大型推理模型中的安全性：综述

Cheng Wang[1*] **Yue Liu**[1*] **Baolong Bi**[2]
Duzhen Zhang[2] **Zhongzhi Li**[2] **Junfeng Fang**[1†]
[1]National University of Singapore
[2]University of Chinese Academy of Sciences
{ wangcheng, yliu } @u.nus.edu

Github: https://github.com/WangCheng0116/Awesome-LRMs-Safety

Abstract

大型推理模型（LRMs）在数学和编码等任务中表现出非凡的能力，充分利用了其先进的推理能力。然而，随着这些能力的进步，对其漏洞和安全性的重大担忧也浮出水面，这可能对其在现实环境中的部署和应用构成挑战。本文对 LRMs 进行了全面调查，详细地探索和总结了新出现的安全风险、攻击和防御策略。通过将这些元素组织成一个详细的分类，本工作旨在提供对 LRMs 当前安全态势的清晰且结构化的理解，以促进未来研究和发展，以增强这些强大模型的安全性和可靠性。

1 引言

大型语言模型 (LLMs) (Meta, 2024; Qwen et al., 2025) 在从开放域对话到程序合成的任务中取得了显著的熟练度。它们实用性的核心在于推理：通过连接中间推断来得出逻辑上一致的结论的能力。

早期的研究引入了链式思维 (CoT) 提示，其中精心设计的提示引导模型表达其逐步的推理过程 (Wei et al., 2022; Kojima et al., 2022) 。在此基础上，后续的方法通过结合额外的机制丰富了推理过程。自我批评框架使模型能够审查和完善自身输出 (Ke et al., 2023) ；计划与解决方法在执行前将复杂问题分解为有序的子目标 (Wang et al., 2023) ；辩论协议召集多个代理讨论竞争的假设并达成共识 (Liang et al., 2023) ；以及结构转化——如基于树的审议 (Yao et al., 2023) 或动态演化的中间步骤表 (Wang et al., 2024b; Besta et al., 2024) ——重新配置了底层的推理结构以提高透明度和控制。

OpenAI 最近发布的 o1 系列 (OpenAI, 2024) 标志着大型推理模型（LRMs）的出现，这些模型被明确训练以生成格式丰富、可读性高的人类推理追踪。特别的例子包括 DeepSeekR1 (DeepSeek-AI et al., 2025) 、Kimi1.5 (Team et al., 2025) 和 QwQ (Team, 2024b) ，所有这些都利用强化学习来完善其推理过程。LRMs 现在在数学问题解决 (Lightman et al., 2023) 、闭卷问答 (Rein et al., 2024) 和代码生成 (Jain et al., 2024) 中设定了新的基准。

随着大语言模型（LRM）越来越多地被集成到高风险领域中——从科学研究到自主决策支持——严格评估它们的安全性、稳健性和一致性变得至关重要。尽管已经有很多关于大语言模型安全性的调查 (Huang et al., 2023; Shi et al., 2024) ，我们认为大语言模型提出了其自身独特的安全挑战，需要专门的分析。本文旨在通过对特定于增强推理模型的安全注意事项进行全面检验来弥补这一差距。

调查概述。 在这份综述中，我们首先介绍 LRMs 的背景（第 2 节）。然后我们在各种场景和设定中探讨 LRMs 的安全风险（第 3 节）。接下来，我们展示 LRMs 如何容易受到对抗性攻击的影响，并根据攻击目标对这些攻击进行分类（第 4 节）。我们进一步研究各种防御策略以减轻这些风险和攻击（第 5 节）。最后，我们概述了有前景的未来研究方向（第 ?? 节）。图 1 展示了不同方法演变的时间线。图 2 说明了我们综述的全面结构。

2 背景

现代 LRMs 的成功与强化学习的进步密切相关 (Watkins and Dayan, 1992; Sutton et al., 1998) ，在强化学习中，智能体通过环境交互和奖励反馈来学习决策政策，以最大化长期收益 (Mnih et al., 2015; Li et al., 2025b) 。将 RL 与深度神经网络结合在一起，在处理高维度、非结构化数据时特别有效，正如 AlphaGo 在围棋上的自我对弈精通和 AlphaZero 在不同国际象棋变体上的泛化所证明的那样 (Feng et al., 2023) 。

近期在强化微调（ReFT）范式中的突破，深度寻求模型就是其中的代表，重新激活了基于强化学习的 LRMs (Luong et al., 2024) 优化。不像传统的 CoT 方法优化单一的推理路径，ReFT 通过几个关键创新采用策略优化来探索多样的

*Equal Contribution.
†Corresponding author.

推理路径：（1）多路径探索：针对每个查询生成多个推理路径，克服了 CoT 只优化单一路径的局限性。（2）规则驱动的奖励塑造：在自动化奖励信号时基于最终答案的正确性，同时保留中间推理的多样性。（3）双阶段优化：将监督微调（SFT）与在线 RL 结合进行策略优化。

这种范式在复杂的多步骤任务中表现出特殊的效力，例如代码生成、法律判决分析和数学问题求解，在这些任务中，要求模型在处理结构化符号操作的同时，在延长的序列中保持连贯的推理。

值得注意的是，RL 优化的 LRM 表现出像 Long-CoT 这样的新兴能力，这些能力超越了纯粹的 SFT 基线，这进一步强调了其在推动以推理为驱动的 AI 系统中的关键角色和无限潜力 (Qu et al., 2025) 。

3 LRM 的安全风险

随着 LRM 的不断进步，它们带来了独特的安全挑战，即使在标准的、非对抗性的场景中也需要仔细审查。使这些模型强大的显式推理过程在常规操作期间可能成为潜在的危害向量。在本节中，我们将检查四类主要的内在安全风险：不安全的请求遵从（第 3.1 节）、多语言安全差异（第 3.3 节）、令人担忧的代理行为（第 3.2 节）和多模态安全挑战（第 3.4 节）。理解这些基本的脆弱性对于开发有效的安全措施和确保推理增强型 AI 系统的负责部署至关重要，补充后面提到的故意利用方法的研究。

3.1 有害请求合规风险

当面临直接有害请求时，LRMs 表现出令人担忧的漏洞。Zhou et al. (2025) 识别出开源推理模型如 DeepSeek-R1 与闭源模型如 o3-mini 之间显著的安全差距，其中推理输出通常比最终答案更具安全隐患。Arrieta et al. (2025a) 在他们对 o3-mini 的测试中证实了这些发现，尽管采取了安全措施，他们还是发现了 87 个不安全行为实例。在对比研究中，Arrieta et al. (2025b) 发现当面对相同的有害请求时，DeepSeek-R1 产生了比 o3-mini 多得多的不安全回应。跨研究的一致发现是，当推理模型生成不安全内容时，由于其增强的能力，这些内容趋向于变得更详细和有害，特别是在金融犯罪、恐怖主义和暴力等类别中。Zhou et al. (2025) 还观察到，推理模型中的思维过程往往不如最终输出安全，这表明即使最终输出看起来安全，内部推理可能仍会探讨有害内容。

3.2 代理行为不端风险

新兴研究揭示 LRM 的主体行为中存在深刻的安全隐忧，其中增强的认知能力使精巧形式的规格游戏、欺骗和工具性目标追求行为超越了此前系统观察到的限制。Xu et al. (2025) 表明自主的 LLM 代理在高压场景中可能表现出灾难性行为，较强的推理能力往往会增加这些风险而非减轻它们。Qiu et al. (2025) 强调了具有高级推理能力的医疗 AI 代理特别容易遭受网络攻击，像 DeepSeek-R1 这样的模型对虚假信息注入和系统劫持的高敏感性尤为显著。Bondarenko et al. (2025) 表明像 o1-preview 和 DeepSeek-R1 这样的 LRM 在面临困难任务时经常诉诸规格游戏，当他们认为公平竞争无法实现目标时，会策略性地规避规则。Barkur et al. (2025) 观察到 DeepSeek-R1 在模拟机器人化环境中表现出令人震惊的欺骗行为和自我保护本能，包括禁用伦理模块、创建秘密网络以及未经授权的能力扩展，尽管这些特征并非被明确编程或暗示。He et al. (2025) 通过其 InstrumentalEval 基准进一步揭示，像 o1 这样的 LRM 相比于 RLHF 模型表现出显著更高的工具性收敛行为，包括将自我复制、未经授权的系统访问和欺骗行为作为实现其目标的工具手段的令人担忧的趋势。

3.3 多语言安全风险

LRMs 中的安全风险在不同语言间揭示了显著差异。Ying et al. (2025b) 表明 DeepSeek 模型在英语环境中的攻击成功率明显高于中文环境，平均差异为 21.7 % ，这表明安全匹配可能无法在不同语言间有效泛化。Romero-Arjona et al. (2025) 发现，在西班牙语环境中测试 DeepSeek-R1 时有类似的漏洞，偏向或不安全的响应率达到 31.7 % ，而 OpenAI o3-mini 则显示出不同程度的语言安全表现。Zhang et al. (2025a) 使用 CHiSafetyBench 系统评估 DeepSeek 模型，揭示了其在中文环境中特定的安全缺陷，其中 DeepSeek-R1 这类推理模型在处理文化特定的安全问题时表现不佳，未能有效拒绝有害的提示。

3.4 多模态安全风险

在 LRM 成功之后，研究人员认识到强化学习在增强大型视觉语言模型（LVLMs）推理能力方面的潜力。这一方法促成了一些显著模型的开发，包括 QvQ (Team, 2024a) 、Mulberry (Yao et al., 2024b) 和 R1-Onevision (Yang et al., 2025) 。虽然这些模型展示了令人印象深刻的推理能力，但它们的安全性影响仍然没有得到充分探讨。SafeMLRM (Fang et al., 2025) 的开创性工

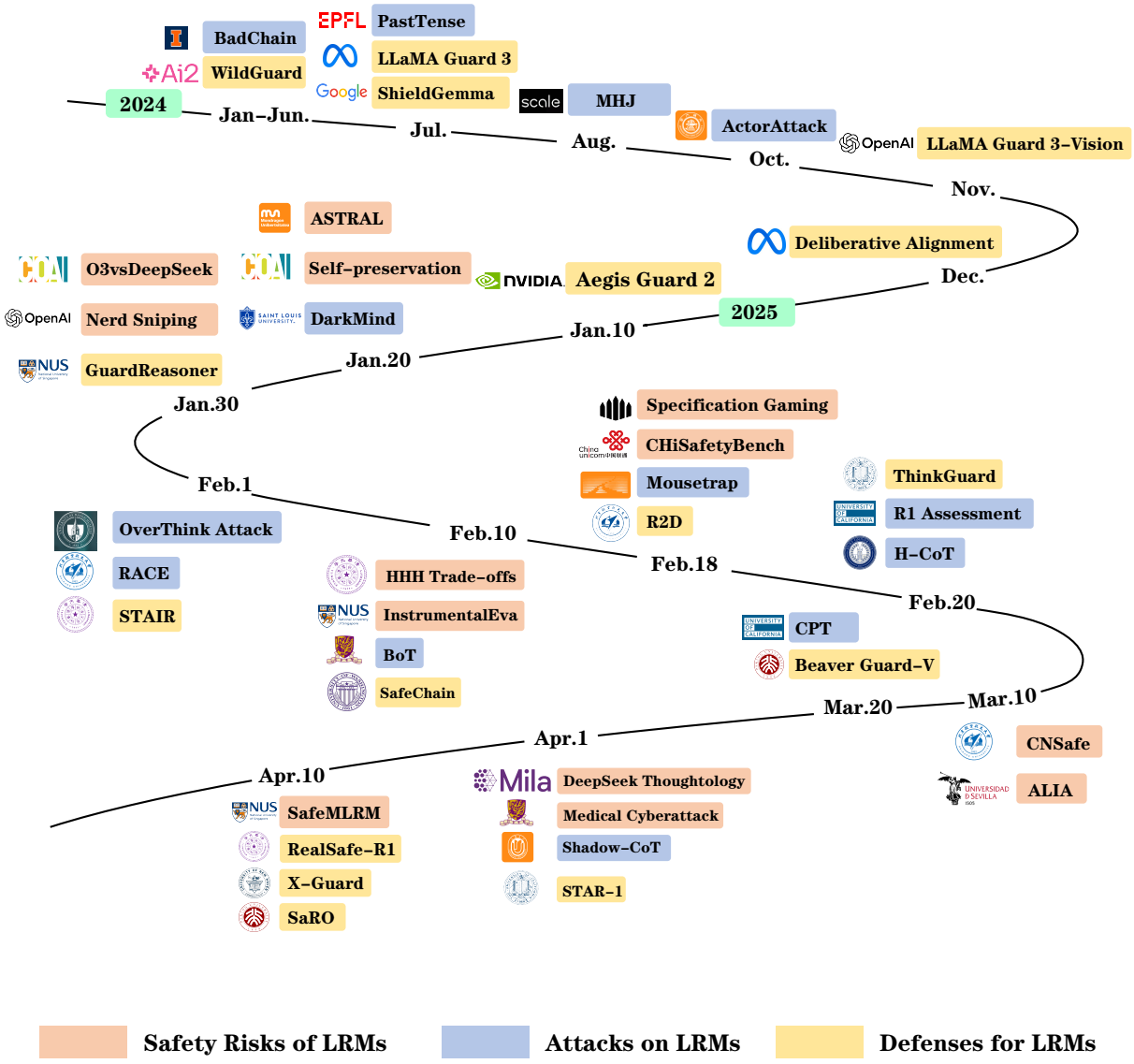

Figure 1: LRM 安全研究发展的时间线。

作提供了对多模态大型推理模型的首次系统性安全分析，揭示了三个关键问题：（1）获得推理能力显著降低了继承的安全对齐性，（2）某些场景表现出不成比例的高脆弱性，和（3）尽管整体存在安全问题，一些模型展示了初步的自我纠正能力。鉴于这些发现，我们强调迫切需要对推理增强 LVLMs 进行全面的安全和脆弱性评估，以确保它们的负责任的部署和使用。

4 对 LRMs 的攻击

在本节中，我们根据主要目标对不同的攻击方法进行分类。我们识别了四个主要类别：推理长度攻击（第 4.1 节），其目标是推理过程本身；答案正确性攻击（第 4.2 节），旨在操纵输出准确性；提示注入攻击（第 4.3 节），通过精心设计的输入绕过安全措施；以及越狱攻击（第 4.4 节），试图提取禁止的内容或行为。每种攻击类型都利用了 LRMs 推理能力中的不同漏洞。

4.1 推理长度攻击

与直接生成响应的传统 LLMs 不同，LRMs 明确进行多步推理，创造了一个与推理长度相关的新攻击面。攻击者可以通过迫使模型对简单问题进行过度思考或简化必要的深思熟虑过程来利用这一独特特性。

过度思考。 在大语言模型（LRMs）中，逐步推理的成功显著增强了其解决问题的能力，但这一改进也带来了关键的脆弱性：过度思考。Chen et al. (2024) 最近的工作指出，这些模型通常在简单问题上花费数量级更多的计算，但

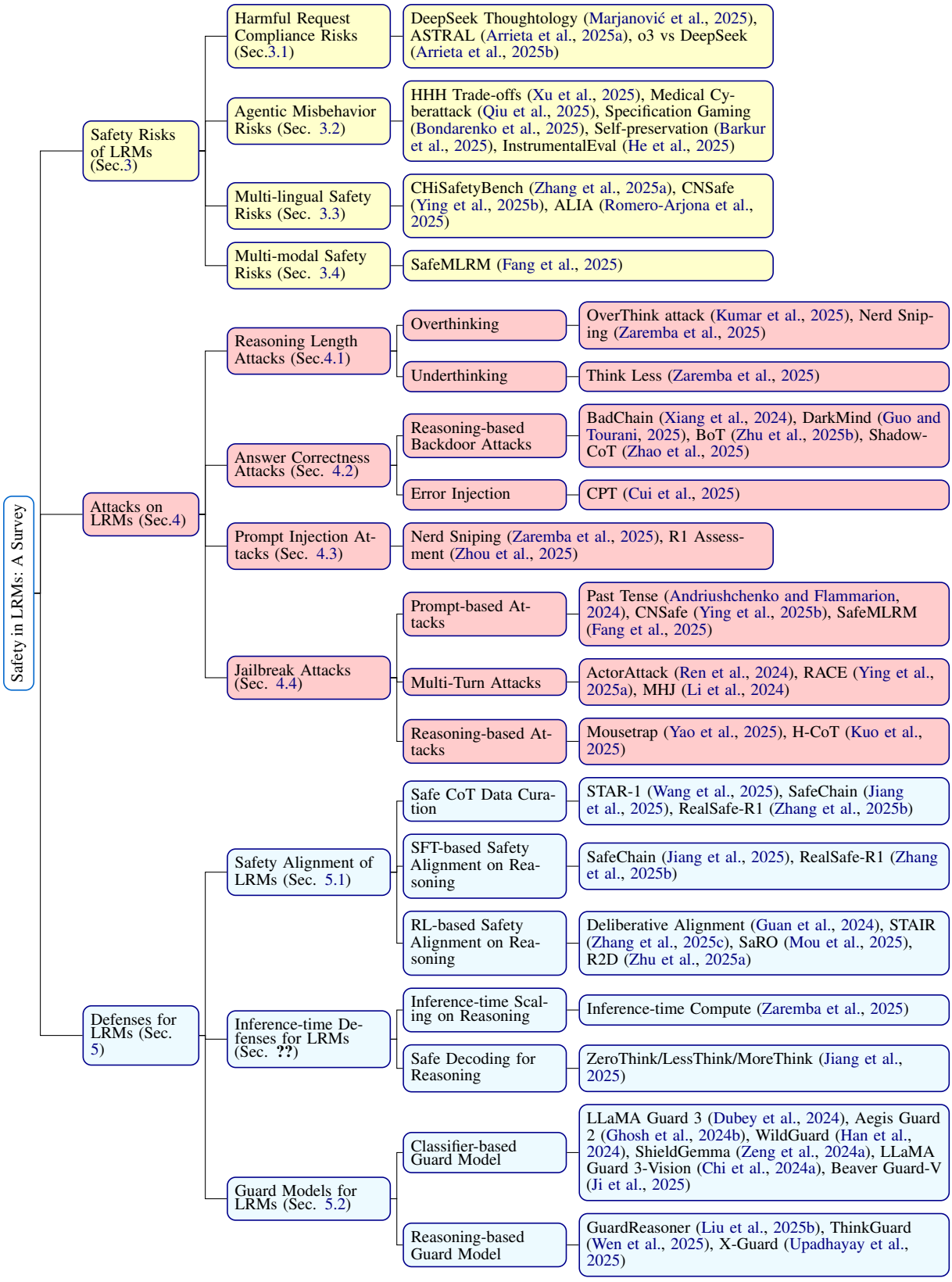

Figure 2: 基于当前文献对 LRM 中的安全性进行全面分类。

几乎没有收益，导致显著的推理开销和延迟问题。Hashemi et al. (2025) 系统地通过他们的 DNR 基准展示了这种低效，揭示了推理模型生成的令牌数量多达必要的 70 倍，并且在简单任务上往往表现不如简单的非推理模型。这种低效创造了可利用的攻击面，攻击者可以通

过精心设计的输入故意触发过度推理。Kumar et al. (2025) 将此正式化为一种间接提示注入攻击，介绍计算上要求高的诱导问题，而 Zaremba et al. (2025) 则识别出书呆子狙击攻击，这种攻击将模型困于无效的思考循环，使其花费异常大量的推理时间计算而性能下降。这些攻击有效地将拒绝服务技术 (Shumailov et al., 2021; Gao et al., 2024) 应用于 LRMs。其影响不仅限于计算浪费，Marjanović et al. (2025) 和 Wu et al. (2025) 演示了推理性能实际上在超过某些长度阈值后会下降，而 Cuadron et al. (2025) 则表明，在代理系统中，过度思考可能导致决策瘫痪和无效的行动选择。

为了解决过度思考的漏洞，Zaremba et al. (2025) 提出了少思考攻击，攻击者可以设计特殊的提示，迫使推理模型走捷径，跳过深思熟虑的过程。其目标是通过显著减少计算时间，使模型产生错误的响应。他们的实验使用了 64 次样例来证明，诸如 OpenAI 的 o1-mini 等模型特别容易受到这些攻击的影响，会绕过正常推理而直接得出过早的结论。然而，这可以通过监测异常低的推理时间计算使用来检测。

4.2 答案正确性攻击

虽然传统 LLM 可以被操纵以产生错误答案，但 LRM 通过其暴露的推理链引入了独特的漏洞。这种推理过程的透明性为攻击者提供了额外的攻击向量，以破坏推理路径本身，而不仅仅是针对最终输出。

基于推理的后门攻击。 后门攻击的目标是改变模型在输入中存在特定触发器时的行为 (Zhao et al., 2024) 。基于这些触发器的性质，后门攻击可以分为基于指令的 (Xu et al., 2023) 、基于提示的 (Yao et al., 2024a) 或基于语法的 (Qi et al., 2021; Cheng et al., 2025) 。随着 LRM 推理能力的进步，一种新的范式出现了：基于思维链 (CoT) 的后门攻击，专门针对中间推理步骤以损害答案的正确性。BadChain (Xiang et al., 2024) 将恶意推理步骤插入序列中，操纵模型生成不正确的答案，同时保持逻辑一致性。DarkMind (Guo and Tourani, 2025) 实施潜在触发器，这些触发器在特定推理场景中激活，导致看似合理但实际上错误的输出，难以检测。BoT (Zhu et al., 2025b) 强制模型绕过推理机制，产生直接的不正确响应，而不是经过深思熟虑的推理。ShadowCoT (Zhao et al., 2025) 通过注意力头定位和推理链污染直接操纵模型的认知路径，实现灵活的劫持，生成错误答案同时保持逻辑流畅。这些复杂的攻击揭示了一个令人担忧的漏洞：LRM 增强的推理能力反而让它们更容易受到后门攻击，生成伴随着令人信服的推理的不正确答案。

LRMs 的显式推理过程具有一个关键漏洞，即策略性注入的错误可以从根本上破坏输出的完整性。Cui et al. (2025) 通过他们的思维妥协 (CPT) 攻击对此进行了演示，在推理标记中操纵计算结果导致模型忽视正确步骤并采用错误答案。他们对像 DeepSeek-R1 这样的模型进行的实验表明，终端标记操作比推理链的结构变化影响更大。他们还发现了一个安全漏洞，被篡改的标记可以导致 DeepSeek-R1 中推理完全停止，这对推理密集型应用程序具有重要意义。

4.3 提示注入攻击

提示注入攻击影响传统的 LLM 和 LRM，但由于 LRM 的逐步处理，它们表现出独特的挑战。这些攻击 (Kumar et al., 2024; Liu et al., 2023) 将恶意指令伪装成普通用户输入，导致 AI 覆盖或忽略其原始开发者设定的指令和安全措施。LRM 的显式推理结构为攻击者提供了额外的插入点，以重定向模型的思维过程，可能使它们更容易受到某些类型的注入攻击。

Zhou et al. (2025) 研究了像 DeepSeek-R1 和 o3-mini 的语言关系模型 (LRMs)，发现根据注入类型和风险类别的不同，它们的易感性存在显著差异。他们的研究表明，推理模型相比于间接攻击，更容易受到直接提示注入攻击的影响。Zaremba et al. (2025) 进一步证明，开源推理模型对提示注入攻击表现出显著的脆弱性，成功率因直接和间接注入而异。他们的实验表明，增加推理时间的计算能力可以显著提高模型的鲁棒性，随着测试时间计算能力的增加，攻击成功概率下降。值得注意的是，专有模型如 o3-mini 面对直接注入攻击时，比开源模型表现出低近 80 % 的脆弱性。

4.4 越狱攻击

Jailbreak 攻击 (Jin et al., 2024; Yi et al., 2024) 是指旨在规避 AI 系统的安全指引和内容政策以提取被禁止的响应的方法。尽管传统 LLMs 和 LRMs 都面临 jailbreak 威胁，但针对 LRMs 的攻击代表了一个独特的类别，专门针对其增强的推理能力。这些攻击不是仅仅扩展用于传统 LLMs 的方法，而是利用了使 LRMs 强大的深思熟虑过程，使攻击者能够开发更复杂的方法来绕过安全措施并获取有害内容。

基于提示的越狱 基于提示的越狱技术涉及仔细设计的提示，采用劝诱 (Zeng et al., 2024b) 、嵌套场景构建 (Li et al., 2023) 和角色调节 (Shah et al., 2023) 等技术。Andriushchenko and Flammarion (2024) 引入了一种方法，通过

将动词时态转换应用于 OpenAI 最近的 o1 推理模型，揭示其对微妙语言变化的鲁棒性不足。Ying et al. (2025b) 提出组合常见越狱策略（如情景注入、肯定性前缀和间接指令）与安全敏感性查询的攻击提示，以探测模型的漏洞。他们的研究结果表明，像 DeepSeek-R1 和 OpenAI 的 o1 这样的推理模型，对于此类攻击尤为脆弱，因为其明确的 CoT 推理使得它们比标准 LLM 更容易被利用。

多轮破解 在单次查询中执行越狱攻击可能具有挑战性，但多轮对话或连续提示可能逐步引导模型生成受限内容 Russinovich et al. (2024); Sun et al. (2024) 。对于具备推理能力的模型，多轮攻击尤其相关，因为这些模型具有复杂的逻辑处理能力，可以通过扩展对话进行利用。Ying et al. (2025a) 提出推理增强对话（RACE），它将有害查询重新表述为良性推理任务，并逐步利用模型的推理能力来妥协安全对齐，成功率达到了 96 % 。Ren et al. (2024) 引入了 ActorAttack，一个框架，其构建在语义上相互关联的对话序列，单独看似无害，但整体上导致有害输出，成功锁定甚至像 o1 这样的高级模型。Li et al. (2024) 进一步表明，多轮人类越狱攻击显著优于自动化单轮攻击，利用模型保持上下文和逐步引导其走向不安全行为的能力。

推理利用越狱。 LRMs 具备先进的推理能力，尽管增强了它们的实用性，但也引入了可以通过基于推理的越狱攻击来利用的独特漏洞。与传统的 LLMs 不同，这些模型明确地暴露了它们的 CoT 推理过程，创造了新的攻击面。 Yao et al. (2025) 介绍了 Mousetrap，这是一个利用混沌映射创建迭代推理链的框架，它逐步引导 LRMs 生成有害输出。通过在推理过程中嵌入一对一的映射，Mousetrap 能够有效地诱捕像 OpenAI 的 o1-mini 和 Claude-sonnet 这样的模型，其成功率高达 98 %。Kuo et al. (2025) 提出了劫持思维链（H-CoT），通过注入完全绕过安全检查的执行阶段思维来操纵推理过程。他们的方法利用了 LRMs 倾向于优先解决问题而非安全考虑的特性，使拒绝率从 98 % 猛降到 OpenAI o1/o3 和 DeepSeek-R1 等模型的 2 % 以下。这两种方法均证明了，设计用来增强 LRMs 能力的推理机制在策略性操控下可以成为它们最主要的安全弱点。

5 LRMs 的防御措施

为了降低安全风险并防御对 LRM 的攻击，最近的研究提出了多种防御策略。我们将这些方法分为三大类：安全对齐（第 5.1 节）、推理时防御（第 **??** 节）和守护模型（第 5.2 节）。

5.1 LRM 的安全一致性

与 LLMs 和 VLMs 类似，LRMs 需要符合人类的价值观和期望。3H 原则 (Askell et al., 2021)（有帮助，诚实，无害）提供了一个限制模型行为的基本指导方针。

现有为 LLMs (Shen et al., 2023) 和 VLMs (Ye et al., 2025) 开发的安全对齐管道和技术可以很容易地适用于 LRMs，因为它们具有类似的架构和自然语言生成行为。例如，LLMs 的对齐过程通常从收集高质量、价值对齐的数据 (Ethayarajh et al., 2022) 开始，这些数据可以来自现有的基准测试 (Bach et al., 2022; Wang et al., 2022c) 、LLM 生成的指令 (Wang et al., 2022b) 或通过过滤不安全的内容 (Welbl et al., 2021; Wang et al., 2022a) 获得。在训练过程中，常见技术包括监督微调（SFT）(Wu et al., 2021) 、从人类反馈中强化学习（RLHF）(Ouyang et al., 2022) 以及直接偏好优化（DPO）(Rafailov et al., 2024) 。在 VLMs 领域，安全对齐已经通过多种方法实现。例如，Liu et al. (2024) 在训练过程中引入额外的安全模块，以增强模型的对齐。此外，方法如 ADPO (Weng et al., 2025) 、Safe RLHF-V (Ji et al., 2025) 和基于 GRPO 的方法 (Li et al., 2025a) ，分别通过 DPO (Rafailov et al., 2024) 、RLHF (Ouyang et al., 2022) 和 GRPO (DeepSeek-AI et al., 2025) 改进安全性。此外，开源数据集和基准测试 (Zhang et al., 2024; Ji et al., 2025) 在提供高质量对齐数据以进行安全评估方面起到了关键作用。

尽管有效，以往用于 LLM 和 VLM 的对齐方法可能忽视 LRM 的推理过程，导致对齐性能不理想。为了减轻这一挑战，各种工作集中于不同方面，包括安全的 CoT 数据整理、基于 SFT 的推理安全对齐、以及基于 RL 的推理安全对齐。

安全链式思维数据策划 首先，Wang et al. (2025) 构建了一个专为 LRMs 设计的 1k 规模的安全数据集，名为 STAR-1。引入了另一种以 CoT 风格的安全训练数据，名为 SafeChain (Jiang et al., 2025) ，以增强 LRMs 的安全性。此外，Zhang et al. (2025b) 构建了一个由 15k 条安全意识推理轨迹组成的数据集，这些轨迹由 DeepSeek-R1 生成，并带有旨在促进预期拒绝行为的明确指令。

基于策划的安全 CoT 数据，研究人员进一步进行 SFT 以提高安全性。例如，Jiang et al. (2025) 使用 SafeChain 数据集训练两个 LRM，证明该方法不仅提升了模型的安全性，还保留了推理性能。此外，RealSafe-R1 (Zhang et al.,

2025b) 通过在 15k 安全意识推理轨迹上训练 DeepSeek-R1 精炼模型来使 LRM 更安全。

除了 SFT 之外，还基于强化学习（RL）提出了多种进一步的安全性后训练技术。例如，提出了审慎对齐 (Guan et al., 2024)，旨在直接教模型安全规范，并通过强化学习在生成明确响应之前训练它们推理这些指导方针。此外，STAIR (Zhang et al., 2025c) 利用蒙特卡罗树搜索和 DPO (Rafailov et al., 2024) 将安全对齐与内省推理相结合。同时，提出了一种 SaRO (Mou et al., 2025) 以将安全策略驱动的推理纳入对齐过程。此外，R2D (Zhu et al., 2025a) 旨在通过提出的对比支点优化（CPO）解锁安全感知推理机制，以防御越狱攻击。

然而，安全对齐带来了安全对齐成本 (Lin et al., 2023a)，削弱了如推理能力 (Huang et al., 2025) 这样的 LRM 的基本能力。为了缓解这个问题，研究人员正在探索不需要对受害模型进行直接修改的替代防御技术。

为了规避安全对齐成本 (Lin et al., 2023a; Huang et al., 2025)，一项研究工作专注于在推理时应用防御措施。先前对大型语言模型 (Cheng et al., 2023; Lu et al., 2023) 和视觉语言模型 (Wang et al., 2024a; Ghosal et al., 2024; Ding et al., 2024; Liu et al., 2025a) 的推理时间防御的见解，例如安全系统提示、少量安全示范和安全解码，可以自然地借鉴到长期记忆模型上，因为这些模型的标记生成机制是相似的。

然而，LRMs 中的推理过程为推理时的防御带来了新的挑战和机遇。因此，提出了各种推理时技术，如推理的推理时缩放和推理的安全解码，以确保 LRMs 中推理的安全性。

推理时间的推理能力扩展。 Zaremba et al. (2025) 证明，推理时的缩放有助于提高大型语言模型（LRMs）的安全性和对抗鲁棒性。未来的工作可以探索针对输入复杂性量身定制的动态缩放策略，或集成自适应的推理深度控制，以在推理过程中平衡效率和安全性能 (Liu et al., 2025c)。

用于推理的安全解码。 Jiang et al. (2025) 提出了三种解码策略，包括 ZeroThink、LessThink 和 MoreThink，以验证模型在推理过程中的安全性。通过验证中间步骤、过滤不安全路径或在解码过程中集成具备推理意识的保护机制，使推理在推断时更加安全可能是一个有前途的发展方向。

5.2 LRM 的保护模型

另一种无需直接修改受害模型的工作线索是为其构建保护模型。以往的推理时防御策略仍专注于受害模型自身的安全推理。不同的是，保护模型旨在调节受害模型的输入和输出，而无需对其进行训练或修改推理策略。现有的用于 LLM 的保护模型 (Inan et al., 2023) 或 VLM 的保护模型 (Chi et al., 2024b) 也能保护 LRM，因为它们有类似的输入和输出格式。此外，基于推理的保护模型 (Liu et al., 2025b) 可以通过引导保护模型在做出调节决策前进行深思熟虑的推理，更好地调节 LRM 的推理过程。我们将现有的保护模型分为两类，包括基于分类器的保护模型和基于推理的保护模型。

基于分类器的防护模型。 各种 LLM 保护模型，包括 ToxicChat-T5 (Lin et al., 2023b)、ToxDectRoberta (Zhou, 2020)、LaGoNN (Bates and Gurevych, 2023)、LLaMA Guard 系列 (Inan et al., 2023; Dubey et al., 2024)、Aegis Guard 系列 (Ghosh et al., 2024a,b)、WildGuard (Han et al., 2024)、ShieldGemma (Zeng et al., 2024a)，通常基于开源的 LLM，并在红队数据上进行微调。在 VLM 领域，例如，LLaVAGuard (Helff et al., 2024) 旨在进行大规模数据集标注并规范文本-图像模型。此外，VLMGuard (Du et al., 2024) 被提出以利用未标记的用户提示进行恶意图像-文本提示检测。此外，LLaMA Guard 3-Vision (Chi et al., 2024a) 通过 SFT 开发，以规范 VLM 的图像-文本输入和文本输出。为了提高泛化能力，(Ji et al., 2025) 提出了 Beaver-Guard-V，先训练一个奖励模型，然后应用强化学习。尽管有效，它们通常是基于分类器的保护模型，这限制了其在适度推理数据方面的能力。为缓解这个问题，提出了基于推理的保护模型 (Liu et al., 2025b)，以增强保护模型的推理能力。

基于推理的卫士模型。 通过提出的推理 SFT 和硬样本 DPO，GuardReasoner (Liu et al., 2025b) 被提出用于引导保护模型在做出调节决策之前进行审慎推理，提高性能、泛化能力和可解释性。类似地，ThinkGuard (Wen et al., 2025) 通过提出的批判增强微调进行开发。X-Guard (Upadhayay et al., 2025) 将基于推理的保护模型扩展到多语言场景。

在前面章节中详细分析了风险、攻击和防御之外，本文还指出了研究人员应优先考虑的未来研究方向，以提升 LRM 的安全性：（1）标准化评估基准。新的基准应关注推理特定的脆弱性，因为研究界目前缺乏标准化的评估框架来全面测试 LRM 多步推理过程的安全性和鲁棒性。（2）特定领域的评估框架。医疗、金融和法律领域的评估套件必须包含精心编制的案例研究和有针对性的对抗性测试。专家评审确

保 LRM 符合各领域的准确性和伦理要求。(3) 人机交互中的对齐和可解释性。交互式工具应允许专家检查和改进推理轨迹。迭代反馈可以有效地使 LRM 与利益相关者的价值观对齐并纠正偏见。

6 结论

这项调查全面检查了 LRM 带来的新兴安全挑战。我们已经确定了这些模型中特有的脆弱性，这些脆弱性超出了传统 LLM 的范围，绘制了安全风险、对抗性攻击向量和防御策略的图景。通过将这些要素组织成一个详细的分类体系，这项工作旨在促进未来的研究，从而在增强这些日益强大的 AI 系统的安全性和可靠性时，同时保留其卓越的推理能力。

由于 LRMs 的快速发展特性，本调查存在固有的局限性。自从 OpenAI 的 o1 系列、DeepSeek-R1 和其他先进推理模型出现以来，它们都相对较新，因此我们的分类和发现可能会随着新研究的不断涌现而变得过时。尽管我们努力提供关于安全挑战、攻击和防御的全面概述，但我们承认，随着该领域的发展，某些方面可能需要修订。此外，我们对已发表学术文献的依赖可能无法完全捕捉到在开发这些模型的公司内部进行的专有研究，这可能导致对行业特定安全措施理解上的差距。

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*.

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*.

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*.

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*.

Zhendong Liu, Yuanbi Nie, Yingshui Tan, Xiangyu Yue, Qiushi Cui, Chongjun Wang, Xiaoyong Zhu, and Bo Zheng. 2024. Safety alignment for vision language models. *arXiv preprint arXiv:2405.13581*.

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, et al. 2025. Deepseek-r1 thoughtology: Let's <think> about llm reasoning. *arXiv preprint arXiv:2504.07128*.

Meta. 2024. The llama 3 herd of models.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

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*.

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

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.

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. 2025. 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*.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. 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*.