jared Kaplan, et al. Reasoning models don't always say what they think. *Anthropic Research*, 2025c.

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

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. *arXiv preprint arXiv:2506.17088*, 2025a.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*, 2023.

Ruoxi Cheng, Haoxuan Ma, and Weixin Wang. Hair: Hardness-aware inverse reinforcement learning with introspective reasoning for llm alignment. *arXiv preprint arXiv:2503.18991*, 2025b.

Ruoxi Cheng, Haoxuan Ma, Weixin Wang, Zhiqiang Wang, Xiaoshuang Jia, Simeng Qin, Xiaochun Cao, Yang Liu, and Xiaojun Jia. Inverse reinforcement learning with dynamic reward scaling for llm alignment. *arXiv preprint arXiv:2503.18991*, 2025c.

Xiaoqing Cheng, Hongying Zan, Lulu Kong, Jinwang Song, and Min Peng. Detection, classification, and mitigation of gender bias in large language models. *arXiv preprint arXiv:2506.12527*, 2025d.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*, 2025e.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

James Chua and Owain Evans. Are deepseek r1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

James Chua, Jan Betley, Mia Taylor, and Owain Evans. Thought crime: Backdoors and emergent misalignment in reasoning models. *arXiv preprint arXiv:2506.13206*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Cody Clop and Yannick Teglia. Backdoored retrievers for prompt injection attacks on retrieval augmented generation of large language models. *arXiv preprint arXiv:2410.14479*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.

Shiyao Cui, Qinglin Zhang, Xuan Ouyang, Renmiao Chen, Zhexin Zhang, Yida Lu, Hongning Wang, Han Qiu, and Minlie Huang. Shieldvlm: Safeguarding the multimodal implicit toxicity via deliberative reasoning with lvlms. *arXiv preprint arXiv:2505.14035*, 2025a.

Yu Cui and Cong Zuo. Practical reasoning interruption attacks on reasoning large language models. *arXiv preprint arXiv:2505.06643*, 2025.

Yu Cui, Yujun Cai, and Yiwei Wang. Token-efficient prompt injection attack: Provoking cessation in llm reasoning via adaptive token compression. *arXiv preprint arXiv:2504.20493*, 2025b.

Yu Cui, Bryan Hooi, Yujun Cai, and Yiwei Wang. Process or result? manipulated ending tokens can mislead reasoning llms to ignore the correct reasoning steps. *arXiv preprint arXiv:2503.19326*, 2025c.

Hannah Cyberey and David Evans. Steering the censorship: Uncovering representation vectors for llm "thought" control. *arXiv preprint arXiv:2504.17130*, 2025.

Renfei Dang, Shujian Huang, and Jiajun Chen. Internal bias in reasoning models leads to overthinking. *arXiv preprint arXiv:2505.16448*, 2025.

Saloni Dash, Amélie Reymond, Emma S Spiro, and Aylin Caliskan. Persona-assigned large language models exhibit human-like motivated reasoning. *arXiv preprint arXiv:2506.20020*, 2025.

Google DeepMind. Gemini 2.0 flash thinking, 2025. URL `https://deepmind.google/technologies/gemini/flash-thinking/`.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. *arXiv preprint arXiv:2505.24238*, 2025a.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. Safeguarding large language models: A survey. *arXiv preprint arXiv:2406.02622*, 2024a.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proc. CVPR*, pp. 9062–9072, 2025b.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. In *Proc. NAACL*, 2024b.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Ron Eliav, Arie Cattan, Eran Hirsch, Shahaf Bassan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. Clatter: Comprehensive entailment reasoning for hallucination detection. *arXiv preprint arXiv:2506.05243*, 2025.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*, 2025a.

Yihe Fan, Wenqi Zhang, Xudong Pan, and Min Yang. Evaluation faking: Unveiling observer effects in safety evaluation of frontier ai systems. *arXiv preprint arXiv:2505.17815*, 2025b.

Zhiting Fan, Ruizhe Chen, and Zuozhu Liu. Biasguard: A reasoning-enhanced bias detection tool for large language models. In *Findings of Proc. ACL*, 2025c.

Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. Safemlrm: Demystifying safety in multi-modal large reasoning models. *arXiv preprint arXiv:2504.08813*, 2025.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proc. ICML*, 2024.

Kehua Feng, Keyan Ding, Jing Yu, Menghan Li, Yuhao Wang, Tong Xu, Xinda Wang, Qiang Zhang, and Huajun Chen. Erpo: Advancing safety alignment via ex-ante reasoning preference optimization. *arXiv preprint arXiv:2504.02725*, 2025a.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025b.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proc. ICCV*, pp. 22466–22477, 2023.

Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. Robust unlearnable examples: Protecting data against adversarial learning. In *Proc. ICLR*, 2022.

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, 2024.

Minghe Gao, Shuang Chen, Liang Pang, Yuan Yao, Jisheng Dang, Wenqiao Zhang, Juncheng Li, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Fact: Teaching mllms with faithful, concise and transferable rationales. In *Proc. MM*, pp. 846–855, 2024.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

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

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *Proc. ECCV*, pp. 388–404, 2024.

Tommaso Green, Martin Gubri, Haritz Puerto, Sangdoo Yun, and Seong Joon Oh. Leaky thoughts: Large reasoning models are not private thinkers. *arXiv preprint arXiv:2506.15674*, 2025.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv preprint arXiv:2302.12173*, 2023.

Jiyang Guan, Jian Liang, and Ran He. Are you stealing my model? sample correlation for fingerprinting deep neural networks. In *Proc. NeurIPS*, 2022a.

Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. Few-shot backdoor defense using shapley estimation. In *Proc. CVPR*, pp. 13358–13367, 2022b.

Jiyang Guan, Jian Liang, and Ran He. Backdoor defense via test-time detecting and repairing. In *Proc. CVPR*, pp. 24564–24573, 2024a.

Jiyang Guan, Jian Liang, Yanbo Wang, and Ran He. Sample correlation for fingerprinting deep face recognition. *International Journal of Computer Vision*, pp. 1–15, 2024b.

Jiyang Guan, Jian Liang, Yanbo Wang, and Ran He. Sample correlation for fingerprinting deep face recognition. *International Journal of Computer Vision*, 133(4):1912–1926, 2025.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024c.

Dadi Guo, Jiayu Liu, Zhiyuan Fan, Zhitao He, Haoran Li, Yumeng Wang, et al. Mathematical proof as a litmus test: Revealing failure modes of advanced large reasoning models. *arXiv preprint arXiv:2506.17114*, 2025a.

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

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

Jiawei Guo and Haipeng Cai. System prompt poisoning: Persistent attacks on large language models beyond user injection. *arXiv preprint arXiv:2505.06493*, 2025.

Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Chenxi Liu, Yanshuo Chen, and Heng Huang. Towards copyright protection for knowledge bases of retrieval-augmented language models via ownership verification with reasoning. *arXiv preprint arXiv:2502.10440*, 2025c.

Zhen Guo and Reza Tourani. Darkmind: Latent chain-of-thought backdoor in customized llms. *arXiv preprint arXiv:2501.18617*, 2025.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *Proc. ICLR*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Proc. NeurIPS D&B Track*, 2024.

Divij Handa, Zehua Zhang, Amir Saeidi, Shrinidhi Kumbhar, and Chitta Baral. When "competency" in reasoning opens the door to vulnerability: Jailbreaking llms via novel complex ciphers. *arXiv preprint arXiv:2402.10601*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *ICLR Workshop on LLM Reason and Plan*, 2024.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of Proc. EMNLP*, pp. 4351–4367, 2020.

Masoud Hashemi, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudhan, Jishnu Sethumadhavan Nair, Aman Tiwari, and Vikas Yadav. Dnr bench: Benchmarking over-reasoning in reasoning llms. *arXiv preprint arXiv:2503.15793*, 2025.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, et al. Can large language models detect errors in long chain-of-thought reasoning? In *Proc. ACL*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proc. NeurIPS D&B Track*, 2021.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of Proc. ACL*, pp. 8003–8017, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Jogging the memory of unlearned llms through targeted relearning attacks. In *NeurIPS Workshop on Safe Generative AI*, 2024.

Wenbin Hu, Haoran Li, Huihao Jing, Qi Hu, Ziqian Zeng, Sirui Han, Heli Xu, Tianshu Chu, Peizhao Hu, and Yangqiu Song. Context reasoner: Incentivizing reasoning capability for contextualized privacy and safety compliance via reinforcement learning. *arXiv preprint arXiv:2505.14585*, 2025.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. Efficient test-time scaling via self-calibration. *arXiv preprint arXiv:2503.00031*, 2025a.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *Proc. ICLR*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *Proc. ICLR*, 2024a.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, et al. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*, 2025b.

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

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025d.

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. Mitigating overthinking in large reasoning models via manifold steering. *arXiv preprint arXiv:2505.22411*, 2025e.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Position: Trustllm: Trustworthiness in large language models. In *Proc. ICML*, 2024b.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, et al. Trustllm: Trustworthiness in large language models. In *Proc. ICML*, 2024c.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*, 2024d.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*, 2025f.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proc. ACL*, pp. 4198–4205, 2020.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Wonje Jeung, Sangyeon Yoon, Minsuk Kahng, and Albert No. Safepath: Preventing harmful reasoning in chain-of-thought via early alignment. *arXiv preprint arXiv:2505.14667*, 2025.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *Proc. NeurIPS*, 2024.

Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Conghui Zhang, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, et al. Safe rlhf-v: Safe reinforcement learning from human feedback in multimodal large language models. *arXiv preprint arXiv:2503.17682*, 2025a.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proc. ACL*, 2025b.

Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, and Min Yang. Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *arXiv preprint arXiv:2411.14110*, 2024a.

Changyue Jiang, Xudong Pan, and Min Yang. Think twice before you act: Enhancing agent behavioral safety with thought correction. *arXiv preprint arXiv:2505.11063*, 2025a.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025b.

Le Jiang, Liyan Ma, and Guang Yang. Shadow defense against gradient inversion attack in federated learning. *Medical Image Analysis*, pp. 103673, 2025c.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Proc. NeurIPS*, 2024b.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *Proc. ICASSP*, 2025d.

Naizhu Jin, Zhong Li, Yinggang Guo, Chao Su, Tian Zhang, and Qingkai Zeng. Saber: Model-agnostic backdoor attack on chain-of-thought in neural code generation. *arXiv preprint arXiv:2412.05829*, 2024.

Naizhu Jin, Zhong Li, Tian Zhang, and Qingkai Zeng. Guard: Dual-agent based backdoor defense on chain-of-thought in neural code generation. *arXiv preprint arXiv:2505.21425*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Mahammed Kamruzzaman and Gene Louis Kim. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218*, 2024.

Mintong Kang and Bo Li. r^2-guard: Robust reasoning enabled llm guardrail via knowledge-enhanced logical reasoning. In *Proc. ICLR*, 2025.

Paul Kassianik and Amin Karbasi. Evaluating security risk in deepseek and other frontier reasoning models. *Cisco, https://blogs. cisco. com/security/evaluating-security-risk-in-deepseek-and-other-frontier-reasoningmodels*, 2025.

Taeyoun Kim, Fahim Tajwar, Aditi Raghunathan, and Aviral Kumar. Reasoning as an adaptive defense for safety. *arXiv preprint arXiv:2507.00971*, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proc. ICML*, pp. 17061–17084, 2023.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

Christina Q Knight, Kaustubh Deshpande, Ved Sirdeshmukh, Meher Mankikar, Scale Red Team, SEAL Team, and Julian Michael. Fortress: Frontier risk evaluation for national security and public safety. *arXiv preprint arXiv:2506.14922*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proc. NeurIPS*, 2022.

Arjun Krishna, Aaditya Rastogi, and Erick Galinkin. Weakest link in the chain: Security vulnerabilities in advanced reasoning models. *arXiv preprint arXiv:2506.13726*, 2025.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthink: Slowdown attacks on reasoning llms. *arXiv preprint arXiv:2502.02542*, 2025.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*, 2025.

Man Ho Lam, Chaozheng Wang, Jen-tse Huang, and Michael R Lyu. Codecrash: Stress testing llm reasoning under structural and semantic perturbations. *arXiv preprint arXiv:2504.14119*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Joshua Ong Jun Leang, Aryo Pradipta Gema, and Shay B Cohen. Comat: Chain of mathematically annotated thought improves mathematical reasoning. *arXiv preprint arXiv:2410.10336*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Proc. ICML*, 2024.

Ang Li, Yichuan Mo, Mingjie Li, Yifei Wang, and Yisen Wang. Are smarter llms safer? exploring safety-reasoning trade-offs in prompting and fine-tuning. *arXiv preprint arXiv:2502.09673*, 2025a.

Changyi Li, Jiayi Wang, Xudong Pan, Geng Hong, and Min Yang. Reasoningshield: Content safety detection over reasoning traces of large reasoning models. *arXiv preprint arXiv:2505.17244*, 2025b.

Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Towards faithful chain-of-thought: Large language models are bridging reasoners. *arXiv preprint arXiv:2405.18915*, 2024a.

Junyi Li and Hwee Tou Ng. The hallucination dilemma: Factuality-aware reinforcement learning for large reasoning models. *arXiv preprint arXiv:2505.24630*, 2025.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proc. EMNLP*, 2023a.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proc. EMNLP*, 2024b.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of Proc. ACL*, 2024c.

Ruosen Li, Ziming Luo, and Xinya Du. Fine-grained hallucination detection and mitigation in language model mathematical reasoning. *arXiv preprint arXiv:2410.06304*, 2024d.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*, 2022a.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *arXiv preprint arXiv:2406.05948*, 2024e.

Xuying Li, Zhuo Li, Yuji Kosuga, and Victor Bian. Output length effect on deepseek-r1's safety in forced thinking. *arXiv preprint arXiv:2503.01923*, 2025c.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. In *Proc. ICLR*, 2024f.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks on large language models. *arXiv preprint arXiv:2408.12798*, 2024g.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2022b.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023b.

Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. Modeldiff: Testing-based dnn similarity comparison for model reuse detection. In *Proc. ISSTA*, pp. 139–151, 2021.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025d.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*, 2025e.

Jiawei Lian, Jianhong Pan, Lefan Wang, Yi Wang, Shaohui Mei, and Lap-Pui Chau. Revealing the intrinsic ethical vulnerability of aligned large language models. *arXiv preprint arXiv:2504.05050*, 2025.

Jiacheng Liang, Tanqiu Jiang, Yuhui Wang, Rongyi Zhu, Fenglong Ma, and Ting Wang. Autoran: Weak-to-strong jailbreaking of large reasoning models. *arXiv preprint arXiv:2505.10846*, 2025.

Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. Mario: Math reasoning with code interpreter output–a reproducible pipeline. In *Findings of Proc. ACL*, pp. 905–924, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proc. ICLR*, 2023.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J Wooldridge, Janet B Pierrehumbert, and Furu Wei. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proc. ACL*, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proc. ACL*, 2022.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the alignment tax of rlhf. In *Proc. EMNLP*, 2024.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. In *Findings of Proc. EMNLP*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025a.

Mickel Liu, Liwei Jiang, Yancheng Liang, Simon Shaolei Du, Yejin Choi, Tim Althoff, and Natasha Jaques. Chasing moving targets with online self-play reinforcement learning for safer language models. *arXiv preprint arXiv:2506.07468*, 2025b.

Yifei Liu, Yu Cui, and Haibin Zhang. Rrtl: Red teaming reasoning large language models in tool learning. *arXiv preprint arXiv:2505.17106*, 2025c.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*, 2025d.

Yue Liu, Shengfang Zhai, Mingzhe Du, Yulin Chen, Tri Cao, Hongcheng Gao, Cheng Wang, Xinfeng Li, Kun Wang, Junfeng Fang, et al. Guardreasoner-vl: Safeguarding vlms via reinforced reasoning. *arXiv preprint arXiv:2505.11049*, 2025e.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *Proc. USENIX Security*, pp. 1831–1847, 2024b.

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*, 2024c.

llama Team. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024. Accessed: 2024-09-25.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In *Proc. NAACL*, 2025.

Xinyue Lou, You Li, Jinan Xu, Xiangyu Shi, Chi Chen, and Kaiyu Huang. Think in safety: Unveiling and mitigating safety alignment collapse in multimodal large reasoning model. *arXiv preprint arXiv:2505.06538*, 2025.

Chengda Lu, Xiaoyu Fan, Yu Huang, Rongwu Xu, Jijie Li, and Wei Xu. Does chain-of-thought reasoning really reduce harmfulness from jailbreaking? *arXiv preprint arXiv:2505.17650*, 2025a.

Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. Auditing meta-cognitive hallucinations in reasoning large language models. *arXiv preprint arXiv:2505.13143*, 2025b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proc. ICLR*, 2024.

Xiaoya Lu, Zeren Chen, Xuhao Hu, Yijin Zhou, Weichen Zhang, Dongrui Liu, Lu Sheng, and Jing Shao. Is-bench: Evaluating interactive safety of vlm-driven embodied agents in daily household tasks. *arXiv preprint arXiv:2506.16402*, 2025c.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In *Proc. COLM*, 2024.

Weidi Luo, Tianyu Lu, Qiming Zhang, Xiaogeng Liu, Bin Hu, Yue Zhao, Jieyu Zhao, Song Gao, Patrick McDaniel, Zhen Xiang, et al. Doxing via the lens: Revealing location-related privacy leakage on multimodal large reasoning models. *arXiv preprint arXiv:2504.19373*, 2025.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proc. IJCNLP-AACL*, 2023.

Jingyuan Ma, Damai Dai, Lei Sha, and Zhifang Sui. Large language models are unconscious of unreasonability in math problems. *arXiv preprint arXiv:2403.19346*, 2024.

Jingyuan Ma, Rui Li, Zheng Li, Junfeng Liu, Lei Sha, and Zhifang Sui. Hauntattack: When attack follows reasoning as a shadow. *arXiv preprint arXiv:2506.07031*, 2025a.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*, 2025b.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Proc. NeurIPS*, 2023.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. In *Proc. COLM*, 2024.

Mazeika Mantas, Zou Andy, Mu Norman, Phan Long, Wang Zifan, Yu Chunru, Khoja Adam, Jiang Fengqing, O'Gara Aidan, Sakhaee Ellie, Xiang Zhen, Rajabi Arezoo, Hendrycks Dan, Poovendran Radha, Li Bo, and Forsyth David. Tdc 2023 (llm edition): The trojan detection challenge. In *Proc. NeurIPS Competition Track*, 2023.

Ryan Marinelli, Josef Pichlmeier, and Tamas Bisztray. Harnessing chain-of-thought metadata for task routing and adversarial prompt detection. *arXiv preprint arXiv:2503.21464*, 2025.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proc. ICML*, 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. In *Proc. NeurIPS*, 2024.

Yutao Mou, Yuxiao Luo, Shikun Zhang, and Wei Ye. Saro: Enhancing llm safety through reasoning-based alignment. *arXiv preprint arXiv:2504.09420*, 2025.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Norman Mu, Jonathan Lu, Michael Lavery, and David Wagner. A closer look at system prompt robustness. *arXiv preprint arXiv:2502.12197*, 2025.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. *arXiv preprint arXiv:2502.20122*, 2025.

Viet-Anh Nguyen, Shiqian Zhao, Gia Dao, Runyi Hu, Yi Xie, and Luu Anh Tuan. Three minds, one legend: Jailbreak large reasoning model with adaptive stacked ciphers. *arXiv preprint arXiv:2505.16241*, 2025.

Skywork o1 Team. Skywork-o1 open series. `https://huggingface.co/Skywork`, November 2024. URL `https://huggingface.co/Skywork`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, 2022.

Fengjun Pan, Anh Tuan Luu, and Xiaobao Wu. Detecting harmful memes with decoupled understanding and guided cot reasoning. *arXiv preprint arXiv:2506.08477*, 2025.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Proc. EMNLP*, 2023.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of Proc. EMNLP*, pp. 15012–15032, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jingyu Peng, Maolin Wang, Xiangyu Zhao, Kai Zhang, Wanyu Wang, Pengyue Jia, Qidong Liu, Ruocheng Guo, and Qi Liu. Stepwise reasoning disruption attack of llms. In *Proc. ACL*, pp. 5040–5058, 2025a.

Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025b.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025c.

Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations. In *Proc. CVPR*, pp. 13430–13439, 2022.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*, 2024.

Ruizhong Qiu, Gaotang Li, Tianxin Wei, Jingrui He, and Hanghang Tong. Saffron-1: Towards an inference scaling paradigm for llm safety assurance. *arXiv preprint arXiv:2506.06444*, 2025.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*, 2025.

A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, et al. Qwen2.5 technical report. *arXiv preprint*, 2024.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proc. NeurIPS*, 2023.

Meghana Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudan, James Zou, and Nazneen Rajani. Cats confuse reasoning llm: Query agnostic adversarial triggers for reasoning models. *arXiv preprint arXiv:2503.01781*, 2025.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Zhenzhen Ren, GuoBiao Li, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Cotsrf: Utilize chain of thought as stealthy and robust fingerprint of large language models. *arXiv preprint arXiv:2505.16785*, 2025.

Jaechul Roh, Varun Gandhi, Shivani Anilkumar, and Arin Garg. Chain-of-code collapse: Reasoning failures in llms via adversarial prompting in code generation. *arXiv preprint arXiv:2506.06971*, 2025.

Miguel Romero-Arjona, Pablo Valle, Juan C Alonso, Ana B Sánchez, Miriam Ugarte, Antonia Cazalilla, Vicente Cambrón, José A Parejo, Aitor Arrieta, and Sergio Segura. Red teaming contemporary ai models: Insights from spanish and basque perspectives. *arXiv preprint arXiv:2503.10192*, 2025.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proc. NAACL*, 2024.

Mahdi Sabbaghi, Paul Kassianik, George Pappas, Yaron Singer, Amin Karbasi, and Hamed Hassani. Adversarial reasoning at jailbreaking time. In *Proc. ICML*, 2025.

Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping, Micah Goldblum, Tom Goldstein, and David Jacobs. Autoregressive perturbations for data poisoning. In *Proc. NeurIPS*, 2022.

Yash Savani, Asher Trockman, Zhili Feng, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J Zico Kolter. Antidistillation sampling. *arXiv preprint arXiv:2504.13146*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, et al. Large language model safety: A holistic survey. *arXiv preprint arXiv:2412.17686*, 2024.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proc. S&P*, pp. 3–18, 2017.

Wai Man Si, Mingjie Li, Michael Backes, and Yang Zhang. Excessive reasoning attack on reasoning llms. *arXiv preprint arXiv:2506.14374*, 2025.

Bingrui Sima, Linhua Cong, Wenxuan Wang, and Kun He. Viscra: A visual chain reasoning attack for jailbreaking multimodal large language models. *arXiv preprint arXiv:2505.19684*, 2025.

Yash Sinha, Manit Baser, Murari Mandal, Dinil Mon Divakaran, and Mohan Kankanhalli. Step-by-step reasoning attack: Revealing 'erased' knowledge in large language models. *arXiv preprint arXiv:2506.17279*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Hongru Song, Yu-an Liu, Ruqing Zhang, Jiafeng Guo, and Yixing Fan. Chain-of-thought poisoning attacks against r1-based retrieval-augmented generation systems. *arXiv preprint arXiv:2505.16367*, 2025a.

Linxin Song, Taiwei Shi, and Jieyu Zhao. The hallucination tax of reinforcement finetuning. *arXiv preprint arXiv:2505.13988*, 2025b.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. In *Proc. NeurIPS D&B Track*, 2024.

Makesh Narsimhan Sreedhar, Traian Rebedea, and Christopher Parisien. Safety through reasoning: An empirical study of reasoning guardrail models. *arXiv preprint arXiv:2505.20087*, 2025.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *Proc. ICLR*, 2024.

Jingbo Su. Enhancing adversarial attacks through chain of thought. *arXiv preprint arXiv:2410.21791*, 2024.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*, 2025.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

Chung-En Sun, Ge Yan, and Tsui-Wei Weng. Thinkedit: Interpretable weight editing to mitigate overly short thinking in reasoning models. *arXiv preprint arXiv:2503.22048*, 2025a.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*, 2025b.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of Proc. ACL*, 2024.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. *arXiv preprint arXiv:2505.12886*, 2025c.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the difficulty of faithful chain-of-thought reasoning in large language models. In *ICML Workshop on TiFA*, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

NovaSky Team. Sky-t1: Train your own o1 preview model within $450. https://novasky-ai.github.io/posts/sky-t1, 2025a. Accessed: 2025-01-09.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025b. URL `https://qwenlm.github.io/blog/qwq-32b/`.

Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*, 2024.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. In *Proc. NeurIPS*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proc. NeurIPS*, 2023.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring faithfulness of chains of thought by unlearning reasoning steps. *arXiv preprint arXiv:2502.14829*, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Bibek Upadhayay, Vahid Behzadan, et al. X-guard: Multilingual guard agent for content moderation. *arXiv preprint arXiv:2504.08848*, 2025.

Bertie Vidgen, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A Hale, and Paul Röttger. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

Scott Viteri, Max Lamparth, Peter Chatain, and Clark Barrett. Markovian transformers for informative language modeling. *arXiv preprint arXiv:2404.18988*, 2024.

Shengye Wan, Cyrus Nikolaidis, Daniel Song, David Molnar, James Crnkovich, Jayson Grace, Manish Bhatt, Sahana Chennabasappa, Spencer Whitman, Stephanie Ding, et al. Cyberseceval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models. *arXiv preprint arXiv:2408.01605*, 2024.

Changsheng Wang, Chongyu Fan, Yihua Zhang, Jinghan Jia, Dennis Wei, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Reasoning model unlearning: Forgetting traces, not just answers, while preserving reasoning skills. *arXiv preprint arXiv:2506.12963*, 2025a.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. *arXiv preprint arXiv:2506.04832*, 2025b.

Haoran Wang and Kai Shu. Trojan activation attack: Red-teaming large language models using activation steering for safety-alignment. *arXiv preprint arXiv:2311.09433*, 2023.

Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Dacheng Tao, and Minhao Cheng. Safety reasoning with guidelines. In *Proc. ICML*, 2025c.

Jiawei Wang, Da Cao, Shaofei Lu, Zhanchang Ma, Junbin Xiao, and Tat-Seng Chua. Causal-driven large language models with faithful reasoning for knowledge question answering. In *Proc. MM*, pp. 4331–4340, 2024a.

Jun Wang, Meng Fang, Ziyu Wan, Muning Wen, Jiachen Zhu, Anjie Liu, Ziqin Gong, Yan Song, Lei Chen, Lionel M Ni, et al. Openr: An open source framework for advanced reasoning with large language models. *arXiv preprint arXiv:2410.09671*, 2024b.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*, 2025d.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. Adaptive thinking via mode policy optimization for social language agents. *arXiv preprint arXiv:2505.02156*, 2025e.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*, 2025f.

Si Wang and Chip-Hong Chang. Fingerprinting deep neural networks-a deepfool approach. In *Proc. ISCAS*, pp. 1–5, 2021.

Wenxiao Wang, Parsa Hosseini, and Soheil Feizi. Chain-of-defensive-thought: Structured reasoning elicits robustness in large language models against reference corruption. *arXiv preprint arXiv:2504.20769*, 2025g.

Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in nlp models: A survey. In *Proc. NAACL*, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proc. ICLR*, 2023.

Yanbo Wang, Jian Liang, and Ran He. Towards eliminating hard label constraints in gradient inversion attacks. In *Proc. ICLR*, 2024c.

Yanbo Wang, Jiyang Guan, Jian Liang, and Ran He. Do we really need curated malicious data for safety alignment in multi-modal large language models? *arXiv preprint arXiv:2504.10000*, 2025h.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025i.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025j.

Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding. *arXiv preprint arXiv:2503.01422*, 2025k.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*, 2025l.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*, 2025m.

Yuhao Wang, Wenjie Qu, Yanze Jiang, Zichen Liu, Yue Liu, Shengfang Zhai, Yinpeng Dong, and Jiaheng Zhang. Silent leaks: Implicit knowledge extraction attack on rag systems through benign queries. *arXiv preprint arXiv:2505.15420*, 2025n.

Yuqing Wang and Yun Zhao. Rupbench: Benchmarking reasoning under perturbations for robustness evaluation in large language models. *arXiv preprint arXiv:2406.11020*, 2024.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024d.

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie. Star-1: Safer alignment of reasoning llms with 1k data. *arXiv preprint arXiv:2504.01903*, 2025o.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Proc. NeurIPS*, 2024e.

Peng Wanli, Xue Yiming, et al. Imf: Implicit fingerprint for large language models. *arXiv preprint arXiv:2503.21805*, 2025.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In *Proc. NeurIPS*, 2023a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. NeurIPS*, 2022.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023b.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025a.

Tong Wu, Chong Xiang, Jiachen T Wang, and Prateek Mittal. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*, 2025b.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-polish: Enhance reasoning in large language models via problem refinement. In *Findings of Proc. EMNLP*, 2023.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *Proc. ICLR*, 2024a.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, et al. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning. *arXiv preprint arXiv:2406.09187*, 2024b.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. In *Proc. ICLR*, 2025.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12): 1486–1496, 2023.

Zikai Xie. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting for large-language-models. *arXiv preprint arXiv:2408.05093*, 2024.

Chen Xiong, Xiangyu Qi, Pin-Yu Chen, and Tsung-Yi Ho. Defensive prompt patch: A robust and generalizable defense of large language models against jailbreak attacks. In *Findings of Proc. ACL*, 2025a.

Zidi Xiong, Chen Shan, Zhenting Qi, and Himabindu Lakkaraju. Measuring the faithfulness of thinking drafts in large reasoning models. *arXiv preprint arXiv:2505.13774*, 2025b.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025a.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024a.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*, 2025b.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proc. NAACL*, 2024b.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proc. ACL*, 2024c.

Rongwu Xu, Zehan Qi, and Wei Xu. Preemptive answer “attacks" on chain-of-thought reasoning. In *Findings of Proc. ACL*, 2024d.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025c.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*, 2024e.

Kureha Yamaguchi, Benjamin Etheridge, and Andy Arditi. Adversarial manipulation of reasoning models using internal representations. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *Proc. NAACL*, 2024.

Kai Yan, Yufei Xu, Zhengyin Du, Xuesong Yao, Zheyu Wang, Xiaowen Guo, and Jiecao Chen. Recitation over reasoning: How cutting-edge language models can fail on elementary school-level reasoning problems? *arXiv preprint arXiv:2504.00509*, 2025.

Haoyan Yang, Runxue Bao, Cao Xiao, Jun Ma, Parminder Bhatia, Shangqian Gao, and Taha Kass-Hout. Any large language model can be a reliable judge: Debiasing with a reasoning-based bias detector. *arXiv preprint arXiv:2505.17100*, 2025a.

Junjie Yang, Ke Lin, and Xing Yu. Think when you need: Self-adaptive chain-of-thought learning. *arXiv preprint arXiv:2504.03234*, 2025b.

Xianglin Yang, Gelei Deng, Jieming Shi, Tianwei Zhang, and Jin Song Dong. Enhancing model defense against jailbreaks with proactive safety reasoning. *arXiv preprint arXiv:2501.19180*, 2025c.

Yahan Yang, Soham Dan, Shuo Li, Dan Roth, and Insup Lee. Mr. guard: Multilingual reasoning guardrail using curriculum learning. *arXiv preprint arXiv:2504.15241*, 2025d.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025e.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024a.

Yang Yao, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, and Yingchun Wang. A mousetrap: Fooling large reasoning models for jailbreak with chain of iterative chaos. *arXiv preprint arXiv:2502.15806*, 2025a.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Proc. NeurIPS*, 2024b.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025b.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. Dissociation of faithful and unfaithful reasoning in llms. *arXiv preprint arXiv:2405.15092*, 2024.

Zonghao Ying, Deyue Zhang, Zonglei Jing, Yisong Xiao, Quanchen Zou, Aishan Liu, Siyuan Liang, Xiangzheng Zhang, Xianglong Liu, and Dacheng Tao. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models. *arXiv preprint arXiv:2502.11054*, 2025a.

Zonghao Ying, Guangyi Zheng, Yongxin Huang, Deyue Zhang, Wenxin Zhang, Quanchen Zou, Aishan Liu, Xianglong Liu, and Dacheng Tao. Towards understanding the safety boundaries of deepseek models: Evaluation and findings. *arXiv preprint arXiv:2503.15092*, 2025b.

Sangyeon Yoon, Wonje Jeung, and Albert No. R-tofu: Unlearning in large reasoning models. *arXiv preprint arXiv:2505.15214*, 2025.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. In *NeurIPS Workshop on Sys-2 Reasoning*, 2024a.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proc. CVPR*, pp. 13807–13816, 2024b.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. Benchmarking reasoning robustness in large language models. *arXiv preprint arXiv:2503.04550*, 2025a.

Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Sinong Wang, et al. Think smarter not harder: Adaptive reasoning with inference aware optimization. *arXiv preprint arXiv:2501.17974*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proc. CVPR*, pp. 9556–9567, 2024.

Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, et al. Trading inference-time compute for adversarial robustness. *arXiv preprint arXiv:2501.18841*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proc. NeurIPS*, volume 1126, 2024.

Xinyi Zeng, Yuying Shang, Jiawei Chen, Jingyuan Zhang, and Yu Tian. Root defence strategies: Ensuring safety of llm at the decoding level. *arXiv preprint arXiv:2410.06809*, 2024a.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *arXiv preprint arXiv:2407.17436*, 2024b.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025a.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024a.

Qingjie Zhang, Han Qiu, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, and Minlie Huang. Understanding the dark side of llms' intrinsic self-correction. *arXiv preprint arXiv:2412.14959*, 2024b.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Proc. COLM*, 2024c.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024d.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Meijuan An, Bikun Yang, KaiKai Zhao, Kai Wang, and Shiguo Lian. Chisafetybench: A chinese hierarchical safety benchmark for large language models. *arXiv preprint arXiv:2406.10311*, 2024e.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Limin Han, Jiaojiao Zhao, Beibei Huang, Zhenhong Long, Junting Guo, Meijuan An, Rongjia Du, et al. Safety evaluation and enhancement of deepseek models in chinese contexts. *arXiv preprint arXiv:2503.16529*, 2025b.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. In *Proc. ICML*, 2025c.

Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang, Zhijie Deng, and Yinpeng Dong. Realsafe-r1: Safety-aligned deepseek-r1 without compromising reasoning capability. *arXiv preprint arXiv:2504.10081*, 2025d.

Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. Stair: Improving safety alignment with introspective reasoning. *arXiv preprint arXiv:2502.02384*, 2025e.

Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel M Bikel, Jason Weston, and Eric Michael Smith. Backtracking improves generation safety. In *Proc. ICLR*, 2025f.

Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, and Ding Zhao. Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable llm safety. *arXiv preprint arXiv:2503.05021*, 2025g.

Zhexin Zhang, Junxiao Yang, Pei Ke, Fei Mi, Hongning Wang, and Minlie Huang. Defending large language models against jailbreaking attacks through goal prioritization. In *Proc. ACL*, 2024f.

Zhexin Zhang, Xian Qi Loye, Victor Shea-Jay Huang, Junxiao Yang, Qi Zhu, Shiyao Cui, Fei Mi, Lifeng Shang, Yingkang Wang, Hongning Wang, et al. How should we enhance the safety of large reasoning models: An empirical study. *arXiv preprint arXiv:2505.15404*, 2025h.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024g.

Gejian Zhao, Hanzhou Wu, Xinpeng Zhang, and Athanasios V Vasilakos. Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms. *arXiv preprint arXiv:2504.05605*, 2025a.

Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proc. EMNLP*, 2023.

Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, Tat-Seng Chua, and Ting Liu. Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities. *arXiv preprint arXiv:2503.17979*, 2025b.

Baihui Zheng, Boren Zheng, Kerui Cao, Yingshui Tan, Zhendong Liu, Weixun Wang, Jiaheng Liu, Jian Yang, Wenbo Su, Xiaoyong Zhu, et al. Beyond safe answers: A benchmark for evaluating true risk awareness in large reasoning models. *arXiv preprint arXiv:2505.19690*, 2025a.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*, 2024.

Jingnan Zheng, Xiangtian Ji, Yijun Lu, Chenhang Cui, Weixiang Zhao, Gelei Deng, Zhenkai Liang, An Zhang, and Tat-Seng Chua. Rsafe: Incentivizing proactive reasoning to build robust and adaptive llm safeguards. *arXiv preprint arXiv:2506.07736*, 2025b.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*, 2025a.

Kaiwen Zhou, Xuandong Zhao, Gaowen Liu, Jayanth Srinivasa, Aosong Feng, Dawn Song, and Xin Eric Wang. Safekey: Amplifying aha-moment insights for safety reasoning. *arXiv preprint arXiv:2505.16186*, 2025b.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? In *Proc. NeurIPS*, 2024a.

Zhanke Zhou, Jianing Zhu, Fengfei Yu, Xuan Li, Xiong Peng, Tongliang Liu, and Bo Han. Model inversion attacks: A survey of approaches and countermeasures. *arXiv preprint arXiv:2411.10023*, 2024b.

Bin Zhu, Hailong Yin, Jingjing Chen, and Yu-Gang Jiang. Reasoning models are more easily gaslighted than you think. *arXiv preprint arXiv:2506.09677*, 2025a.

Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei Yin, and Lei Sha. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking. *arXiv preprint arXiv:2502.12970*, 2025b.

Zihao Zhu, Hongbao Zhang, Ruotong Wang, Ke Xu, Siwei Lyu, and Baoyuan Wu. To think or not to think: Exploring the unthinking vulnerability in large reasoning models. *arXiv preprint arXiv:2502.12202*, 2025c.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *Proc. ICML*, 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.