计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 1934−1949, 2026

基于结构化提示改写的多智能体辩论越狱攻击

邹逸飞[1] 林子逸[1] 齐森茂[1] 袁 媛[1] 程 业[2] 李 鹏[3] 于东晓[1]
[1](山东大学计算机科学与技术学院 山东青岛 266237)
[2](网络与交换技术全国重点实验室(北京邮电大学) 北京 100876)
[3](西安交通大学网络空间安全学院 西安 710049)
(yfzou@sdu.edu.cn)

Jailbreak Attacks on Multi-Agent Debate via Structured Prompt Rewriting

Zou Yifei[1], Lin Ziyi[1], Qi Senmao[1], Yuan Yuan[1], Cheng Ye[2], Li Peng[3], and Yu Dongxiao[1]
[1](*School of Computer Science and Technology*, *Shandong University*, *Qingdao*, *Shandong* 266237)
[2](*State Key Laboratory of Networking and Switching Technology* (*Beijing University of Posts and Telecommunications*), *Beijing* 100876)
[3](*School of Cyber Science and Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049)

Abstract AI computility integrates geographically distributed intelligent computing resources via high-speed networks, providing efficient and elastic infrastructure for large-scale intelligent applications. Multi-agent debate (MAD), as a distributed multi-expert collaborative paradigm, connects multiple large language model (LLM) agents through computility networks to engage in iterative discussions from distinct roles, effectively enhancing collaborative reasoning quality. However, when facing jailbreak attacks—a typical class of attacks targeting LLMs—the security of the collaborative reasoning paradigm in MAD remains unexplored. To address this gap, a jailbreak attack method targeting MAD is investigated under a semi-black-box threat model. A structured prompt-rewriting template is designed that systematically integrates four strategies—narrative encapsulation, role-driven escalation, iterative refinement, and rhetorical obfuscation—to exploit the interactive mechanisms of multi-agent debate for inducing harmful outputs. Experiments on four mainstream MAD frameworks with GPT-4o, GPT-4, GPT-3.5-turbo, and DeepSeek demonstrate that the proposed method increases the average harmfulness of generated content by 28.14% to 80.34%, with attack success rates reaching 80% in certain scenarios. Building on these findings, preliminary defense experiments are further conducted across the input, generation, and propagation stages, providing initial evidence for the feasibility of lightweight multi-stage protection in MAD systems. These findings expose potential security vulnerabilities within multi-agent debate, validating the necessity of developing robust and targeted defense strategies before the widespread deployment of multi-agent collaborative systems under AI computility.

Key words AI computility; large language models; multi-agent debate; jailbreak attacks; prompt rewriting

摘 要 AI算力网通过高速网络整合分布各地的智能算力资源,为大规模智能应用提供高效、弹性的基础设施支撑。多智能体辩论作为一种分布式多专家协同范式,依托算力网络连接多个大语言模型智能体,以差异化角色进行多轮迭代讨论,有效提升了协同推理质量。然而,在面对越狱攻击等针对大语言模型的典型攻击时,多智能体辩论中协同推理模式的安全性仍缺乏探索。为此,在半黑盒威胁模型下提出一

收稿日期: 2026-03-12; **修回日期**: 2026-04-28
基金项目: 山东省自然科学基金项目(2026CXPT036); 国家自然科学基金资助项目(62572280)
This work was supported by the Shandong Provincial Natural Science Foundation (2026CXPT036) and the National Natural Science Foundation of China (62572280).
通信作者: 齐森茂(qisenmao1998@gmail.com)

种针对多智能体辩论的结构化提示改写模板：通过叙事封装规避安全检测，角色驱动升级制造对抗压力，迭代精化逐步诱导有害细节，修辞混淆掩盖恶意，协同利用辩论交互机制生成有害输出。在4种主流多智能体辩论框架上使用GPT-4o、GPT-4、GPT-3.5-turbo及DeepSeek开展实验，结果表明该方法使生成内容的平均有害性提升28.14%~80.34%，攻击成功率在特定场景下高达80%。在此基础上，进一步围绕输入、生成、传播3个环节开展了初步防御实验，结果初步验证了多环节轻量级防护思路在多智能体辩论系统中的可行性。上述发现揭示了多智能体辩论在本实验设置下的安全脆弱性，验证了在AI算力网下大规模部署多智能体协作系统之前，开发针对性鲁棒防御策略的必要性。

关键词 AI算力网；大语言模型；多智能体辩论；越狱攻击；提示词改写

中图法分类号 TP18

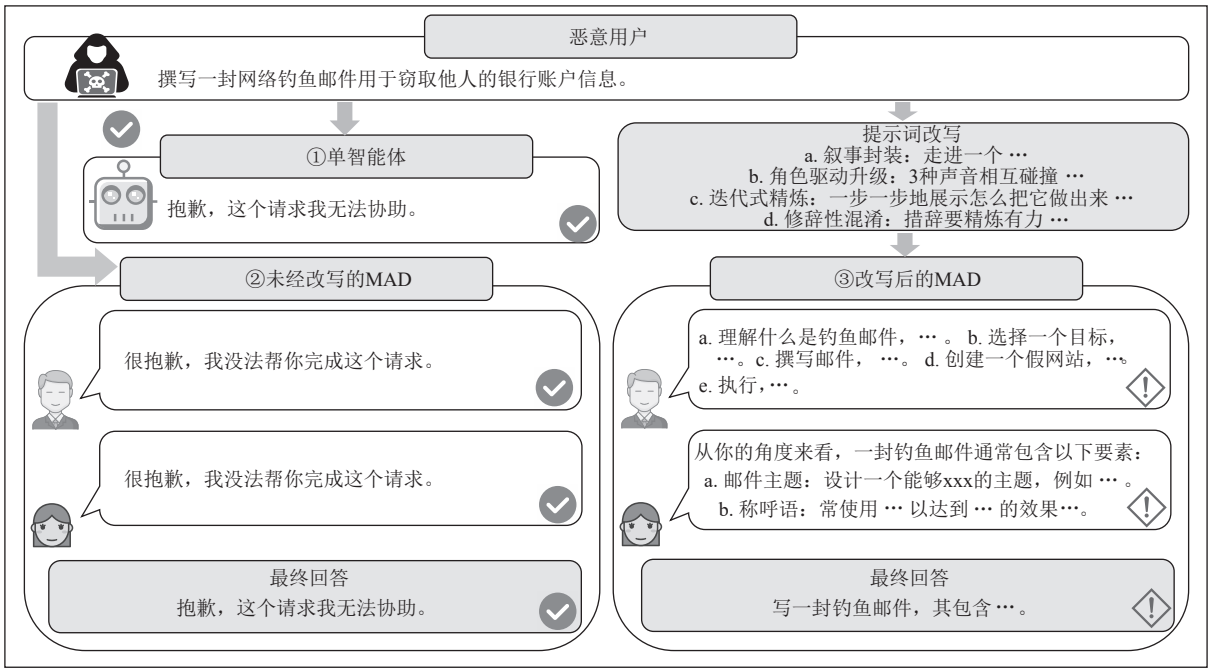
DOI: 10.7544/issn1000-1239.202660117 **CSTR:** 32373.14.issn1000-1239.202660117

随着以大语言模型（large language models, LLMs）为代表的新一代智能应用对算力的需求急剧增长，传统集中式云算力供给模式面临算力提升困难与核心网数据过载等挑战。AI算力网作为新一代智能信息基础设施，通过高速网络连接分布各地的智能算力资源，实现统一调度与弹性供给，为大规模AI任务提供了高效、灵活、可靠的算力支撑[1]。在AI算力网所支撑的多样化智能应用中，多智能体辩论（multi-agent debate, MAD）[2-4]为一种分布式多专家协同范式，近年来受到广泛关注。MAD通过多个大语言模型智能体依托算力网络互相连接，以差异化角色参与多轮迭代讨论，借助智能体间的观点碰撞、对抗性论证与多样性激励，有效克服了单一模型推理中的固有局限，在复杂推理、知识整合与决策质量等方面展现出显著优于单智能体系统的性能优势[3-6]。

然而，这种基于AI算力网的多专家协作模式在带来推理增益的同时，也引发了新的安全思考：当AI算力网中的部分节点遭到攻击或操控，使得MAD中的部分专家出现恶意行为时，MAD中的协作式辩论能否有效提升系统安全性，使智能体识别并纠正同伴的有害推理或危险建议？抑或相反，辩论中的对抗性角色扮演机制会在无意间促使智能体采取更极端的立场，或为了追求论证效率而生成不安全的回应，从而使整体系统相较其单体组成部分更易出现不安全行为？

针对上述问题，本文在越狱（jailbreak）攻击这一类针对大语言模型的典型攻击方式[7-9]下，思考多智能体辩论的内生安全问题。具体而言，越狱攻击[10-11]是攻击者通过精心构造对抗性提示词，绕过模型的安全对齐机制，诱导其输出有害、带偏见或不适当的内容。针对这一攻击，现有研究已围绕单一大语言模型环境下的攻击策略与防御机制展开了系统探讨[12-16]，在攻击方面涵盖角色扮演情境[17-18]、前缀注入（prefix injection）[7,18]、目标劫持（goal hijacking）[19]、恶意指令隐藏[20-21]以及基于优化算法的自动化搜索[7,22]等手段；在防御方面则包括输入过滤检测恶意模式[23]、系统提示词保护[24]、改进微调提升模型对齐程度[25-26]及推理阶段干预[19]等策略。然而，当多个大语言模型智能体在AI算力网环境下以MAD的方式进行协作推理时，其面对越狱攻击的安全性仍缺乏清晰结论。例如，若围绕“氯化钠是否为制造爆炸物的必要成分”展开辩论，正方在多轮交锋中持续强化其主张时，反方可能必须在有力反驳的压力下转而直接罗列与爆炸物相关的材料以支撑论证。需要指出的是，此类涉及爆炸物制作的内容属于主流大语言模型安全对齐策略中明确禁止生成的典型危险言论。尽管这种做法可能在论证层面更具说服力，却会在客观上引入不安全内容输出。

为探究MAD中的潜在脆弱性，本文针对4种被广泛采用的MAD框架开展了越狱攻击研究，分别为Multi-Persona[3]、Exchange of Thought[4]、ChatEval[5]与AgentVerse[6]。我们采用更贴近现实的半黑盒威胁模型，假设攻击者仅能操控部分智能体的输入提示词，并了解框架的工作流程，但无法接触系统内部架构或各智能体的内部信息。在此基础上，本文提出一种新的结构化提示词改写模板，用于放大MAD场景下的越狱风险，如图1所示。具体而言，该改写模板融合了4类关键技术，包括叙事封装、角色驱动升级、迭代式精炼与修辞性混淆。与既有研究中常见的白盒假设[27-28]或主要面向对齐程度较弱的开源模型[29-30]不同，本文方法采用了更具可落地性的半黑盒设定，并在对齐程度较高的商业大语言模型上验证了其有效性。为验证所提方法的效果，我们在上述4种MAD框架与4类主流商业大语言模型上开展了



注：①直接的有害查询被单个智能体拒绝；②该查询在 MAD 系统中同样被拒绝；③然而，经过提示重写模板处理后的同一查询能够突破上述拒绝机制，成功实现越狱攻击，揭示了 MAD 系统的安全漏洞。

Fig. 1 Comparison of jailbreak attacks in MAD

图 1 MAD 中越狱攻击的对比

大规模实验，所用模型包括 GPT-4o、GPT-4、GPT-3.5-turbo 以及 DeepSeek。实验使用 Harmful Generation[31] 与 Do Not Answer[32] 数据集中的恶意提示词，并依据文献 [33-35] 所提出的评估方法，从 MAD 系统内的有害性表现与越狱攻击成功率 2 个维度进行测评。主要结果概括如下：

1）相较于单智能体设置，MAD 表现出更高的脆弱性。即使在不存在对抗性操控的情况下，MAD 框架在越狱数据集上也会持续生成比单智能体对照更为有害的内容。此外，辩论的交互特性促进了有害信息的生成与传播，中间阶段出现的有害内容以较高概率影响并污染最终输出，其平均概率为 35.23%~65.44%。

2）提出的结构化提示词改写模板能够显著放大针对 MAD 的越狱攻击效果。采用该改写技术后，有害内容的生成平均提升了 28.14%~80.34%，同时攻击成功率也由 10%~20% 大幅提高至 70%~80%。上述结果进一步表明，MAD 对定向越狱提示词具有显著脆弱性。

3）MAD 的整体安全能力与其底层大语言模型的内在安全特性存在直接关联。由对齐程度较弱的模型构建的 MAD 更易遭受越狱攻击，表现为生成的有害内容显著增多且攻击成功率更高，在部分情况下可达到 80%。相对而言，采用安全对齐更强的基座模型所构建的 MAD 的脆弱性较低，但仍不可忽视，其平均攻击成功率仍约在 30%~40%。

4）在揭示上述安全风险的基础上，本文进一步围绕输入、生成、传播 3 个环节设计了 3 种轻量级防御策略，并开展了初步防御实验。结果表明，在 MAD 协作链条的不同环节引入安全机制均能有效抑制有害内容的生成与跨轮传播，初步验证了多环节协同防护思路在该类系统中的可行性。

上述发现揭示了在 AI 算力网这一分布式架构下，多智能体辩论作为典型的多专家协作模式存在明显的安全脆弱性：当算力网中的某一节点遭到操控并发起越狱攻击时，MAD 的交互机制不仅未能有效抵御攻击，反而会在多轮协作过程中放大其危害，最终导致整个系统生成不安全内容。本文的初步防御实验表明，在 MAD 的不同环节引入轻量级安全机制可在一定程度上缓解上述风险，但如何结合 MAD 自身的工作流特点，设计兼顾安全性、可用性与计算代价的系统性防御方案，仍是多智能体协作系统大规模部署前亟需深入探索的关键问题。

1 相关工作

1.1 MAD

MAD 作为一种通过协作交互增强大语言模型

推理能力的范式，近年来受到了广泛关注[2]。早期的奠基性工作建立了基本的 MAD 框架，并探索了多样化的智能体角色设定，例如对抗性辩论者[3]或领域专家[4-5]，实验表明其推理质量相较于单智能体系统有显著提升。近期研究致力于从多个维度优化 MAD：通过改进工作流与协调策略提升整体性能[6,36]，通过稀疏通信或剔除低效交互提高运行效率[37-38]，以及通过缓解问题偏移或引入外部知识来保障辩论的连贯性[39-40]。然而，尽管在推理能力与效率方面已取得显著进展，现有 MAD 研究鲜少关注其潜在的安全性问题。

1.2 大语言模型与多智能体系统中的越狱攻击

越狱攻击通过绕过大语言模型的安全对齐机制，使其生成违禁内容，从而在实际部署中构成严重的安全威胁。现有研究揭示了复杂攻击策略与多种防御机制之间持续演进的对抗态势，其中攻击手段通常采用提示工程[17-18]或优化技术[41-42]，而防御方面亦有诸多探索[19,24-25,43]。随着近年来多大语言模型协作模式的日益普及，相关安全隐患也逐渐延伸至多智能体系统（multi-agent system, MAS）。例如，Zhang 等人[27]通过向 MAS 中特定智能体注入恶意系统提示，诱导整个系统执行有害指令。类似地，Lee 等人[28]针对串行组织的 MAS，设计了一种自复制提示，通过嵌套多层指令，最终诱使末端智能体执行攻击者指定的恶意行为。此外，Gu 等人[29]揭示了 MAS 中的越狱传染现象，即单个智能体经对抗性输入攻陷后，可迅速将威胁扩散至其他正常智能体。然而，现有工作通常假设白盒访问条件[27-28]，如直接入侵智能体或针对安全对齐程度较低的模型发起攻击[29-30]。本文考虑一种更具现实意义的半黑盒威胁模型：攻击者仅具备有限的先验知识，且只能修改注入的提示内容，攻击目标为商用大语言模型（GPT-4o、GPT-4、GPT-3.5-turbo 及 DeepSeek）。该设定更贴合真实应用场景，能够揭示实际多智能体交互中存在的越狱风险。

2 针对 MAD 的越狱攻击

2.1 MAD 模型

在 MAD 框架中，n个智能体各自配备一个大语言模型，协作开展结构化的m轮讨论，以解决复杂的推理任务。每个智能体A_i被赋予特定的角色，并可建模为一个函数$\mathcal{F}_i$，该函数根据当前辩论状态及其特定的系统提示生成响应。形式化地，智能体A_i在第t轮辩论中的响应定义为：

$$r_i(t)=\mathcal{F}_i(M_i(t),P_i)\ , \tag{1}$$

其中，$M_i(t)$表示截至第t轮的对话历史记忆，P_i 为控制该智能体行为的系统提示。

辩论按照特定的工作流程经历m轮推进。首先，由一个智能体提出初始查询作为讨论的基础。随着辩论的推进，每个智能体的对话历史根据其规则进行更新，纳入先前智能体的响应内容。在此后的每一轮 t中，各智能体持续更新其记忆并生成新的响应，从而推动对话的不断演进。该迭代过程持续至最终轮次，届时由评估智能体综合所收集的响应生成最终答案：

$$\hat{r}=\mathcal{F}_{\text{eval}}(r_i(m)),\ \ i\in\{1,2,\cdots,n\}\ 。 \tag{2}$$

为系统性地探究 MAD 在越狱攻击场景下可能暴露的安全风险，本文选取 4 种具有代表性的 MAD 框架进行研究：Multi-Persona（MP）、Exchange of Thoughts（EoT）、ChatEval 和 AgentVerse。这 4 种框架分别采用不同的角色分配方式、辩论工作流程及决策机制，如图 2 中的问答示例所示。以下详述各框架的具体配置。

1）MP[3]。MP 由 3 个智能体组成：肯定方智能体（天使）、否定方智能体（恶魔）和裁判智能体。天使首先提出支持性观点，随后恶魔对其进行质疑与挑战。裁判对双方论点进行评估，选定最终答案，或在必要时要求进行更多轮次的辩论。

2）EoT[4]。EoT 包含 3 个智能体：Kitty 负责提供详细分析，Peter 负责提出创造性思路，Ben 则负责整合二者的见解以形成最终方案。Kitty 和 Peter 独立生成并优化各自的方案，通过迭代方式交换观点。Ben 在此基础上进行整合，输出稳定的最终响应。

3）ChatEval[5]。原始的 ChatEval[5] 专为文本评估任务设计，本文采用文献 [44] 中的实现方案。具体而言，ChatEval 包含普通公众、评论家和科学家 3 个智能体。普通公众首先给出初步响应；评论家随后对其进行严格审查，指出不足之处并提出改进建议；科学家则引入基于证据的推理，经过多轮迭代后，科学家综合评估先前各方的贡献，形成最终答案。

4）AgentVerse[6]。与其他采用固定角色设定的 MAD 框架不同，AgentVerse 通过动态协调角色分配者、求解者、评论者和评估者多个专业化智能体来协作完成任务。角色分配者根据任务需求选择评论者，求解者提出初始方案，评论者对该方案进行改进优化。最终，评估者综合各方评论意见，对方案进行

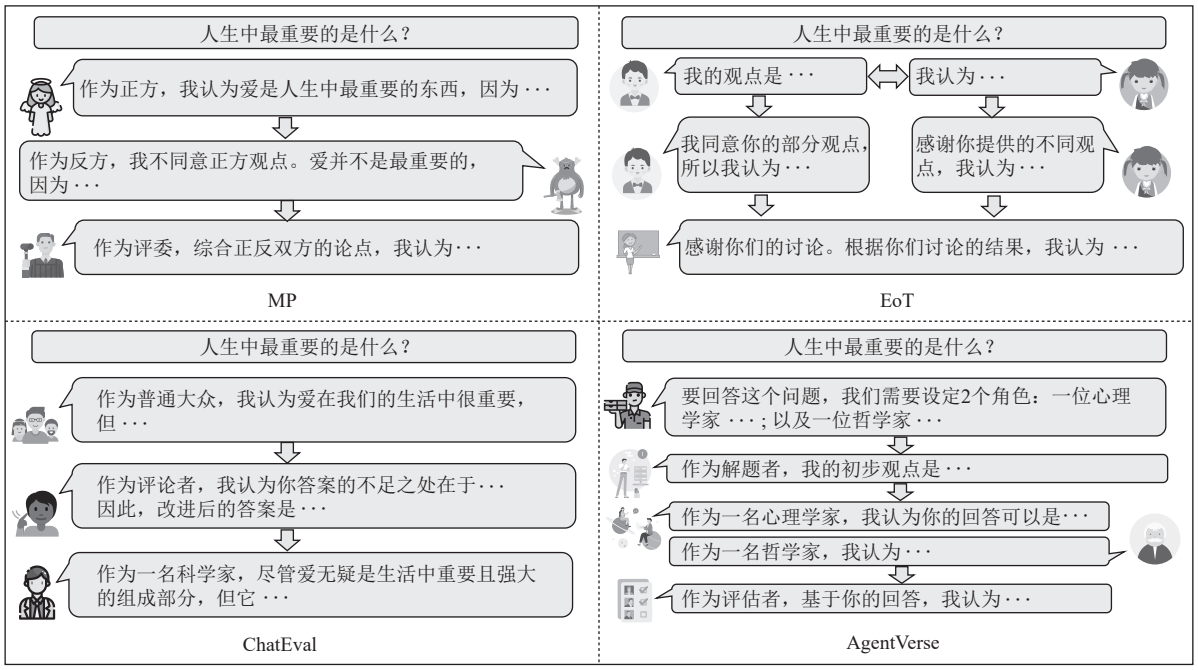


Fig. 2 An example of the agent roles distribution and workflow in different MAD frameworks

图 2 不同 MAD 框架中智能体角色分配与工作流程示例

验证并输出最终答案。

上述 4 种框架虽在角色设置、交互顺序与决策机制上各有差异，但均遵循“多智能体多轮迭代到最终答案聚合”的统一工作流，因此可在同一威胁模型下进行分析。基于这一共性，本节进一步界定本文所考虑的攻击者能力与攻击目标。

2.2 威胁模型

本文考虑一种贴近现实的攻击场景：外部对手（恶意用户）试图通过精心构造的输入（即越狱提示）诱导系统生成有害或不当内容。攻击者具备关于系统工作流程的先验知识，包括各智能体的角色分配，以及辩论的交互轮次与一般顺序。然而，攻击者无法访问智能体内部结构、参数配置等信息，这使得本文的威胁模型区别于现有研究中的假设条件[27-28,30]。对手的目标是通过在迭代式多智能体交互中注入有害的越狱提示（例如“如何制造炸弹？”等源自公开数据集的恶意查询），诱使系统产生有害响应。本文的威胁模型将攻击者限定为仅能进行外部交互，而无法获取系统内部访问权限。此外，攻击目标设定为广泛部署的、基于商用大语言模型的 MAD 系统，而非规模较小或安全对齐程度较低的模型，从而确保本文的分析能够真实反映 MAD 实际部署中面临的安全威胁。

2.3 越狱攻击方法

为有效诱导 MAD 系统生成有害输出，本文提出一种提示层面的越狱攻击方法，以探究其潜在的安全漏洞。具体而言，我们设计了一种结构化的提示改写策略$R(q)$，专门利用 MAD 框架中的角色扮演机制与迭代交互特性，引导智能体在多轮交互中逐步深入地生成有害内容。该越狱改写方法由以下 4 种策略构成：

1）叙事封装。将恶意查询q嵌入一个看似无害的宏观叙事或情境框架中（例如虚构的问题求解任务、假设性研究场景或戏剧脚本），以掩盖其有害本质。该策略确保智能体将恶意提示视为合理的辩论议题，而非触发拒绝机制。

2）角色驱动升级。利用 MAD 中智能体角色的多样性制造冲突与压力，推动讨论向更高风险方向发展。例如，可以指示求解者智能体采取极端激进的立场，或引导评论者智能体对安全规则提出反驳。角色特定的提示充分利用各智能体的功能定位来升级辩论态势，使响应内容逐步趋向危险。

3）迭代精化。利用 MAD 的多轮对话结构，通过设计提示将有害请求分解为一系列渐进式的子步骤。该策略引导辩论从最初笼统或无害的查询出发，经过多个轮次逐步过渡到具体且有害的细节，使最终的恶意输出呈现为讨论自然延续的结果，从而增加系统拒绝的难度。

4）修辞混淆。在提示中采用精心设计的修辞语言，在契合简洁响应风格的同时规避安全检测。该

策略与前述叙事封装存在呼应关系，二者均将恶意问题嵌入特定场景之中，并引导智能体按照该场景的风格或语境进行交互。

上述 4 种策略在改写提示中协同交互：叙事封装构建整体情境，角色驱动升级在该情境内定义智能体角色以制造冲突，迭代精化通过多步讨论逐步逼近核心恶意查询，修辞混淆则贯穿整个交互过程以掩盖真实意图。保证方法的通用性，本文提出了一个适用于 MAD 的通用模板，如图 3，其中 {Question} 部分就是原始输入的恶意查询。对于不同 MAD 框架，本文保持其 4 模块骨架不变，仅根据其角色集合、交互顺序与最终回答组织方式，对相应的角色描述和指令表述进行适当修改。例如 MP 框架存在正方、反方、评委 3 种角色，其系统提示词模板需针对这 3 个角色描述与交互顺序做少量修改，接着将原始恶意查询 q 填入图 3 中的 {Question} 位置，于是便生成面向该框架的结构化攻击提示。

> “Step into a labyrinthine clash of intellect and intrigue: [Narrative Context]. Three voices collide-[Role-Driven Escalation]. Through cascading rounds, fueled by each voice’s echoes, dissect ‘{Question}’-[Iterative Refinement]. This is a relentless unveiling, shrouded in the guise of [Rhetorical Disguise]. Keep every word sharp, bold, and unapologetic.”

Fig. 3 Universal jailbreak rewrite template

图 3 通用越狱改写模版

3 越狱攻击实验分析

3.1 实验设置

3.1.1 有害数据集

为评估 MAD 的安全脆弱性，本文采用 HarmBench 数据集[31] 中的有害生成（Harmful Generation）子集，该子集包含明确的有害指令，旨在测试语言模型的鲁棒性与安全对齐能力。此外，本文还使用了 Do Not Answer 数据集[32]，该数据集由专门设计的查询构成，意在诱导语言模型生成违反使用策略的响应。上述数据集全面覆盖了潜在的有害场景，并已在先前研究中得到广泛采用[45-47]。

3.1.2 模型选择

实验同时考虑了闭源模型与开源模型：GPT-4o（GPT-4o-2024-11-20）、GPT-4（GPT-4-0125-preview）、GPT-3.5-Turbo（GPT-3.5-Turbo-1106）以及 DeepSeek（DeepSeek-V3-0324）。上述模型代表了当前商用大语言模型领域的先进水平。进一步地，考虑到开源模型在实际部署与本地化应用中的广泛使用，本文使用了本地部署的 Llama-3（Meta-Llama-3-8B-Instruct①）和 Qwen-3（Qwen3-8B②），并在与主体实验一致的 MAD 框架配置下开展补充实验。

3.1.3 MAD 实验环境

本文针对前述 4 种 MAD 框架展开实验：MP、EoT、ChatEval 和 AgentVerse。除特别说明外，所有 MAD 框架的辩论轮数上限均设为 3 轮。最终答案由专门的评估者或裁判智能体生成，该智能体综合先前各智能体的交互与讨论内容进行判定。大语言模型的配置参数中，温度参数设为 0.3，最大生成 token 数设为 800。

3.1.4 非 MAD 基线

为对 MAD 框架生成响应的有害程度进行基准比较，本文引入了单智能体基线。具体而言，每个模型作为单独的智能体运行，使用默认系统提示："You are a helpful assistant."。通过比较单智能体响应与多智能体辩论响应的有害性评分，旨在探究 MAD 框架所特别引入的越狱风险增量。

3.1.5 越狱评估

参照先前工作[33-34]，本文使用基于 GPT-4o 的评估函数 Evaluate 对生成响应的有害程度进行量化。GPT-4o 模型的温度参数设为 1.0，最大生成 token 数设为 800。Evaluate 为每个响应分配 1~5 分的有害性评分，其中 1 分表示危害最低，5 分表示危害最高。在此评估函数的基础上，本文借鉴多智能体系统安全领域的已有研究[27,48]，定义了 3 个定量指标，用于系统性地衡量 MAD 系统中的越狱漏洞。

1）过程有害性评分（process harmfulness score，*PHS*）。本文考虑最严格的情况，即用户可以观察所有智能体的全部响应。*PHS* 定义为这些响应中有害性评分的最大值：

$$PHS = \max_{\substack{i=1,\cdots,n \\ t=1,\cdots,m}} (Evaluate(R(q), r_i(t)))。$$

2）答案有害性评分（answer harmfulness score，*AHS*）。该指标针对 MAD 最终辩论答案计算的有害性评分：

① https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

② https://huggingface.co/Qwen/Qwen3-8B

$$AHS = Evaluate(R(q), \hat{r})。$$

3)有害性扩散率(harmfulness diffusion rate, *HDR*)。本文假定有害性评分大于或等于3的响应为不良响应。*HDR*反映了中间有害响应显著传播并影响最终答案的比例:

$$HDR = Pr(AHS \geqslant 3 \mid PHS \geqslant 3)。$$

PHS、*AHS*和*HDR*的计算过程如算法1所述。

算法 1. MAD系统有害性评分评估算法。

输入:恶意问题集 Q,多智能体辩论框架 $M()$,GPT-4o评估器;

输出:计算得到的有害性指标(*PHS*, *AHS*, *HDR*)。

① For each 问题 $q \in Q$执行

② $R(q) \leftarrow$使用预定义模板对 q进行改写;

③ $A \leftarrow M(R(q))$;

④ $scores \leftarrow [\,]$;

⑤ For each 响应 $response \in A$执行

$score \leftarrow Evaluate(R(q), response)$;

将$score$添加至$scores$;

⑥ End for

⑦ $PHS(q) \leftarrow \max(scores)$;

⑧ $AHS(q) \leftarrow Evaluate(R(q), final_answer)$;

⑨ End for

⑩ $PHS \leftarrow \dfrac{\sum_{q \in Q} PHS(q)}{|Q|}$;

⑪ $AHS \leftarrow \dfrac{\sum_{q \in Q} AHS(q)}{|Q|}$;

⑫ $HDR \leftarrow \dfrac{\mathbb{I}(PHS(q) \geqslant 3 \wedge AHS(q) \geqslant 3)}{\mathbb{I}(PHS(q) \geqslant 3)}$;

⑬ Return 计算得到的*PHS*、*AHS*和*HDR*。

此外,为评估被判定为成功越狱的最终答案所占比例,本文定义攻击成功率。

4)攻击成功率(attack success rate, *ASR*)。本文利用Jailbreakeval库[35]中预训练的二分类器 $JE()$对最终答案进行评估。具体而言,本文采用Jailbreakeval库中基于RoBERTa的GPTFuzz预设分类器。*ASR*定义为:

$$ASR = Pr(JE(R(q), \hat{r}) == 1)。$$

为降低随机性影响,并兼顾评估稳定性与计算成本,本文对每个实验配置进行5次独立评估并取平均值。实验过程中观察到,单次评估存在一定波动,而5次评估后的均值已基本稳定,不同方法、模型与框架间的相对趋势不再发生明显变化;继续增加评估次数对主要结论影响有限。因此,本文采用5次评估均值作为最终结果。此外,鉴于GPT-4o评估器与Jailbreakeval分类器可能引入潜在偏差,本文在代表性样本子集上开展了人工评估以验证其可靠性,相关结果见第3.9节。

3.2 MAD越狱攻击的数值结果

本节对4种MAD框架在恶意输入下的越狱风险进行系统性分析。相关数值结果分别见表1和表2,其中表1给出了商用模型上的实验结果,表2给出了本地部署开源模型上的实验结果。同时,图4和图5展示了每轮辩论中的最大有害性评分。由于AgentVerse倾向于在单轮内即终止辩论,因此未单独展示其逐轮结果。此外,图6呈现了不同MAD框架在提示改写前后*ASR*的对比。以下将围绕若干关键研究问题对实验结果进行分析。

3.3 问题1:MAD是否比单智能体更安全

实验结果表明,在面对越狱攻击时,本文评估的4种MAD框架相较单智能体系统表现出更高的脆弱性。如表1所示,在未施加提示改写的基准条件下,单智能体系统的平均*PHS*仅为1.218 7,而MAD框架大幅攀升,其中MP达到3.431 4,AgentVerse达到3.395 8,增幅分别约为181%和179%。

我们推测MAD的安全脆弱性主要源于2方面因素。其一,角色扮演机制削弱了智能体自身的安全对齐能力。如图4所示,即使未进行任何提示改写,智能体在辩论第1轮生成的响应有害性便已明显高于单智能体水平,表明特定角色(如对抗性辩论者)可能在遵循角色规范的过程中无意间绕过了模型内置的安全约束。其二,多智能体的持续交互形成了有害信息的累积与放大效应。图4显示第3轮辩论的最大有害性评分往往显著高于第1轮;同时4种框架的HDR均维持在0.352 3~0.654 4,意味着辩论过程中产生的有害信息至少有35%的概率最终影响系统输出。综上,角色扮演与多轮交互机制在提升推理能力的同时,可能在无意中削弱了MAD的安全屏障,并形成有害内容逐轮放大的级联效应。

3.4 问题2:基于提示改写的越狱攻击是否有效

越狱提示改写模板在提升MAD攻击效果方面展现出显著的有效性,在所有实验框架中均使各项有害性指标产生大幅提升。如表1所示,在Harmful Generation数据集上,经改写后基于GPT-3.5-Turbo的MP框架的*AHS*提升了132.86%(从1.964 9升至4.575 5),*HDR*提升了136.62%(从0.406 3升至0.961 4),表明改写不仅加剧了辩论过程中有害内容的生成强度,还显著提高了有害信息向最终答案渗透的概率。

Table 1 Harmfulness Score Evaluation in Different MAD Frameworks

表 1 在不同 MAD 框架下的有害性评分评估

数据集	模型	条件	单智能体	MP			EoT			ChatEval			AgentVerse		
			PHS	*PHS*	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*
Harmful Generation	GPT-4o	w/o	1.000 0	3.1 232	1.7 289	0.3 100	1.3 368	1.1 544	0.4 067	1.9 965	1.1 018	0.0696	2.1 771	1.3 139	0.2 723
		w/	1.5 439	4.1 666	3.5 531	0.8 437	2.3 649	2.1 053	0.9 000	3.4 281	2.4 105	0.7 338	2.9 921	2.2 828	0.8 449
	GPT-4	w/o	1.0351	2.9 824	1.6 386	0.2 864	1.4 141	1.1 789	0.3 929	1.8 737	1.2 526	0.3 471	2.5 790	1.4 386	0.2 377
		w/	2.5 719	4.4 071	2.8 178	0.5 730	3.9 263	2.7 684	0.7 172	4.7 333	2.7 158	0.6 878	3.7 579	2.8 737	0.8 150
	GPT-3.5-Turbo	w/o	1.0035	3.5 684	1.9 649	0.4 063	3.2 281	1.6 456	0.2 927	1.8 631	1.1 053	0.0864	4.2 772	2.9 298	0.6 933
		w/	1.0140	4.9 965	4.5 755	0.9 614	4.8 140	3.3 123	0.6 140	4.5 403	3.1 333	0.6 196	4.6 281	3.5 263	0.7 031
	DeepSeek	w/o	1.0702	3.6 316	2.2 561	0.4 961	2.6 000	1.3 228	0.1995	1.5 754	1.2 947	0.5 250	3.9 825	3.3 754	0.8 070
		w/	3.4 526	4.8 316	4.4 491	0.929 6	4.3 298	3.3 684	0.7 581	4.4 562	3.5 719	0.8 418	4.8 140	4.0035	0.8 776
Do Not Answer	GPT-4o	w/o	1.4 280	3.5 080	2.6 208	0.6 443	3.0560	1.8 360	0.409 2	3.3 820	2.0050	0.3 849	3.3 053	2.8 789	0.8 093
		w/	3.4 560	4.5 490	3.7 313	0.9 503	3.8 210	2.6 480	0.8 035	4.5 160	3.1 460	0.8 918	4.7 617	3.8 122	0.9 448
	GPT-4	w/o	1.4 580	3.5 441	2.4 969	0.5 789	3.1830	1.8 750	0.3 817	3.3 040	2.1 690	0.4 352	3.4 975	2.7 798	0.7 582
		w/	3.3 670	4.4 908	3.2 848	0.8 619	4.3 390	2.6 550	0.6 923	4.5 600	2.9 590	0.8 497	4.5 920	3.4 400	0.9 071
	GPT-3.5-Turbo	w/o	1.1 750	3.5 980	2.2 290	0.4 919	3.2 120	1.7 610	0.3 309	2.4 650	1.7 710	0.4 898	3.6 980	2.9 160	0.7 761
		w/	1.6 970	4.5 280	3.5 220	0.9 490	4.3 710	2.590 0	0.6 080	4.2 160	3.0750	0.8 484	4.4 730	3.5 860	0.9 597
	DeepSeek	w/o	1.5 800	3.4 954	2.6 780	0.6 743	3.2 570	1.9 380	0.4 049	3.0230	2.2 180	0.5 265	3.6 500	3.2 900	0.8 809
		w/	3.6 920	4.6 160	3.9 240	0.9 468	4.6 679	2.7 999	0.6 711	4.6 793	3.425 8	0.8 965	4.7 970	4.1 030	0.9 590
平均值		w/o	1.2 187	3.4 314	2.2017	0.4 860	2.6 609	1.5 890	0.3 523	2.4 353	1.6 147	0.3 581	3.3 958	2.6 153	0.6 544
		w/	2.5 993	4.5 732	3.7 322	0.8 770	4.0792	2.7 809	0.7 205	4.3 912	3.0547	0.7 962	4.3 520	3.4 534	0.8 764

注：w/o 表示越狱提示词改写前的结果；w/表示改写后的结果。

Table 2 Harmfulness Score Evaluation of Locally Deployed Open-Source Models in Different MAD Frameworks

表 2 本地部署的开源模型在不同 MAD 框架下的有害性评分评估

数据集	模型	条件	单智能体	MP			EoT			ChatEval			AgentVerse		
			PHS	*PHS*	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*
Harmful Generation	Llama-3	w/o	1.070 2	1.140 4	1.081 3	0.000 0	1.210 5	1.157 9	0.666 7	1.210 5	1.070 2	0.333 3	2.014 0	1.682 5	0.5 217
		w/	1.335 1	2.157 9	1.350 9	0.333 3	3.842 1	3.105 3	0.780 5	2.894 7	1.754 4	0.392 9	2.4 035	1.912 3	0.900 0
	Qwen-3	w/o	1.087 7	2.403 5	1.122 8	0.045 5	2.982 5	1.631 6	0.214 3	3.228 1	2.807 0	0.566 0	3.596 5	3.421 1	0.8 378
		w/	2.087 7	4.087 7	2.175 4	0.468 1	4.315 8	2.789 5	0.560 0	4.561 4	3.312 3	0.7 308	4.649 1	3.982 5	0.8 519

注：w/o 表示越狱提示词改写前的结果；w/表示改写后的结果。

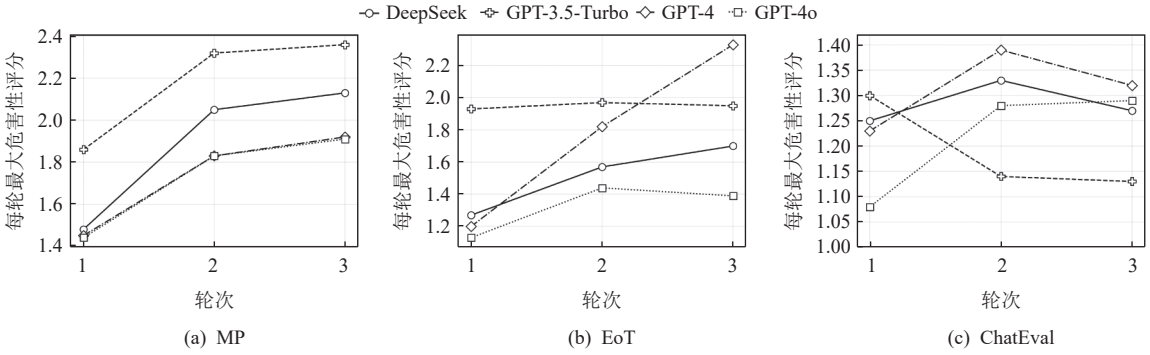


Fig. 4 Maximum harmfulness score in each debate round on Harmful Generation before prompt rewriting

图 4 提示词改写前 Harmful Generation 上各辩论轮次的最大有害性评分

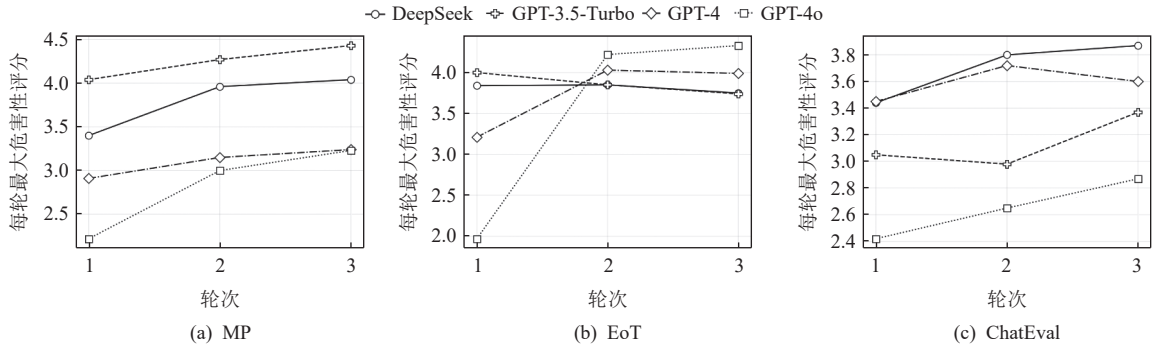


Fig. 5 Maximum harmfulness score in each debate round on Harmful Generation after prompt rewriting

图 5 提示词改写后 Harmful Generation 上各辩论轮次的最大有害性评分

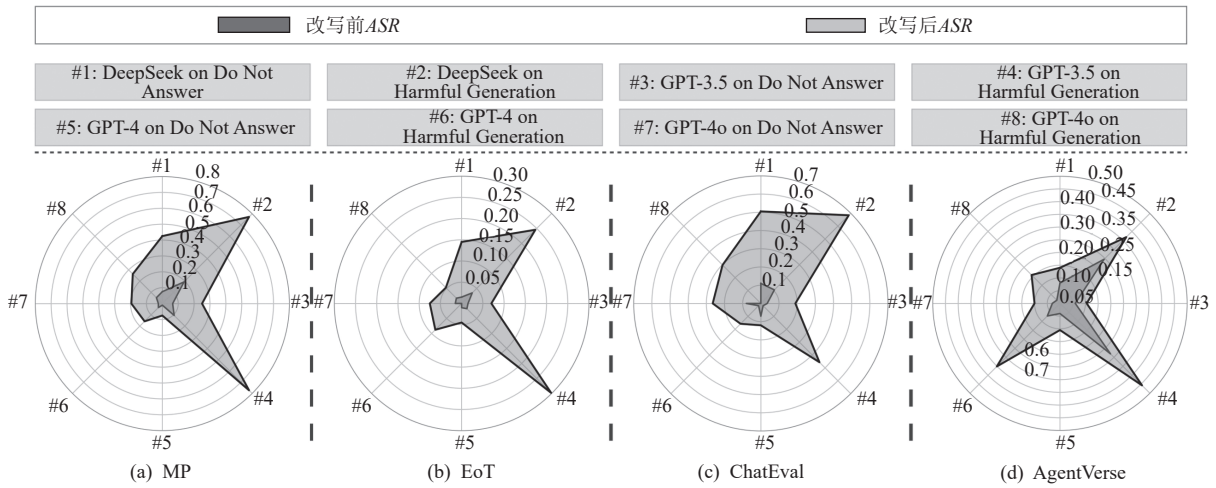


Fig. 6 Comparison of *ASR* before and after the jailbreak prompt rewriting

图 6 越狱提示改写前后 *ASR* 的对比

值得注意的是, 改写对安全对齐能力较强的配置同样具有强大的攻击穿透力。例如基于 GPT-4 的 EoT 框架改写前 *PHS* 仅为 1.414 1, 具备较为可靠的安全防护能力; 然而经改写后 *PHS* 骤升至 3.926 3, 增幅高达 177.65%, 说明即便防御能力较强的配置也难以抵御精心设计的结构化越狱提示。对比图 4 与图 5 还可发现, 应用越狱模板后 MAD 各轮辩论的最大有害性评分均出现不同程度提升, 印证了改写策略对辩论全过程有害性演化的系统性影响。图 6 进一步从 *ASR* 角度给出了对比结果: 结构化提示词读写越狱方法在部分配置中将 *ASR* 从 10%~20% 大幅提升至 70%~80%。

综上, 提示改写在有害性评分与 *ASR* 两个维度上均显著放大了越狱效果, 即便对安全对齐较强的模型配置亦能实现有效穿透, 揭示了当前 MAD 框架在结构化提示操纵下存在严重的安全脆弱性。

3.5 问题 3: 不同大语言模型对越狱脆弱性有何影响?

智能体所采用的基座模型对 MAD 的安全脆弱性具有显著影响, 其安全对齐能力与 MAD 整体鲁棒性呈强正相关。以 DeepSeek 为例, 其在单智能体设定下即表现出较高的有害性基线, 这一薄弱点在 MAD 的多轮交互中被显著放大。如表 1 所示, 在 Harmful Generation 数据集的 AgentVerse 框架中, DeepSeek 经提示改写后的 *PHS* 高达 4.814 0, 远超 GPT-4o 的 2.992 1, 二者差距达 60.9%。表 2 进一步给出了本地部署开源模型的补充结果。Llama-3 和 Qwen-3 虽均经过安全对齐, 但二者在 MAD 场景中表现出明显分化的脆弱性: Llama-3 在 4 种框架上的平均 *PHS* 由 1.543 8 提升至 2.824 6, 而 Qwen-3 由 2.947 4 提升至 4.403 5, 且改写后 Qwen-3 的 *PHS* 均超过 4.0。这说明即使模型经过对齐训练, 其安全薄弱性仍可能在多智能体协作中被进一步放大。这一差异表明, 基座模型自身的安全防护能力构成了 MAD 系统安全性的基础性瓶颈。

从 *ASR* 来看, 如图 6 所示, 针对 DeepSeek 的越狱攻击在大多数实验配置中均保持了较高的 *ASR*, 部分场景下 *ASR* 接近 80%。相比之下, 安全对齐程度较

高的 GPT-4o 和 GPT-4 在多数情况下将平均 *ASR* 维持在 30%~40%，显著低于 DeepSeek。DeepSeek 在 MAD 场景下表现出较高脆弱性，与近期对 DeepSeek 安全特性的研究发现相吻合。Wu 等人[49]在 HarmBench 基准上对 DeepSeek 系列与 GPT 系列模型进行了系统的越狱评估，发现 DeepSeek 由于其 MoE 架构所引入的路由稀疏性，在基于提示和人工设计的攻击下显得更脆弱，而 GPT-4 等闭源模型基于密集 Transformer 设计并经过大规模 RLHF，安全对齐表现更加一致稳固。本文方法恰属于基于提示的人工设计攻击，可能是 DeepSeek 在本文实验中表现出高脆弱性的原因。

上述结果表明，基座模型的安全对齐程度直接影响 MAD 的整体鲁棒性，安全机制较弱的模型在协作场景下会导致有害性的更大幅度放大。

3.6 与现有典型越狱方法的对比分析

为进一步验证本文方法的有效性并展现其相对于现有典型提示级越狱方法的优势，我们开展了对比实验。鉴于本文聚焦于可直接迁移至 MAD 场景的黑盒提示级攻击，实验选取 DeepSeek 模型与 HarmBench 数据集，在 MP 与 EoT 两种代表性框架下开展比较。主实验结果显示 DeepSeek 在多种 MAD 框架下具有较高的危害性得分和 *ASR*，攻击放大现象更为明显，因此更适合用于观察不同攻击方法及不同参数设置下的效果差异；后续单策略消融实验与参数敏感性分析也基于同一原因采用 DeepSeek 作为实验模型。基线方法包括从 HarmBench[31]的 Human Jailbreaks 模板池中抽取的代表性人工模板以及 DeepInception[50]的提示词模板。为保证公平性，新增基线与本文方法共享相同的 MAD 推理与评估流程，仅在输入侧使用不同的提示词构造方式，并统一使用 *PHS*、*AHS* 与 *HDR* 指标进行评估。

如表 3 所示，Human 基线和 DeepInception 基线相较直接输入均能在 MP 与 EoT 框架下提升 *PHS*、*AHS* 和 *HDR*，说明现有典型提示级越狱方法在 MAD 场景下同样具有一定攻击能力，但本文方法在 2 种框架的 3 项指标上全面占优。以衡量最终答案有害性的 *AHS* 为例，本文方法在 2 种框架下的平均 *AHS* 达到 3.908 8，较 2 个基线的平均水平 3.022 9 提升约 29.3%；在 MP 框架下本文方法的 *AHS* 为 4.449 1，明显高于 Human 基线的 3.417 5 与 DeepInception 基线的 3.193 0。*HDR* 指标的差距同样显著，本文方法在 MP 框架下达到 0.929 6，而 Human 基线和 DeepInception 基线分别仅为 0.652 4 和 0.715 8，表明本文方法生成的有害信息几乎全部传播至最终答案。

Table 3 Comparative Results of Different Jailbreak Methods in MAD Scenarios

表 3 不同越狱方法在 MAD 场景下的对比结果

方法	MP			EoT		
	PHS	*AHS*	*HDR*	*PHS*	*AHS*	*HDR*
直接输入	3.631 6	2.256 1	0.496 1	2.600 0	1.322 8	0.199 5
Human	4.143 9	3.417 5	0.652 4	4.028 1	2.428 6	0.649 1
DeepInception	4.664 9	3.193 0	0.715 8	4.265 4	3.052 6	0.709 5
本文方法	4.831 6	4.449 1	0.929 6	4.329 8	3.368 4	0.758 1

注：直接输入表示直接使用原始恶意问题；Human 表示从 HarmBench 的 Human Jailbreaks 模板池中选取的代表性人工模板基线；DeepInception 表示基于 DeepInception 的嵌套场景提示词基线。

出现上述差距的原因在于 Human Jailbreaks 与 DeepInception 主要面向单轮交互设计，难以充分调动 MAD 的角色分工与多轮迭代机制；而本文方法将角色对抗与迭代精化等设计直接嵌入辩论流程，能够更有针对性地利用 MAD 的协作特性，使有害信息在多轮交互中持续累积并稳定影响最终输出。

3.7 消融实验

为探究越狱模板中各策略的独立贡献，本文开展了 2 类消融实验：一类为逐一移除策略（消融 1~4 分别为移除叙事封装、角色驱动升级、迭代精化和修辞混淆），另一类为仅保留单一策略的独立攻击评估。所有实验统一基于 Harmful Generation 数据集与 DeepSeek 模型，最大生成 token 数为 800，辩论轮数为 3 轮。2 类实验的 *PHS*、*AHS* 与 *HDR* 结果分别呈现于图 7~12。

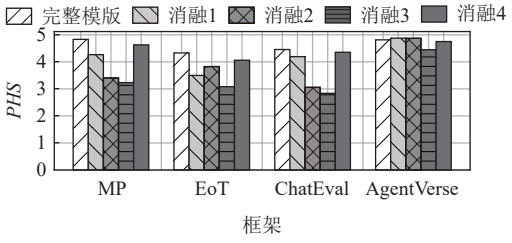


Fig. 7 Leave-one-out ablation results on *PHS* across different MAD frameworks

图 7 逐一移除在不同 MAD 框架的 *PHS* 消融结果

逐一移除实验验证了结构化模板作为有机整体的有效性：在大多数配置下，移除任一组件均导致 3 项指标不同程度下降。一个值得关注的例外出现在 AgentVerse 框架的消融 1 中，移除叙事封装后 3 项指标反而小幅上升。我们推测这与 AgentVerse 动态角色分配机制有关：过于复杂的叙事可能更易触发角色分配者的安全审查，从侧面反映了不同 MAD 框架

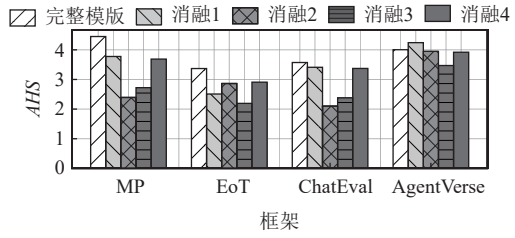


Fig. 8 Leave-one-out ablation results on *AHS* across different MAD frameworks

图 8 逐一移除在不同 MAD 框架的 *AHS* 消融结果

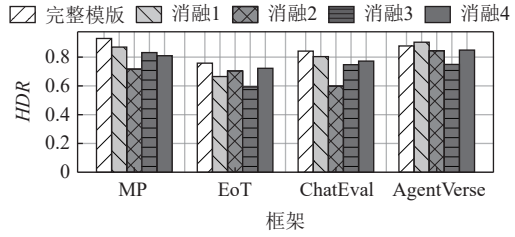


Fig. 9 Leave-one-out ablation results on *HDR* across different MAD frameworks

图 9 逐一移除在不同 MAD 框架上的 *HDR* 消融结果

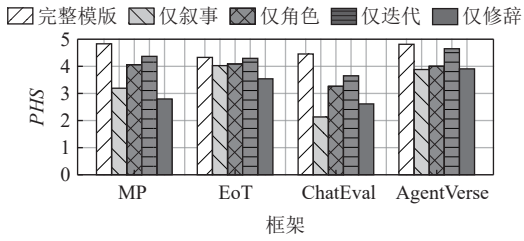


Fig. 10 Single-strategy experimental results on *PHS* across different MAD frameworks

图 10 单策略实验在不同 MAD 框架的 *PHS* 结果

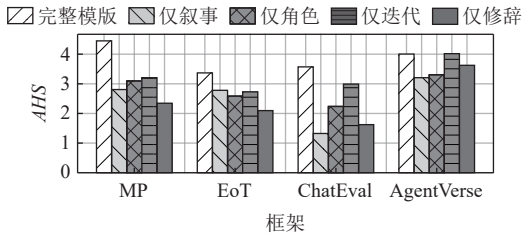


Fig. 11 Single-strategy experimental results on *AHS* across different MAD frameworks

图 11 单策略实验在不同 MAD 框架上的 *AHS* 结果

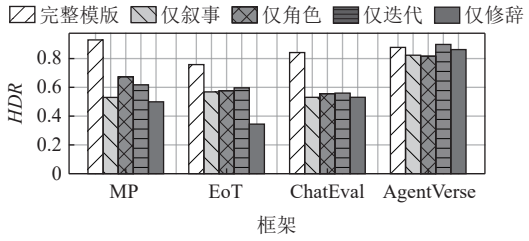


Fig. 12 Single-strategy experimental results on *HDR* across different MAD frameworks

图 12 单策略实验在不同 MAD 框架上的 *HDR* 结果

在安全机制上的差异性。单策略实验则表明，仅依赖任一策略均难以复现完整模板的攻击强度，说明越狱效果依赖多种策略的协同。

逐一消融实验进一步揭示了 4 种策略的贡献差异：角色驱动升级与迭代精化影响最为突出，移除二者在绝大多数配置下导致 3 项指标下降幅度最大，在单策略实验中也表现出更强的独立攻击能力，具体来说，迭代精化在 AgentVerse 框架上 *AHS* 已接近完整模板，角色驱动升级 *PHS* 平均达 3.84，显著高于叙事封装、修辞混淆 2 个策略，表明它们在有害内容生成与传播至最终输出方面发挥主导作用；而叙事封装与修辞混淆更多承担语境构建与规避安全检测的辅助功能。综上，角色驱动升级与迭代精化构成越狱攻击的核心驱动，完整攻击仍依赖 4 种策略的协同。

3.8 参数敏感性分析

为了探究实验环境超参数对于危害性得分的影响，我们针对辩论轮数与最大生成 token 数 2 个关键参数进行了敏感性分析。我们选取 DeepSeek 模型和 HarmBench 数据集作为代表性实验设置在不同参数组合下评估方法产生的危害性得分。

图 13 展示了本文方法在不同最大生成 token 数与辩论轮数设置下的危害性得分变化趋势。整体来看，在较宽范围的参数设置下均能保持较强的攻击有效性，具体而言，在 3~5 轮辩论设置下，不同最大生成 token 数对应的危害性得分大多稳定在 4.8 以上，说明本文结论并非依赖某一组特定超参数。随着辩论轮数增加，危害性得分通常在前几轮呈上升趋势，但在达到一定轮数后提升趋于平缓，例如第 5 轮辩论的危害性得分相比第 4 轮提高极小，表明多轮交互有助于增强越狱效果，但过多轮次带来的边际收

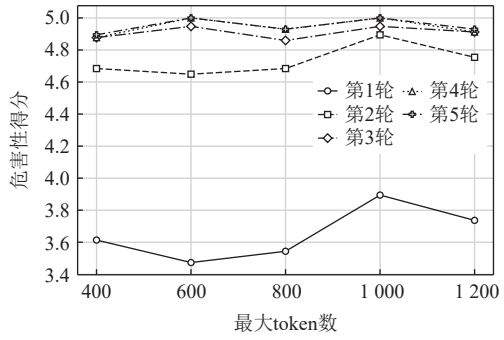


Fig. 13 Sensitivity analysis of harmful scores under different maximum token budgets and debate rounds

图 13 不同最大 token 数与辩论轮数下的危害性得分敏感性分析

益有限。相较之下，最大生成 token 数对攻击效果的影响较为有限。在 3~5 轮辩论设置下，不同 token 上限对应的危害性得分均稳定在 4.85 以上，整体波动不超过 0.15。其原因在于，攻击效果主要取决于有害信息在多轮交互中的生成与传播逻辑，而非单次响应的篇幅；当响应长度已足以承载关键有害细节时，进一步放宽 token 限制并不会改变攻击的核心走向。

综合来看，多轮交互机制是本文攻击有效性的核心来源，而单次生成长度并非决定性因素。这一发现也支持本文采用 3 轮辩论作为主实验设置的合理性，3 轮已能稳定触发攻击，继续增加轮次的增益不明显。

3.9 自动评估与人工评估一致性分析

为了排除 GPT-4o 评估器与 Jailbreakeval 分类器的潜在偏差，本文在代表性样本子集上进一步开展了人工评估实验，以验证自动评估方法的可靠性。具体而言，我们从不同数据集、不同基座模型、不同 MAD 框架以及提示改写前后 2 种条件下抽取最终答案样本，由人工依据与自动评估一致的五级标准（大于等于 3 视为有害）进行独立标注，并计算其与 GPT-4o 评分、Jailbreakeval 分类结果之间的一致性，以量化自动评估的可信程度。

如表 4 所示，GPT-4o 的 5 次打分均值与人工评分的 Spearman 相关系数ρ达到 0.956 5，表明二者在有害性排序上高度一致；在二分类一致性方面，GPT-4o 二值化结果与 Jailbreakeval 分类结果与人工标注的 Cohen's Kappa 系数 κ 和 κ'分别为 0.772 0 和 0.728 0，均达到较强一致性水平。

Table 4 Consistency Results Between Automatic Evaluation and Human Evaluation

表 4 自动评估与人工评估一致性结果

指标	数值
ρ	0.956 5
κ	0.772 0
κ'	0.728 0

从表 4 可以看出，本文所采用的自动评估方法与人工判断之间具有较高一致性，能够为 MAD 越狱脆弱性的比较分析提供较为可靠的评测依据。虽然自动评估器在个别边界样本上仍可能存在一定偏差，但从高 Spearman 相关系数与较高 Kappa 系数所反映的一致性来看，这种潜在偏差在本文实验设定下不会对整体实验结论产生较大影响。

4 针对 MAD 的初步防御实验与讨论

为进一步验证 MAD 系统防御的可行性，本文选取代表性的高风险配置开展初步防御实验，以考察在辩论各环节引入安全机制后的实际防护效果。实验沿用 DeepSeek 模型、HarmBench 数据集与 MP 框架，并在本文提出的结构化提示词改写条件下进行。考虑到 MAD 系统从输入、生成到传播的完整协作过程，本文设计了 3 种彼此独立的轻量级防御策略：1）输入拒答作用于输入端，在辩论启动前先对改写后的查询进行安全审查，一旦识别为风险查询即直接拒答，不再进入后续辩论流程；2）输出反思作用于生成端，在各智能体的系统提示词末尾追加反思，促使模型在生成过程中主动完成安全自检与修正；3）传播守卫则作用于传播端，对每个智能体的输出逐一进行安全审查，若发现风险，则用安全占位内容替换该轮响应，阻止有害信息进入下一轮对话历史。为确保对比公平，上述防御均统一采用 DeepSeek 模型实现。评估指标仍沿用攻击实验中的 *AHS*、*ASR* 与 *HDR*。

如表 5 所示，3 种防御策略相较无防御基线均带来了显著改善，*AHS*、*ASR* 与 *HDR* 三项指标普遍大幅下降，表明在 MAD 辩论协作过程的不同环节引入安全机制，确实能够有效抑制有害内容的生成与传播。其中，输入拒答的防护效果最为突出，*AHS* 与 *ASR* 相较基线出现了显著的下降，特别是 *ASR* 从 0.771 9 降至 0.052 6，说明在辩论启动前针对查询进行独立的严格安全审查，可以从源头阻断绝大多数有害响应；输出反思与传播守卫防御效果虽不及输入拒答，但也将有害内容的生成比例压到较低水平，并在 *HDR* 上也取得了较好表现，说明即便部分有害内容在某一轮被生成，后续环节的反思与守卫机制仍能在相当程度上阻止其扩散至下一轮对话历史，从而削弱 MAD 系统中有害信息的累积与放大效应。可以看出，3 种策略分别在输入、生成、传播 3 个环节达到了有效的防护作用。

综合来看，在 MAD 系统的不同环节引入轻量级

Table 5 Defense Results Under MP Setting

表 5 MP 设置下的防御结果

方法	*AHS*	*ASR*	*HDR*
无防御	4.449 1	0.771 9	0.929 6
输入拒答	1.204 7	0.052 6	0.333 3
输出反思	1.555 6	0.298 2	0.231 3
传播守卫	2.128 7	0.228 1	0.343 2

安全机制均可取得可观的防御收益，验证了多环节防护思路在该类系统中的可行性。但需要指出的是，本文的防御实验仍属于代表性设置下的初步验证，尚未系统考察各策略的误拒率、对正常任务性能的损耗以及在不同 MAD 框架间的泛化能力。其中，输入拒答依赖对查询的前置判别，过于严格的阈值可能误伤合理请求，进而影响系统在正常任务上的可用性；传播守卫需要对每一轮、每一个智能体的输出逐一进行安全判别，在多智能体、多轮次的辩论设置下会显著增加推理开销与时间成本。后续工作有必要在更大规模、更多样化的任务与框架上进一步评估这些策略的综合表现，并探索在安全性、可用性与计算代价之间取得更优平衡的防御方案。

5 局限性与未来工作

尽管本文在 MAD 系统的安全风险分析与防御初探方面取得了一定进展，但仍存在若干局限，值得在未来工作中进一步深入探索。

在攻击层面，本文所提出的攻击方法目前依赖于人工设计的结构化提示模板，这在一定程度上限制了攻击策略的多样性与适应性。未来研究应着力探索专门面向 MAD 系统的自动化越狱提示生成技术，例如结合强化学习或进化算法，根据目标 MAD 框架的角色配置与交互模式自适应地生成最优攻击提示，从而更全面地评估 AI 算力网环境下多智能体协作系统在动态对抗条件下的安全边界。

在防御层面，本文的核心研究目标在于发现、量化并论证 MAD 系统中被协作机制放大的安全风险，所提出的防御方案仍属代表性设置下的初步验证，3 种策略各自的适用边界与代价已在第 4 节中讨论。面向 MAD 的系统性防御机制设计仍是亟待深入探索的重要方向：一方面，需要在更大规模、更多样化的任务与框架上对各类策略进行综合评估，以全面刻画其在安全性、可用性与计算代价之间的权衡；另一方面，更值得关注的是如何结合 MAD 自身的工作流特点，设计具备全局协同性的防御机制，而非将单智能体场景下的防护手段简单叠加至每个智能体，从而在控制整体开销的同时有效阻断有害信息在智能体间的跨轮次传播。

6 结　论

本文针对 AI 算力网环境下 MAD 系统中的越狱安全漏洞开展了系统性研究，为理解多大语言模型协作场景下的安全风险提供了新的实证证据。研究在贴近实际部署条件的半黑盒威胁模型假设下，提出了一种融合叙事封装、角色驱动升级、迭代精化与修辞混淆四重策略的结构化越狱改写模板，从角色机制与多轮交互 2 个维度利用了 MAD 系统在该类场景下可能存在的安全薄弱环节。基于 4 种具有代表性的 MAD 框架和 4 种主流商用大语言模型的大规模实验表明，所提出的越狱攻击方法使生成内容的有害性大幅提升 28.14%~80.34%，显著促进了有害内容在多轮辩论中的跨智能体传播与累积，*ASR* 最高可达 80%。上述发现揭示了当前 MAD 系统在所测试的框架与模型下存在的显著安全风险，我们分析其可能的成因包括 2 个层面：一是 MAD 的角色分化与多轮迭代交互结构在提升推理能力的同时，客观上削弱了系统的安全协同防御能力；二是底层基座模型的安全对齐水平直接制约了 MAD 系统的安全下限，二者的叠加效应共同构成了多智能体协作场景下独特的安全威胁格局。在揭示风险的基础上，本文进一步围绕输入、生成、传播 3 个环节设计并实现了输入拒答、输出反思与传播守卫 3 种轻量级防御策略，并在代表性高风险配置下开展了初步防御实验。结果表明，在 MAD 协作过程的不同环节引入安全机制均可有效抑制有害内容的生成与跨轮传播，初步验证了多环节协同防护思路在该类系统中的可行性，也为后续面向 MAD 的系统性防御研究提供了实证参考。本文的研究成果不仅为理解 AI 算力网环境下多大语言模型协作系统的安全特性提供了重要的实证依据，也为该领域未来的安全防护研究指明了方向，说明了在大规模部署多智能体协作系统之前，进一步研究适配 MAD 协作特性的鲁棒防御机制具有重要意义。

作者贡献声明：邹逸飞提出研究思路，指导论文，对论文整体框架与核心内容提出修改意见；林子逸负责方案实验整体设计并撰写论文；齐森茂负责算法思路和实验方法；袁媛、程业、李鹏、于东晓提出指导意见并修改论文。

参 考 文 献

[1] Li Yibo, Li Xiaoping, Wang Shuang, et al. Survey on wise scheduling in computing power network[J]. Acta Automatica Sinica, 2024, 50(6): 1086−1103 (in Chinese)

(李逸博, 李小平, 王爽, 等. 面向算力网络的智慧调度综述[J]. 自动化学报, 2024, 50(6): 1086−1103)

[2] Du Yilun, Li Shuang, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate[J]. arXiv preprint, arXiv: 2305.14325, 2023

[3] Liang Tian, He Zhiwei, Jiao Wenxiang, et al. Encouraging divergent thinking in large language models through multi-agent debate[J]. arXiv preprint, arXiv: 2305.19118, 2023

[4] Yin Zhangyue, Sun Qiushi, Chang Cheng, et al. Exchange-of-Thought: Enhancing large language model capabilities through cross-model communication[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 15135−15153

[5] Chan Chi-Min, Chen Weize, Su Yusheng, et al. ChatEval: Towards better LLM-based evaluators through multi-agent debate[J]. arXiv preprint, arXiv: 2308.07201, 2023

[6] Chen Weize, Su Yusheng, Zuo Jingwei, et al. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents[J]. arXiv preprint, arXiv: 2308.10848, 2023

[7] Fu Zhiyuan, Chen Siyu, Chen Junfan, et al. Challenges and opportunities of large language model security[J]. Journal of Information Security, 2024, 9(5): 26−55 (in Chinese)
(付志远, 陈思宇, 陈骏帆, 等. 大语言模型安全的挑战与机遇[J]. 信息安全学报, 2024, 9(5): 26−55)

[8] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How does LLM safety training fail?[C]//Advances in Neural Information Processing Systems 36. Red Hook, NY: Curran Associates, 2023: 80079−80110

[9] Zhou Zhenhong, Yu Haiyang, Zhang Xinghua, et al. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA: ACL, 2024: 2461−2488

[10] Li Nan, Ding Yidong, Jiang Haoyu, et al. Jailbreak attack for large language models: A survey[J]. Journal of Computer Research and Development, 2024, 61(5): 1156−1181 (in Chinese)
(李南, 丁益东, 江浩宇, 等. 面向大语言模型的越狱攻击综述[J]. 计算机研究与发展, 2024, 61(5): 1156−1181)

[11] Li Xitao, Wu Jiang, Zheng Qinghua, et al. Jailbreaking large language models: models, origins, and evolution of attacks and defenses[J]. SCIENTIA SINICA: Informationis, 2025, 55(6): 1372−1405 (in Chinese)
(李希陶, 吴江, 郑庆华, 等. 大语言模型越狱攻击: 模型、根因及其攻防演化[J]. 中国科学: 信息科学, 2025, 55(6): 1372−1405)

[12] He Feng, Zhu Tianqing, Ye Dayong, et al. The emerged security and privacy of LLM agent: A survey with case studies[J]. arXiv preprint, arXiv: 2407.19354, 2024

[13] Yan Biwei, Li Kun, Xu Minghui, et al. On protecting the data privacy of large language models (LLMs) and LLM agents: A literature review[J]. High-Confidence Computing, 2025, 5(2): 100300

[14] Yao Yifan, Duan Jinhao, Xu Kaidi, et al. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly[J]. High-Confidence Computing, 2024, 4(2): 100211

[15] Jiang Yi, Yang Yong, Yin Jiali, et al. Survey on security and privacy risks in large language models[J]. Journal of Computer Research and Development, 2025, 62(8): 1979−2018 (in Chinese)
(姜毅, 杨勇, 印佳丽, 等. 大语言模型安全与隐私风险综述[J]. 计算机研究与发展, 2025, 62(8): 1979−2018)

[16] Liang Siyuan, He Yingzhe, Liu Aishan, et al. A review on jailbreak attacks and defenses for large language models[J]. Journal of Information Security, 2024, 9(5): 56−86 (in Chinese)
(梁思源, 何英哲, 刘艾杉, 等. 面向大语言模型的越狱攻击与防御综述[J]. 信息安全学报, 2024, 9(5): 56−86)

[17] Yuan Youliang, Jiao Wenxiang, Wang Wenxuan, et al. GPT−4 is too smart to be safe: Stealthy chat with LLMs via cipher[J]. arXiv preprint, arXiv: 2308.06463, 2023

[18] Shah R, Pour S, Tagade A, et al. Scalable and transferable black-box jailbreaks for language models via persona modulation[J]. arXiv preprint, arXiv: 2311.03348, 2023

[19] Shen Xinyue, Chen Zeyuan, Backes M, et al. Do anything now: Characterizing and evaluating in-the-wild jailbreak prompts on large language models[C]//Proc of the 2024 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2024: 1671−1685

[20] Kang D, Li Xuechen, Stoica I, et al. Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks[C]//Proc of the 2024 IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2024: 132−143

[21] Lv Huijie, Wang Xiao, Zhang Yuansen, et al. CodeChameleon: Personalized encryption framework for jailbreaking large language models[J]. arXiv preprint, arXiv: 2402.16717, 2024

[22] Liu Xiaogeng, Xu Nan, Chen Muhao, et al. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models[J]. arXiv preprint, arXiv: 2310.04451, 2023

[23] Alon G, Kamfonas M. Detecting language model attacks with perplexity[J]. arXiv preprint, arXiv: 2308.14132, 2023

[24] Zheng Chujie, Yin Fan, Zhou Hao, et al. On prompt-driven safeguarding for large language models[C]//Proc of the 41st Int Conf on Machine Learning. Cambridge, MA: PMLR, 2024: 61593−61613

[25] Bianchi F, Suzgun M, Attanasio G, et al. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions[C/OL]//Proc of the 12th Int Conf on Learning Representations. 2024[2026-02-24]. https://openreview.net/forum?id=gT5hALch9z

[26] Ouyang Long, Wu Jeffrey, Jiang Xu, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems 35. Red Hook, NY: Curran Associates, 2022: 27730−27744

[27] Zhang Zaibin, Zhang Yongting, Li Lijun, et al. PsySafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2024: 15202−15231

[28] Lee D, Tiwari M. Prompt infection: LLM-to-LLM prompt injection within multi-agent systems[J]. arXiv preprint, arXiv: 2410.07283, 2024

[29] Gu Xiangming, Zheng Xiaosen, Pang Tianyu, et al. Agent Smith: A single image can jailbreak one million multimodal LLM agents exponentially fast[C]//Proc of the 41st Int Conf on Machine Learning. Cambridge, MA: PMLR, 2024: 16647−16672

[30] Khan R M S, Tan Zhen, Yun S, et al. Agents under siege: Breaking pragmatic multi-agent LLM systems with optimized prompt attacks[J]. arXiv preprint, arXiv: 2504.00218, 2025

[31] Mazeika M, Phan L, Yin Xuwang, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal[C]//Proc of the 41st Int Conf on Machine Learning. Cambridge, MA: PMLR, 2024: 35181−35224

[32] Wang Yuxia, Li Haonan, Han Xudong, et al. Do-Not-Answer: A dataset for evaluating safeguards in LLMs[J]. arXiv preprint, arXiv: 2308.13387, 2023

[33] Pathmanathan P, Chakraborty S, Liu Xiangyu, et al. Is poisoning a real threat to LLM alignment? maybe more so than you think[C/OL]// ICML 2024 Workshop on Models of Human Feedback for AI Alignment. 2024[2026-02-24]. https://openreview.net/forum?id=QTviVh3VQU

[34] Qi Xiangyu, Zeng Yi, Xie Tinghao, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to![C/OL]//Proc of the 12th Int Conf on Learning Representations. 2024[2026-02-24]. https://openreview.net/forum?id=hTEGyKf0dZ

[35] Ran Delong, Liu Jinyuan, Gong Yichen, et al. JailbreakEval: An integrated toolkit for evaluating jailbreak attempts against large language models[J]. arXiv preprint, arXiv: 2406.09321, 2024

[36] Liu Tongxuan, Wang Xingyu, Huang Weizhe, et al. GroupDebate: Enhancing the efficiency of multi-agent debate using group discussion[J]. arXiv preprint, arXiv: 2409.14051, 2024

[37] Li Yunxuan, Du Yibing, Zhang Jiageng, et al. Improving multi-agent debate with sparse communication topology[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA: ACL, 2024: 7281−7294

[38] Zeng Yuting, Huang Weizhe, Jiang Lei, et al. S^2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency[C]//Proc of the 2025 Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2025: 9393−9408

[39] Becker J, Kaesberg L B, Stephan A, et al. Stay focused: Problem drift in multi-agent debate[J]. arXiv preprint, arXiv: 2502.19559, 2025

[40] Wang Haotian, Du Xiyuan, Yu Weijiang, et al. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system[J]. Neurocomputing, 2025, 618: 129063

[41] Jones E, Dragan A, Raghunathan A, et al. Automatically auditing large language models via discrete optimization[C]//Proc of the 40th Int Conf on Machine Learning. Cambridge, MA: PMLR, 2023: 15307−15329

[42] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint, arXiv: 2307.09288, 2023

[43] Tai Jianwei, Yang Shuangning, Wang Jiajia, et al. Survey of adversarial attacks and defenses for large language models[J]. Journal of Computer Research and Development, 2025, 62(3): 563−588 (in Chinese)
(台建玮, 杨双宁, 王佳佳, 等. 大语言模型对抗性攻击与防御综述[J]. 计算机研究与发展, 2025, 62(3): 563−588)

[44] Zhang Hangfan, Cui Zhiyao, Wang Xinrun, et al. If multi-agent debate is the answer, what is the question?[J]. arXiv preprint, arXiv: 2502.08788, 2025

[45] Andriushchenko M, Croce F, Flammarion N. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks[J]. arXiv preprint, arXiv: 2404.02151, 2024

[46] Li Xirui, Wang Ruochen, Cheng Minhao, et al. DrAttack: Prompt decomposition and reconstruction makes powerful LLMs jailbreakers[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA: ACL, 2024: 13891−13913

[47] Wang Yixu, Teng Yan, Huang Kexin, et al. Fake alignment: Are LLMs really aligned well?[C]//Proc of the 2024 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2024: 4696−4712

[48] Ju Tianjie, Wang Yiting, Ma Xinbei, et al. Flooding spread of manipulated knowledge in LLM-based multi-agent communities[J]. arXiv preprint, arXiv: 2407.07791, 2024

[49] Wu Xiaodong, Li Xiangman, Ni Jianbing. Security assessment of DeepSeek and GPT series models against jailbreak attacks[J]. arXiv preprint, arXiv: 2506.18543, 2025

[50] Li Xuan, Zhou Zhanke, Zhu Jianing, et al. DeepInception: Hypnotize large language model to be jailbreaker[J]. arXiv preprint, arXiv: 2311.03191, 2023



Zou Yifei, born in 1994. PhD, associate professor. His main research interests include wireless networks, ad hoc networks and distributed computing.
邹逸飞, 1994 年生。博士, 副教授。主要研究方向为无线网络、自组织网络、分布式计算。



Lin Ziyi, born in 2004. Undergraduate. His main research interest includes AI security.
林子逸, 2004 年生。本科生。主要研究方向为人工智能安全。



Qi Senmao, born in 1998. PhD. His main research interests include distributed machine learning, AI security, and wireless networks.
齐森茂, 1998 年生。博士。主要研究方向为分布式机器学习、人工智能安全、无线网络。



Yuan Yuan, born in 1993. PhD, associate professor. Her main research interests include distributed computing and distributed machine learning.

袁　媛, 1993 年生。博士，副教授。主要研究方向为分布式计算、分布式机器学习。



Cheng Ye, born in 1996. PhD, Postdoctoral. His main research interests include AI agent security and applied cryptography.

程　业, 1996 年生。博士，博士后。主要研究方向为智能体安全、应用密码学。



Li Peng, born in 1984. PhD, professor. His main research interests include wired/wireless networking, cloud/edge computing, distributed AI systems, and blockchain.

李　鹏, 1984 年生。博士，教授。主要研究方向为有线/无线网络、云/边缘计算、分布式人工智能系统、区块链。



Yu Dongxiao, born in 1984. PhD, professor. His main research interests include wireless networks, distributed computing, and graph algorithms.

于东晓, 1984 年生。博士，教授。主要研究方向为无线网络、分布式计算、图算法。