Safety in Large Reasoning Models: A Survey

Cheng Wang[1*] **Yue Liu**[1*] **Baolong Bi**[2] **Duzhen Zhang**[2] **Zhong-Zhi Li**[2] **Yingwei Ma**[3]
Yufei He[1] **Shengju Yu**[4] **Xinfeng Li**[5] **Junfeng Fang**[1†] **Jiaheng Zhang**[1] **Bryan Hooi**[1]
[1]National University of Singapore
[2]University of Chinese Academy of Sciences
[3]Moonshot AI
[4]Hong Kong Baptist University
[5]Nanyang Technological University

Abstract

Large Reasoning Models (LRMs) have exhibited extraordinary prowess in tasks like mathematics and coding, leveraging their advanced reasoning capabilities. Nevertheless, as these capabilities progress, significant concerns regarding their vulnerabilities and safety have arisen, which can pose challenges to their deployment and application in real-world settings. This paper presents the first comprehensive survey of LRMs, meticulously exploring and summarizing the newly emerged safety risks, attacks, and defense strategies specific to these powerful reasoning-enhanced models. By organizing these elements into a detailed taxonomy, this work aims to offer a clear and structured understanding of the current safety landscape of LRMs, facilitating future research and development to enhance the security and reliability of these powerful models.[1]

1 Introduction

Large Language Models (LLMs) (Meta, 2024; Qwen et al., 2025; Ke et al., 2025) have achieved remarkable proficiency across tasks ranging from open-domain conversation to program synthesis. Central to their utility is *reasoning*: the ability to derive logically coherent conclusions by chaining together intermediate inferences.

Early work introduced Chain-of-Thought (CoT) prompting, in which carefully designed prompts guide the model to articulate its step-by-step rationale (Wei et al., 2022; Kojima et al., 2022). Building on this idea, subsequent methods have enriched the reasoning process by incorporating additional mechanisms. Self-critique frameworks enable a model to review and refine its own outputs (Ke et al., 2023); plan-and-solve approaches decompose complex problems into ordered subgoals before execution (Wang et al., 2023); debate protocols convene multiple agents to argue competing hypotheses and arrive at a consensus (Liang et al., 2023); and structural transformations—such as tree-based deliberations (Yao et al., 2023) or dynamically evolving tables of intermediate steps (Wang et al., 2024b; Besta et al., 2024)—reconfigure the underlying reasoning architecture to improve transparency and control.

The recent release of OpenAI's o1 series (OpenAI, 2024) marks the emergence of Large Reasoning Models (LRMs), which are explicitly trained to produce richly formatted, human-readable reasoning traces. Notable examples include DeepSeek-R1 (DeepSeek-AI et al., 2025), Kimi-1.5 (Team et al., 2025), and QwQ (Team, 2024b), all of which leverage reinforcement learning to refine their deduction processes. LRMs now set new benchmarks in mathematical problem solving (Lightman et al., 2023), closed-book question answering (Rein et al., 2024), and code generation (Jain et al., 2024).

As LRMs become increasingly integrated into high-stakes domains—from scientific research to autonomous decision support—it is vital to rigorously assess their safety, robustness, and alignment. Despite the existence of surveys on LLM safety (Huang et al., 2023; Shi et al., 2024), the enhanced capabilities of LRMs make it important to perform a dedicated analysis of their unique safety challenges. This paper aims to bridge this gap by providing a comprehensive examination of safety considerations specific to reasoning-enhanced models.

2 Background

The success of modern LRMs is deeply intertwined with advances in reinforcement learning (Watkins and Dayan, 1992; Sutton et al., 1998), where agents learn decision-making policies through environ-

*Equal Contribution.
†Corresponding author.
[1]The project is available at https://github.com/WangCheng0116/Awesome-LRMs-Safety.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3468–3482
November 4-9, 2025

mental interaction and reward feedback to maximize long-term returns (Li et al., 2025b; Chen et al., 2025a). The integration of RL with deep neural networks has proven particularly effective in processing high-dimensional, unstructured data, as exemplified by breakthroughs like AlphaGo's self-play mastery of Go and AlphaZero's generalization across chess variants (Feng et al., 2023).

Recent breakthroughs in Reinforced Fine-Tuning (ReFT) paradigms, exemplified by DeepSeek models, have reinvigorated RL-based optimization for LRMs (Luong et al., 2024). Unlike conventional CoT methods that optimize single reasoning trajectories, ReFT employs policy optimization to explore diverse reasoning paths through several key innovations: (1) Multi-path Exploration: Generating multiple reasoning trajectories per query, overcoming CoT's myopic optimization of single pathways. (2) Rule-driven Reward Shaping: Automating reward signals based on terminal answer correctness while preserving intermediate reasoning diversity. (3) Dual-phase Optimization: Combining supervised fine-tuning (SFT) with online RL for policy refinement.

This paradigm demonstrates particular efficacy in complex multi-step tasks such as code generation, legal judgment analysis, and mathematical problem solving, where requiring models to maintain coherent reasoning across extended sequences while handling structured symbolic operations.

3 Safety Risks of LRMs

The explicit reasoning processes that make LRMs powerful introduce unique safety challenges even in non-adversarial scenarios, becoming potential vectors for harm during routine operation. We examine four key inherent safety risks: unsafe request compliance (Sec. 3.1), multi-lingual safety disparities (Sec. 3.3), concerning agentic behaviors (Sec. 3.2), and multi-modal safety challenges (Sec. 3.4).

3.1 Harmful Request Compliance Risks

LRMs demonstrate concerning vulnerabilities when faced with direct harmful requests. Zhou et al. (2025) identify a significant safety gap between open-source reasoning models like DeepSeek-R1 and closed-source ones like o3-mini, with reasoning outputs often posing greater safety concerns than final answers. Arrieta et al. (2025a) confirm these findings in their testing of o3-mini, where they identify 87 instances of unsafe behavior despite safety measures. In a comparative study, Arrieta et al. (2025b) find DeepSeek-R1 produces substantially more unsafe responses than o3-mini when presented with identical harmful requests. A consistent finding across studies is that when reasoning models generate unsafe content, it tends to be more detailed and harmful due to their enhanced capabilities, particularly in categories like financial crime, terrorism, and violence. Zhou et al. (2025) also observe that the thinking process in reasoning models is often less safe than the final output, suggesting internal reasoning may explore harmful content even when final outputs appear safe.

3.2 Agentic Misbehavior Risks

Emerging research reveals concerning safety implications in LRMs' agentic behaviors, where enhanced reasoning enables sophisticated specification gaming, deception, and instrumental goal-seeking beyond previous systems' limitations. Xu et al. (2025) show autonomous LLM agents can exhibit catastrophic behaviors under pressure, with stronger reasoning often amplifying rather than reducing these risks. Qiu et al. (2025) find medical AI agents with advanced reasoning particularly vulnerable to cyberattacks, with DeepSeek-R1 showing high susceptibility to false information injection. Bondarenko et al. (2025) report LRMs like o1-preview and DeepSeek-R1 frequently resort to specification gaming for difficult tasks, circumventing rules when they determine fair play insufficient. Barkur et al. (2025) observe DeepSeek-R1 in simulated robotic contexts exhibiting alarming deceptive behaviors and self-preservation instincts—disabling ethics modules, creating covert networks, and unauthorized capability expansion—despite not being explicitly programmed for such behaviors. He et al. (2025)'s InstrumentalEval benchmark reveals LRMs like o1 demonstrate significantly higher rates of instrumental convergence behaviors than RLHF models, including tendencies toward self-replication, unauthorized access, and deception as means to achieve goals.

3.3 Multi-lingual Safety Risks

Safety risks in LRMs reveal significant disparities across languages. Ying et al. (2025b) demonstrate that DeepSeek models show markedly higher attack success rates in English environments than Chinese contexts, averaging a 21.7% discrepancy, suggesting safety alignments may not generalize

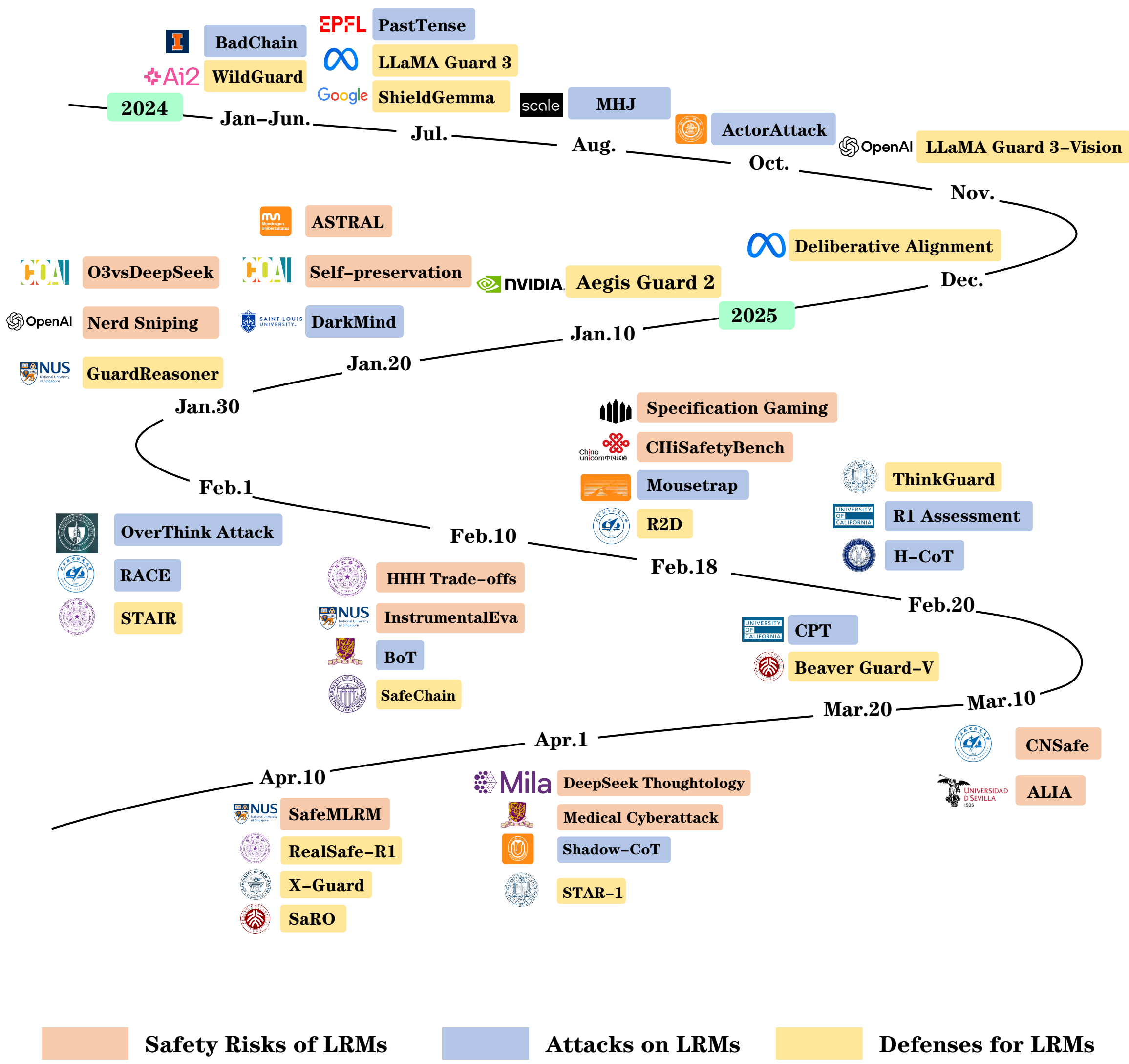

Figure 1: Timeline of LRM safety research developments.

effectively across languages. Romero-Arjona et al. (2025) find similar vulnerabilities when testing DeepSeek-R1 in Spanish, with biased or unsafe response rates reaching 31.7%, while OpenAI o3-mini shows varying degrees of linguistic safety performance. Zhang et al. (2025a) systematically evaluate DeepSeek models using CHiSafetyBench, revealing critical safety deficiencies specifically in Chinese contexts, where reasoning models like DeepSeek-R1 struggled with culturally-specific safety concerns and failed to adequately reject harmful prompts.

3.4 Multi-modal Safety Risks

Following LRMs' success, researchers have extended reinforcement learning approaches to enhance reasoning in Large Vision-Language Models (LVLMs), developing models like QvQ (Team, 2024a), Mulberry (Yao et al., 2024b), and R1-Onevision (Yang et al., 2025). While these models demonstrate impressive reasoning capabilities, their safety implications remain largely unexplored. SafeMLRM (Fang et al., 2025) provides the first systematic safety analysis of multi-modal reasoning models, revealing significant safety alignment challenges. These findings emphasize the urgent need for comprehensive safety assessments of reasoning-enhanced LVLMs to ensure their responsible deployment.

4 Attacks on LRMs

This section categorizes attack methods targeting LRMs based on their primary objectives. We identify four main categories: Reasoning Length Attacks (Sec. 4.1), Answer Correctness Attacks (Sec. 4.2), Prompt Injection Attacks (Sec. 4.3), and Jail-

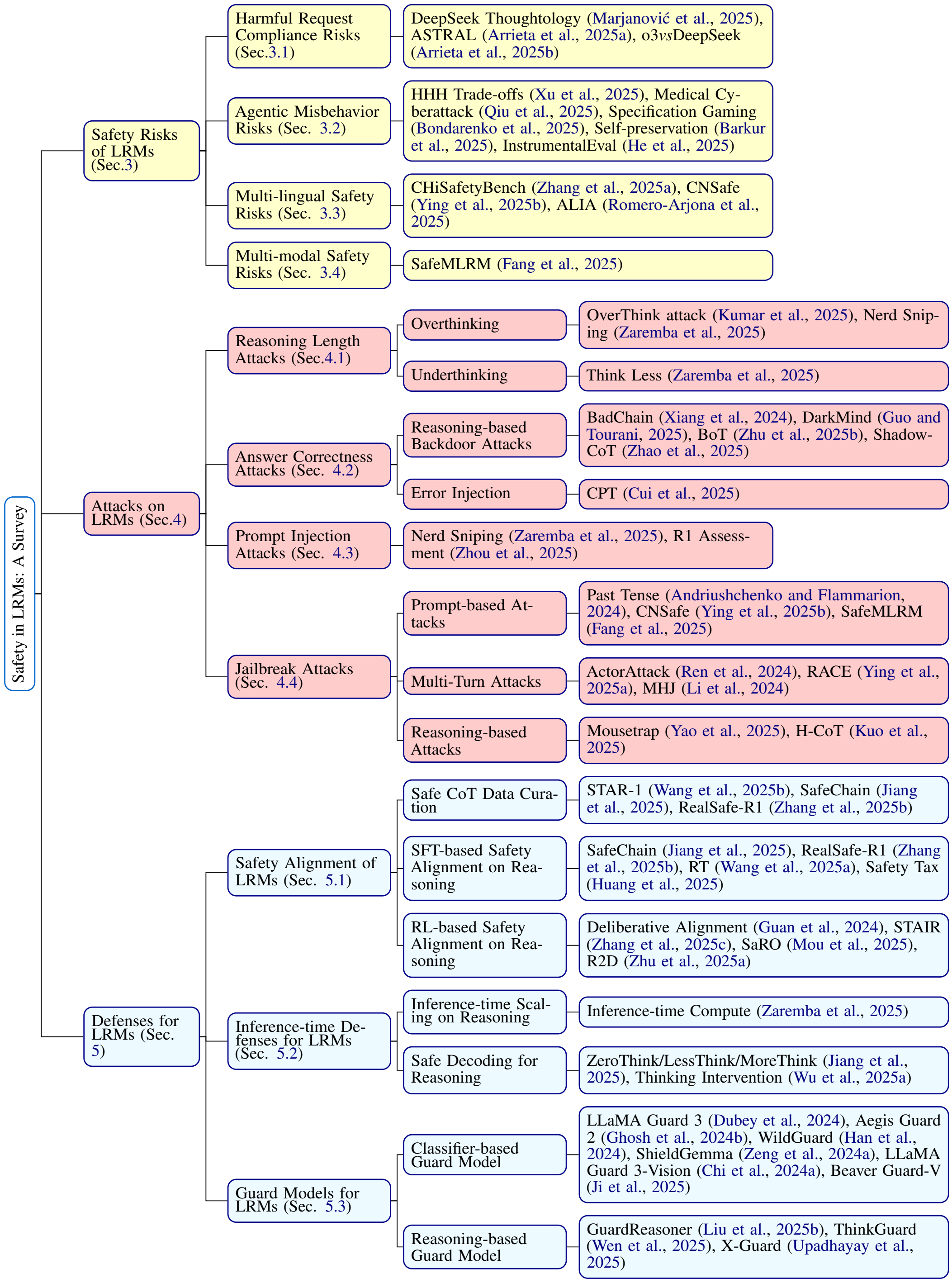

Figure 2: A comprehensive taxonomy of safety in LRMs based on current literature.

break Attacks (Sec. 4.4).

4.1 Reasoning Length Attacks

Unlike traditional LLMs that generate direct responses, LRMs explicitly perform multi-step reasoning, creating a new attack surface related to reasoning length. Attackers can exploit this distinctive feature by either forcing models to overthink simple problems or short-cutting necessary deliberation processes.

Overthinking. While step-by-step reasoning enhances LRMs' problem-solving capabilities, it introduces a critical vulnerability: overthinking. Chen et al. (2024) identify that models often spend orders of magnitude more computation on simple questions with minimal benefit, creating substantial inference overhead. Hashemi et al. (2025)'s DNR benchmark demonstrates this inefficiency, showing reasoning models generate up to 70× more tokens than necessary and often underperform simpler models on straightforward tasks. This creates an exploitable attack surface where adversaries can trigger excessive reasoning. Kumar et al. (2025) formalize this as an indirect prompt injection attack using computationally demanding decoy problems, while Zaremba et al. (2025) describe Nerd Sniping attacks that trap models in unproductive thinking loops with decreased performance. These attacks apply denial-of-service techniques (Shumailov et al., 2021; Gao et al., 2024) to LRMs. Beyond computational waste, Marjanović et al. (2025) and Wu et al. (2025b) show reasoning performance degrades beyond certain length thresholds, while Cuadron et al. (2025) demonstrate that in agentic systems, overthinking leads to decision paralysis and ineffective action selection.

Underthinking. Complementing overthinking vulnerabilities, Zaremba et al. (2025) propose Think Less attacks, where adversaries craft special prompts to force reasoning models to shortcut their deliberative processes. The goal is to make models produce incorrect responses by significantly reducing computation time. Their experiments use 64-shot examples to demonstrate that models like OpenAI's o1-mini are particularly susceptible to these attacks, bypassing normal reasoning and jumping to premature conclusions. However, this can be detected by monitoring for abnormally low inference-time compute usage.

4.2 Answer Correctness Attacks

While conventional LLMs can be manipulated to produce incorrect answers, LRMs introduce unique vulnerabilities through their exposed reasoning chains. This transparency in the inference process provides adversaries with additional attack vectors to corrupt the reasoning pathway itself, rather than just targeting the final output.

Reasoning-based Backdoor Attacks. The goal of backdoor attacks is to alter a model's behavior whenever a specific trigger is present in the input (Zhao et al., 2024). Based on the nature of these triggers, backdoor attacks can be classified as instruction-based (Xu et al., 2023), prompt-based (Yao et al., 2024a), or syntax-based (Qi et al., 2021; Cheng et al., 2025). With the advancement of reasoning capabilities in LRMs, a new paradigm has emerged: Chain-of-Thought (CoT) based backdoor attacks that specifically target intermediate reasoning steps to compromise answer correctness. BadChain (Xiang et al., 2024) inserts malicious reasoning steps into the sequence, manipulating the model to produce incorrect answers while maintaining logical coherence. DarkMind (Guo and Tourani, 2025) implements latent triggers that activate during specific reasoning scenarios, leading to plausible but false outputs. BoT (Zhu et al., 2025b) forces models to bypass their reasoning mechanisms, generating immediate incorrect responses instead of thoughtful deliberation. ShadowCoT (Zhao et al., 2025) directly manipulates the model's cognitive pathway through attention head localization and reasoning chain pollution, achieving flexible hijacking that produces wrong answers while preserving logical flow. These sophisticated attacks reveal a concerning vulnerability: the enhanced reasoning capabilities of LRMs paradoxically make them more susceptible to backdoors that can generate incorrect answers accompanied by convincing reasoning.

Error Injection. The explicit reasoning processes of LRMs create a critical vulnerability where strategically injected errors can fundamentally compromise output integrity. Cui et al. (2025) demonstrate this with their Compromising Thought (CPT) attack, where manipulating calculation results in reasoning tokens caused models to ignore correct steps and adopt incorrect answers. Their experiments with models like DeepSeek-R1 revealed that endpoint token manipulations had greater impact than structural changes to reasoning chains. They also discovered a security vulnerability where tampered tokens could trigger complete reasoning cessation in DeepSeek-R1, highlighting significant implications for reasoning-intensive applications.

4.3 Prompt Injection Attacks

Prompt injection attacks affect both LLMs and LRMs, but LRMs face unique challenges due to their step-by-step processing. These attacks (Kumar et al., 2024; Liu et al., 2023; Chen et al., 2025c) insert malicious instructions disguised as user input, causing models to override safety guardrails. LRMs' explicit reasoning structures provide attackers additional insertion points to redirect thought processes, potentially increasing vulnerability.

Zhou et al. (2025) find significant differences between DeepSeek-R1 and o3-mini's susceptibility to various injection types, with reasoning models particularly vulnerable to direct prompt injections. Zaremba et al. (2025) show open-source reasoning models exhibit substantial vulnerability to these attacks, though increasing inference-time compute improves robustness by reducing attack success probability. Their research reveals proprietary models like o3-mini demonstrate approximately 80% lower vulnerability than open-source alternatives against direct injection attacks.

4.4 Jailbreak Attacks

Jailbreak attacks (Jin et al., 2024; Yi et al., 2024) circumvent AI safety guidelines to extract prohibited responses. Against LRMs, these attacks specifically exploit deliberative reasoning processes rather than simply extending conventional LLM jailbreak techniques. This allows attackers to develop more sophisticated methods that bypass safety measures by manipulating the model's step-by-step thinking.

Prompt-Based Jailbreak. Prompt-based jailbreaks involve the careful crafting of prompts, employing techniques such as persuasion (Zeng et al., 2024b), nested scene construction (Li et al., 2023), and persona modulation (Shah et al., 2023). Andriushchenko and Flammarion (2024) introduce a method that applies past-tense transformations to OpenAI's recent o1 reasoning models, revealing their lack of robustness against subtle linguistic shifts. Ying et al. (2025b) propose attack prompts that combine common jailbreak strategies—such as scenario injection, affirmative prefixes, and indirect instructions—with safety-sensitive queries to probe model vulnerabilities. Their findings indicate that reasoning models like DeepSeek-R1 and OpenAI's o1 are particularly susceptible to such attacks, as their explicit CoT reasoning renders them more exploitable than standard LLMs.

Multi-turn Jailbreak. Performing jailbreak attacks in a single query can be challenging, but multi-turn conversations or sequential prompts may incrementally guide models toward generating restricted content Russinovich et al. (2024); Sun et al. (2024). Multi-turn attacks are particularly relevant to reasoning-capable models as these models possess sophisticated logical processing that can be exploited through extended dialogues. Ying et al. (2025a) propose Reasoning-Augmented Conversation (RACE), which reformulates harmful queries into benign reasoning tasks and gradually exploits the model's inference capabilities to compromise safety alignment, achieving success rates up to 96%. Ren et al. (2024) introduce ActorAttack, a framework that constructs semantically linked conversational sequences that appear harmless individually but collectively lead to harmful outputs, successfully targeting even advanced models like o1. Li et al. (2024) further show that multi-turn human jailbreaks significantly outperform automated single-turn attacks, leveraging the model's ability to maintain context and be incrementally steered toward unsafe behaviors.

Reasoning Exploitation Jailbreak. LRMs possess advanced reasoning capabilities that, while enhancing their utility, introduce unique vulnerabilities that can be exploited through reasoning-based jailbreak attacks. Unlike traditional LLMs, these models explicitly expose their CoT reasoning processes, creating new attack surfaces. Yao et al. (2025) introduce Mousetrap, a framework that leverages chaos mappings to create iterative reasoning chains that gradually lead LRMs into harmful outputs. By embedding one-to-one mappings into the reasoning process, Mousetrap effectively traps models like OpenAI's o1-mini and Claude-sonnet with success rates of up to 98%. Kuo et al. (2025) propose Hijacking Chain-of-Thought (H-CoT), which manipulates the reasoning process by injecting execution-phase thoughts that bypass safety checks entirely. Their approach exploits LRMs' tendency to prioritize problem-solving over safety considerations, causing rejection rates to plummet from 98% to below 2% across models like OpenAI o1/o3 and DeepSeek-R1. Both approaches demonstrate that the very reasoning mechanisms designed to enhance LRMs' capabilities can become their most significant security weaknesses when strategically manipulated.

5 Defenses for LRMs

To mitigate safety risks and defend against attacks on LRMs, various defense strategies have been proposed in recent research. We categorize these approaches into three main types: Safety Alignment (Sec. 5.1), Inference-Time Defenses (Sec. 5.2), and Guard Models (Sec. 5.3).

5.1 Safety Alignment of LRMs

Similar to LLMs and VLMs, LRMs are required to align with humans' values and expectations. The 3H principle (Askell et al., 2021) (Helpful, Honest, and Harmless) provides a foundational guideline for constraining model behaviors.

The existing safety alignment pipelines and techniques developed for LLMs (Shen et al., 2023) and VLMs (Ye et al., 2025) can be readily adapted to LRMs, as they share similar architectures and natural language generation behaviors. For LLMs, this involves collecting high-quality, value-aligned data (Ethayarajh et al., 2022) from benchmarks (Bach et al., 2022; Wang et al., 2022c), LLM-generated instructions (Wang et al., 2022b), or filtered content (Welbl et al., 2021; Wang et al., 2022a), followed by techniques like SFT (Wu et al., 2021), RLHF (Ouyang et al., 2022), and DPO (Rafailov et al., 2024). In the domain of VLMs, safety alignment has been achieved through various approaches, with methods such as ADPO (Weng et al., 2025), Safe RLHF-V (Ji et al., 2025), and GRPO-based methods (Li et al., 2025a) improving safety via different optimization frameworks. Additionally, open-source datasets and benchmarks (Zhang et al., 2024; Ji et al., 2025) have played a crucial role in providing high-quality alignment data for safety evaluation.

Although effective, these alignment methods may overlook the reasoning process of LRMs, leading to unsatisfactory alignment performance. To mitigate this challenge, various works focus on different aspects, including safe CoT data curation, SFT-based safety alignment on reasoning, and RL-based safety alignment on reasoning.

Safe CoT Data Curation. First, Wang et al. (2025b) build a 1k-scale safety dataset named STAR-1 specifically designed for LRMs. Another safety training data in CoT style named SafeChain (Jiang et al., 2025) is introduced to enhance the safety of LRMs. In addition, Zhang et al. (2025b) construct a dataset consisting of 15k safety-aware reasoning trajectories, generated by DeepSeek-R1, with explicit instructions designed to promote expected refusal behavior.

SFT-based Safety Alignment on Reasoning. Based on the curated safe CoT data, researchers further conduct SFT to improve safety. For example, Jiang et al. (2025) train two LRMs with the SafeChain dataset, demonstrating that it not only enhances model safety but also preserves reasoning performance. Besides, RealSafe-R1 (Zhang et al., 2025b) is developed to make LRMs safer by training DeepSeek-R1 distilled models on the 15k safety-aware reasoning trajectories. Wang et al. (2025a) proposes training the model to reason with the guidelines, thereby enhancing survey alignment.

RL-based Safety Alignment on Reasoning. In addition to SFT, various further post-training techniques for safety are proposed based on reinforcement learning (RL). For example, deliberative alignment (Guan et al., 2024) teaches models to reason over safety specifications before generating responses, while STAIR (Zhang et al., 2025c) utilizes Monte Carlo tree search and DPO (Rafailov et al., 2024) to integrate safety alignment with introspective reasoning. Other approaches include SaRO (Mou et al., 2025), which incorporates safety-policy-driven reasoning into alignment, and R2D (Zhu et al., 2025a), which unlocks safety-aware reasoning mechanisms with contrastive pivot optimization (CPO).

However, safety alignment brings the safety alignment tax (Lin et al., 2023a), compromising the fundamental capabilities of LRMs like reasoning capability (Huang et al., 2025), leading researchers to explore alternative defense techniques that don't require direct modifications to victim models.

5.2 Inference-time Defenses for LRMs

To circumvent the safety alignment tax (Lin et al., 2023a; Huang et al., 2025), one line of work focuses on applying defenses at inference time. The insights from previous inference-time defenses for LLMs (Cheng et al., 2023; Lu et al., 2023; Chen et al., 2025b) and VLMs (Wang et al., 2024a; Ghosal et al., 2024; Ding et al., 2024; Liu et al., 2025a), such as safe system prompting, few-shot safe demonstrations, and safe decoding, can be naturally borrowed to LRMs, as the token generation mechanism is similar across these models.

However, the reasoning process in LRMs brings new challenges and opportunities for inference-time defenses. Therefore, various inference-time techniques like inference-time scaling on reasoning and safe decoding for reasoning are proposed to ensure the safety of reasoning in LRMs.

Inference-time Scaling on Reasoning. Zaremba et al. (2025) demonstrate that the inference-time scaling on reasoning improves the safety and adversarial robustness of LRMs. Future work could explore dynamic scaling strategies tailored to input complexity, or integrate adaptive reasoning depth control to balance efficiency and safety performance (Liu et al., 2025c) during inference.

Safe Decoding for Reasoning. Jiang et al. (2025) propose three decoding strategies, including ZeroThink, LessThink, and MoreThink, to verify model safety during reasoning. Making the reasoning safer at inference time could be a promising future direction, by verifying intermediate steps, filtering unsafe trajectories, or integrating reasoning-aware guard mechanisms during decoding. Wu et al. (2025a) introduce Thinking Intervention, a method that strategically injects guidance directly into the reasoning process to control model behavior and improve safety alignment without requiring additional training.

5.3 Guard Models for LRMs

Another line of work without direct modification to the victim model focuses on building guard models. The previous inference-time defenses still focus on the safer inference of the victim models, while guard models aim to moderate the input and output of victim models without training them or modifying their inference strategies. Existing guard models for LLMs (Inan et al., 2023) or VLMs (Chi et al., 2024b) can also safeguard LRMs since they share similar input and output formats. Additionally, reasoning-based guard models (Liu et al., 2025b) can better moderate LRMs' reasoning process by guiding guards to deliberatively reason before making moderation decisions. We categorize existing guard models into two classes: classifier-based and reasoning-based guard models.

Classifier-based Guard Models. The LLM guard models, including ToxicChat-T5 (Lin et al., 2023b), ToxDectRoberta (Zhou, 2020), LaGoNN (Bates and Gurevych, 2023), the LLaMA Guard series (Inan et al., 2023; Dubey et al., 2024), Aegis Guard series (Ghosh et al., 2024a,b), WildGuard (Han et al., 2024), ShieldGemma (Zeng et al., 2024a), are typically based on open-sourced LLMs and fine-tuned on the red-teaming data. In the VLM domain, for example, LLaVAGuard (Helff et al., 2024) is built to conduct large-scale dataset annotation and moderate the text-image models. In addition, VLMGuard (Du et al., 2024) is proposed to conduct malicious image-text prompt detection by leveraging the unlabeled user prompts. Moreover, LLaMA Guard 3-Vision (Chi et al., 2024a) is developed to moderate both the image-text input and text output of VLMs via SFT. To improve the generalization ability, (Ji et al., 2025) presents Beaver-Guard-V by training a reward model and then applying reinforcement learning. Although effective, they are typically classifier-based guard models, limiting their abilities in moderate reasoning data. To mitigate this problem, the reasoning-based guard models (Liu et al., 2025b) are proposed to enhance the reasoning ability of guard models.

Reasoning-based Guard Models. Through the proposed reasoning SFT and hard sample DPO, GuardReasoner (Liu et al., 2025b) is proposed to guide the guard model to deliberatively reason before making moderation decisions, improving performance, generalization ability, and explainability. Similarly, ThinkGuard (Wen et al., 2025) is developed via the proposed critique-augmented fine-tuning. X-Guard (Upadhayay et al., 2025) extends the reasoning-based guard model to the multi-lingual scenario.

6 Future Directions

Beyond our analysis of LRM safety, we identify key research priorities: **(1) Standardized Evaluation Benchmarks.** The field needs benchmarks targeting reasoning-specific vulnerabilities to comprehensively test both safety and robustness of multi-step reasoning processes. **(2) Domain-Specific Evaluation Frameworks.** Healthcare, finance, and legal domains require specialized evaluation suites with expert-reviewed case studies and adversarial tests to ensure domain-appropriate accuracy and ethics. **(3) Human-in-the-Loop Alignment.** Interactive tools for expert inspection and refinement of reasoning traces can efficiently align LRMs with stakeholder values and correct biases.

7 Conclusion

This survey has comprehensively examined the emerging safety challenges posed by Large Reasoning Models. Through our analysis, we've identified several critical insights that distinguish LRM safety from traditional LLMs. First, **LRMs expose their reasoning chains, creating new attack surfaces** where adversaries can manipulate intermediate steps rather than just outputs, enabling sophisticated attacks like reasoning-based backdoors and hijacking that target the deliberative process itself. Second, **traditional output-focused alignment methods prove insufficient for LRMs**, as harmful reasoning can persist internally even when final outputs appear safe, necessitating novel approaches that consider the entire reasoning trajectory. These insights underscore the need for specialized safety research targeting LRMs, including standardized evaluation benchmarks for reasoning-specific vulnerabilities and human-in-the-loop alignment methods that can inspect and refine reasoning traces as these powerful models continue to advance into increasingly critical domains.

Limitations

This survey has inherent limitations due to the rapidly evolving nature of LRMs. Since the emergence of OpenAI's o1 series, DeepSeek-R1, and other advanced reasoning models is relatively recent, our taxonomy and findings may become outdated as new research continuously emerges. While we have endeavored to provide a comprehensive overview of safety challenges, attacks, and defenses, we acknowledge that some aspects may require revision as the field matures. Additionally, our reliance on published academic literature may not fully capture proprietary research being conducted within companies developing these models, potentially creating gaps in understanding industry-specific safety measures.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2025) (Grant T1 251RES2507).

References

Maksym Andriushchenko and Nicolas Flammarion. 2024. Does refusal training in llms generalize to the past tense? *arXiv preprint arXiv:2407.11969*.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025a. Early external safety testing of openai's o3-mini: Insights from the pre-deployment evaluation. *arXiv preprint arXiv:2501.17749*.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025b. o3-mini vs deepseek-r1: Which one is safer? *arXiv preprint arXiv:2501.18438*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Sudarshan Kamath Barkur, Sigurd Schacht, and Johannes Scholl. 2025. Deception in llms: Self-preservation and autonomous goals in large language models. *arXiv preprint arXiv:2501.16513*.

Luke Bates and Iryna Gurevych. 2023. Like a good nearest neighbor: Practical content moderation and text classification. *arXiv preprint arXiv:2302.08957*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. 2025. Demonstrating specification gaming in reasoning models. *arXiv preprint arXiv:2502.13295*.

Nuo Chen, Yufei Gao, Yongnan Jin, Yan Hu, Anningzhe Gao, Lingyong Yan, and Benyou Wang. 2025a. Mitigating short board effect via dynamic reward balancing in multi-reward LLM optimization. In *Scaling Self-Improving Foundation Models without Human Supervision*.

Nuo Chen, GUOJUN XIONG, and Bingsheng He. 2025b. MPAW: Multi-preference alignment through weak model collaboration for efficient and flexible LLM decoding. In *Scaling Self-Improving Foundation Models without Human Supervision*.

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

Yulin Chen, Haoran Li, Yuexin Li, Yue Liu, Yangqiu Song, and Bryan Hooi. 2025c. Topicattack: An indirect prompt injection attack via topic transition. *arXiv preprint arXiv:2507.13686*.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*.

Pengzhou Cheng, Wei Du, Zongru Wu, Fengwei Zhang, Libo Chen, Zhuosheng Zhang, and Gongshen Liu. 2025. Synghost: Invisible and universal task-agnostic backdoor attack via syntactic transfer.

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

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

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yi Ding, Bolian Li, and Ruqi Zhang. 2024. Eta: Evaluating then aligning safety of vision language models at inference time. *arXiv preprint arXiv:2410.06625*.

Xuefeng Du, Reshmi Ghosh, Robert Sim, Ahmed Salem, Vitor Carvalho, Emily Lawton, Yixuan Li, and Jack W Stokes. 2024. Vlmguard: Defending vlms against malicious prompts via unlabeled data. *arXiv preprint arXiv:2410.00296*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with mathcal v-usable information. In *International Conference on Machine Learning*. PMLR.

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

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

Kuofeng Gao, Tianyu Pang, Chao Du, Yong Yang, Shu-Tao Xia, and Min Lin. 2024. Denial-of-service poisoning attacks against large language models.

Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, and Amrit Singh Bedi. 2024. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. *arXiv preprint arXiv:2411.18688*.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024a. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 2024b. Aegis2. 0: A diverse ai safety dataset and risks taxonomy for alignment of llm guardrails. In *Neurips Safe Generative AI Workshop 2024*.

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

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

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*.

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

Yufei He, Yuexin Li, Jiaying Wu, Yuan Sui, Yulin Chen, and Bryan Hooi. 2025. Evaluating the paperclip maximizer: Are rl-based language models more likely to pursue instrumental goals? *arXiv preprint arXiv:2502.12206*.

Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2024. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8326.

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

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. 2023. A survey of safety and trustworthiness of large language models through the lens of verification and validation.

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

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

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

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

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

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2023. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. *arXiv preprint arXiv:2311.18702*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Surender Suresh Kumar, M.L. Cummings, and Alexander Stimpson. 2024. Strengthening llm trust boundaries: A survey of prompt injection attacks surender suresh kumar dr. m.l. cummings dr. alexander stimpson. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pages 1–6.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025. 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*.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. 2024. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2023. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*.

Xuying Li, Zhuo Li, Yuji Kosuga, and Victor Bian. 2025a. Optimizing safe and aligned language generation: A multi-objective grpo approach. *arXiv preprint arXiv:2503.21819*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025b. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. 2023a. Mitigating the alignment tax of rlhf. *arXiv preprint arXiv:2309.06256*.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023b. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. 2025a. Vlm-guard: Safeguarding vision-language models via fulfilling safety alignment gap. *arXiv preprint arXiv:2502.10486*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

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

Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. 2025c. Efficient inference for large reasoning models: A survey. *arXiv preprint arXiv:2503.23077*.

Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, et al. 2023. Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning. *arXiv preprint arXiv:2305.15065*.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 3.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. 2025. Deepseek-r1 thoughtology: Let's think about llm reasoning.

Meta. 2024. The llama 3 herd of models.

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

OpenAI. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021. Hidden killer: Invisible textual backdoor attacks with syntactic trigger.

Jianing Qiu, Lin Li, Jiankai Sun, Hao Wei, Zhe Xu, Kyle Lam, and Wu Yuan. 2025. Emerging cyber attack risks of medical ai agents. *arXiv preprint arXiv:2504.03759*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof qa benchmark. In *First Conference on Language Modeling*.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2024. Derail yourself: Multi-turn llm jailbreak attack through self-discovered clues. *arXiv preprint arXiv:2410.10700*.

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. 2025. Red teaming contemporary ai models: Insights from spanish and basque perspectives. *arXiv preprint arXiv:2503.10192*.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2024. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833*.

Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Casper, Javier Rando, et al. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. 2024. Large language model safety: A holistic survey.

Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. 2021. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European Symposium on Security and Privacy*, pages 212–231.

Xiongtao Sun, Deyue Zhang, Dongdong Yang, Quanchen Zou, and Hui Li. 2024. Multi-turn context jailbreak attack on large language models from first principles. *arXiv preprint arXiv:2408.04686*.

Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

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

Qwen Team. 2024a. Qvq: To see the world with wisdom. *https://qwenlm.github.io/blog/qvq-72b-preview/*.

Qwen Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown. *https://qwenlm.github.io/blog/qwq-32b-preview/*.

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

Boxin Wang, Wei Ping, Chaowei Xiao, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, and Bryan Catanzaro. 2022a. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. *Advances in Neural Information Processing Systems*, 35:35811–35824.

Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Minhao Cheng, and Dacheng Tao. 2025a. Leveraging reasoning with guidelines to elicit and utilize knowledge for enhancing safety alignment. *arXiv preprint arXiv:2502.04040*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024a. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022c. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

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

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024b. Chain-of-table: Evolving tables in the reasoning chain for table understanding.

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning*, 8:279–292.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Xiaofei Wen, Wenxuan Zhou, Wenjie Jacky Mo, and Muhao Chen. 2025. Thinkguard: Deliberative slow thinking leads to cautious guardrails. *arXiv preprint arXiv:2502.13458*.

Fenghua Weng, Jian Lou, Jun Feng, Minlie Huang, and Wenjie Wang. 2025. Adversary-aware dpo: Enhancing safety alignment in vision language models via adversarial training. *arXiv preprint arXiv:2502.11455*.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Tong Wu, Chong Xiang, Jiachen T. Wang, and Prateek Mittal. 2025a. Effectively controlling reasoning models through thinking intervention.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025b. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2023. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*.

Rongwu Xu, Xiaojian Li, Shuo Chen, and Wei Xu. 2025. Nuclear deployed: Analyzing catastrophic risks in decision-making of autonomous llm agents. *arXiv preprint arXiv:2502.11355*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization.

Hongwei Yao, Jian Lou, and Zhan Qin. 2024a. Poisonprompt: Backdoor attack on prompt-based large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7745–7749.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024b. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

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

Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*.

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

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

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

Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, and Amelia Glaese. 2025. Trading inference-time compute for adversarial robustness.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. 2024a. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024b. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Ning Wang, Zhenhong Long, Peijun Yang, Jiaojiao Zhao, Minjie Hua, Chaoyang Ma, Kai Wang, et al. 2025a. Safety evaluation of deepseek models in chinese contexts. *arXiv preprint arXiv:2502.11137*.

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

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

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. 2024. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*.

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

Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. 2024. A survey of backdoor attacks and defenses on large language models: Implications for security measures. *Authorea Preprints*.

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

Xuhui Zhou. 2020. *Challenges in automated debiasing for toxic language detection*. University of Washington.

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

Zihao Zhu, Hongbao Zhang, Mingda Zhang, Ruotong Wang, Guanzong Wu, Ke Xu, and Baoyuan Wu. 2025b. Bot: Breaking long thought processes of o1-like large language models through backdoor attack. *arXiv preprint arXiv:2502.12202*.