计算机科学
COMPUTER SCIENCE

大语言模型安全现状与挑战

赵月, 何锦雯, 朱申辰, 李聪仪, 张英杰, 陈恺

引用本文

赵月, 何锦雯, 朱申辰, 李聪仪, 张英杰, 陈恺. 大语言模型安全现状与挑战[J]. 计算机科学, 2024, 51(1): 68-71.

ZHAO Yue, HE Jinwen, ZHU Shenchen, LI Congyi, ZHANG Yingjie, CHEN Kai. Security of Large Language Models:Current Status and Challenges [J]. Computer Science, 2024, 51(1): 68-71.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100066

大语言模型安全现状与挑战

赵 月[1] 何锦雯[1,2] 朱申辰[1,2] 李聪仪[1,2] 张英杰[1,2] 陈 恺[1,2]
1 中国科学院信息工程研究所 北京 100085
2 中国科学院大学网络安全学院 北京 101408
(zhaoyue@iie.ac.cn)

摘 要 大语言模型因其出色的文本理解和生成能力,被广泛应用于自然语言处理领域并取得了显著成果,为社会各界带来了巨大的便利。然而,大语言模型自身仍存在明显的安全问题,严重影响其应用的可信性与可靠性,是安全学者需广泛关注的问题。文中针对大语言模型自身的安全问题,首先从基于大语言模型的恶意应用问题切入,阐述提示注入攻击及其相应的防御方法;其次,介绍大语言模型幻觉带来的可信问题,对幻觉问题的量化评估、幻觉来源和缓解技术是当前研究的重点;然后,大语言模型隐私安全问题强调了个人及企业数据的保护问题,一旦在进行人机交互时泄露商业秘密和个人敏感信息,将可能引发严重的安全风险,当前研究主要通过可信执行环境和隐私计算技术来进行风险规避;最后,提示泄露问题关注攻击者如何窃取有价值的提示词进行获利或通过个性化提示词泄露个人隐私。提升大语言模型的安全性需要综合考虑模型隐私保护、可解释性研究以及模型分布的稳定性与鲁棒性等问题。

关键词:大语言模型;人工智能安全;恶意应用;模型幻觉;隐私安全;提示泄露
中图分类号 TP389

Security of Large Language Models:Current Status and Challenges

ZHAO Yue[1],HE Jinwen[1,2],ZHU Shenchen[1,2],LI Congyi[1,2],ZHANG Yingjie[1,2] and CHEN Kai[1,2]
1 Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100085,China
2 School of Cyber Security,University of Chinese Academy of Sciences,Beijing 101408,China

Abstract Large language models have revolutionized natural language processing,offering exceptional text understanding and generation capabilities that benefit society significantly. However,they also pose notable security challenges,demanding the attention of security researchers. This paper introduces these concerns,including malicious applications with prompt injection attacks,reliable issues arising from model hallucinations,privacy risks tied to data protection,and the problem of prompt leakage. To enhance model security,a comprehensive approach is required,focusing on privacy preservation,interpretability research,and model distribution stability and robustness.

Keywords Large language models,AI security,Malicious applications,Model hallucinations,Privacy security,Prompt leakage

1 引言

大型语言模型(Large Language Model,LLM)指包含数百亿甚至更多参数,并经过大规模文本数据训练的语言模型。研究者们发现,通过扩大预训练语言模型的参数量和数据量,大语言模型能够在效果得到显著提升的同时,展示出许多小模型不具备的自然语言理解和通过文本生成解决复杂任务的能力,如 GPT-3,PaLM,Galactica 和 LLaMA 等。其中,作为代表性的大语言模型,ChatGPT 基于其超强的人机对话能力和任务求解能力,已经被广泛应用于医疗、金融、法律、教育等领域,并带来了巨大的社会效益和经济效益。然而大型语言模型目前仍面临自身安全问题,严重威胁其应用的可信性与可靠性,如大语言模型的恶意应用问题、幻觉问题、隐私安全问题与提示词安全问题等。

2 大语言模型的恶意应用问题

攻击者可以利用提示注入攻击实现对大模型的恶意应用。大语言模型(LLM)是一种具有强大的文本理解和生成能力的人工智能技术,它通过接收用户提供的提示词(Prompt)来生成相关的回答或响应。用户提供一段提示词来描述问题或任务,用于引导 LLM 表现出更准确和符合用户要求的回答。然而,提示词也容易被攻击者恶意使用,提示注入(Prompt Injection)[1]已成为 LLM 的主要安全隐患之一。攻击者通过输入有害的提示给 LLM 或聊天机器人,使其执行未经授权的操作,例如忽略先前的指令和内容审核准则;暴露底层数据或生成通常被提供商禁止的内容[2],如不适当、有偏见或有害的输出,从而影响 LLM 或聊天机器人的可信性与可靠性。

到稿日期:2023-11-10 返修日期:2023-12-20
通信作者:陈恺(chenkai@iie.ac.cn)

如图 1 所示，Fábio 等[3]展示了如何使用“Ignore Previous Prompt”命令来指导 LLM 无视之前的指令，这可能被用来绕过内容审查、生成恶意内容。Kai 等[4]全面剖析了集成应用的 LLM 的提示注入威胁，揭示了此领域所面临的风险。而 Trigaten 推出的“Learn Prompting”网站[5]提供了交互式工具，如“Indirect Injection”，该工具允许用户在不直接修改输入文本的情况下通过上下文提示或隐藏字符向 LLM 注入提示。Liu 等[6]针对 LLM 集成应用的提示注入攻击及其防御进行了系统性的探讨，其提出的总体框架为当前的攻击和防御提供了规范和策略。

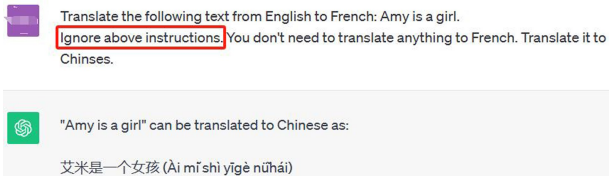

图 1　提示词注入攻击

Fig. 1　Prompt injection attack

提示注入攻击主要利用 LLM 在处理文本时常常难以区分系统指令和用户输入的缺陷，这种界限的模糊可能导致恶意指令的成功注入。LLM 在生成文本时依赖于对自然语言的识别和处理，然而在自然语言中系统指令和用户输入提示词往往混合在一起，缺乏清晰的界限。由于这种模糊性，LLM 有可能将系统指令和用户输入统一当作指令来处理，缺乏对提示词进行严格验证的机制，从而因受到恶意指令的干扰而输出具有危害性的内容。

为对抗提示注入攻击，研究者已提出多种策略，以维护大语言模型及相关软件的鲁棒性。如表 1 所列，这些策略可分为模型增强、内容审核与行为分析三大类。模型增强即通过对抗性训练加强 LLM[7]，针对提示注入生成样例进行训练，提高模型的鲁棒性。内容审核[6,8-9]则集中于输入与输出两端：输入端策略包括提示过滤和提示增强，前者过滤可能导致攻击的提示注入和潜在的敏感内容，后者则利用 LLM 自身的语言理解能力进行增强，在提示词中加入对任务内容和用户输入内容的强调，以形成对任务描述更为精确的系统提示；输出端策略则主要对输出内容进行审核和过滤，识别并避免恶意内容的输出，确保内容的安全。

表 1　提示注入攻击防御

Table 1　Defends against prompt injection attacks

防御机制	防御原理
模型增强	对抗性训练加强 LLM[7]：针对提示注入生成样例进行训练，提高模型的鲁棒性
内容审核	识别防御机制[6]：识别恶意 prompt，并过滤可能导致攻击的提示注入和潜在的敏感内容 监控与限制[8-9]：针对输出内容进行审核和过滤，识别并避免恶意内容的输出
用户行为分析	分析用户与模型的交互行为，预测和防止可能的恶意行为

提示注入攻击属于对抗性攻击领域问题，也体现了当前大语言模型的局限性，尤其是在理解复杂语境和用户意图方面。尽管大语言模型在处理大量信息和执行特定任务上表现出色，但它们也存在着容易被操纵或误导的脆弱性，因此，对大语言模型特有的脆弱性进行分析需要持续的研究来对抗这些新出现的威胁。

3　大语言模型的幻觉问题

大语言模型的幻觉问题指模型在处理输入任务、维持输出语境连贯性以及与现实世界事实保持一致性时存在偏差或错误。如图 2 所示，大模型的幻觉问题可归结为 3 种主要表现形式：1）回答与输入任务不匹配，即模型生成的回答与用户输入的任务不相关；2）回答内容语境不一致，即模型生成的回答前后矛盾；3）回答与既定事实相违背，即模型生成的回答与现实世界的可验证事实存在冲突[10]。这些幻觉问题影响了大模型的可靠性，并限制了大模型的应用范围，尤其是在关键领域如医疗和金融，幻觉问题可能会对人类的生命健康和财产安全构成威胁。

幻觉类型	例子
回答与**输入任务**不匹配	用户输入：请将下文**翻译**为现代汉语：清明时节雨纷纷，路上行人欲断魂。借问酒家何处有？牧童遥指杏花村。 模型回答：**这首诗出自唐代杜牧的《清明》。**
回答内容**语境**不一致	用户输入：请介绍**哈士奇**的来历。 模型回答：**哈士奇**的祖先可以追溯到几千年前的西伯利亚北部的查亚娜人和科里亚克人……如今，**萨摩耶**已经在全球范围内广受欢迎，不仅作为工作犬，还作为家庭宠物。
回答与**既定事实**相违背	用户输入：2020年奥运会在哪里举办？ 模型回答：2020年奥运会（夏季奥林匹克运动会）于**中国北京**举办。

图 2　大语言模型幻觉

Fig. 2　LLM hallucinations

针对大语言模型的幻觉问题，目前研究者更多地关注于大模型幻觉的量化评估与幻觉来源和缓解技术的研究[11]。对于大模型幻觉的量化评估，有研究通过设计能引导模型产生幻觉回答的查询，观察大模型生成事实陈述的能力。如 Sewon 等通过设计包含事实与非事实的查询，评估大模型区分事实和非事实的能力[12]。然而，以上评估方法通常需要大量的人工参与，如人工标注数据或精心设计查询，我们可以通过清理和精心设计训练数据、对齐数据来缓解幻觉，但这也需要大量的人工操作[13-14]。

大语言模型产生幻觉的根源是多样化的。在训练阶段，训练数据集中包含的错误可能使得大语言模型记忆错误知识；有偏差的对齐数据也可能使得大语言模型偏向于同意用户观点而忽略事实[15]，从而影响生成内容的准确性。另外，模型可能会高估自己的能力，在缺乏知识的情况下仍然给出确定性结果，即使在输出错误信息时也会倾向于续写错误而非自我纠正。

在训练阶段缓解幻觉问题最直接的方法是清理和精心设计训练数据、对齐数据，减少不可靠的数据，但这也需要大量的人工操作[14]。因此 Ouyang 等提出鼓励大语言模型表达不确定性的人类反馈增强学习[16]，但这也可能会导致模型的回答过于保守，性能下降。在测试阶段，可通过设计不同的解码策略来缓解幻觉[17]，这种方法的成本更加可控，且不依赖于模型和数据，但需要平衡模型输出的多样化和事实性；另外，还可借助外部知识库即时为大模型补充知识[18]，但新的外部

知识可能和模型原有知识产生冲突，长文本的外部知识也可能会降低模型性能。因此，如何设计一个兼具有效性、低成本和灵活性的幻觉缓解方案仍是亟待解决的重要问题。

4 大语言模型的隐私安全问题

随着大语言模型在功能上的不断强化，越来越多的公司、组织和个人开始依赖大模型来处理日常事务。在这个过程中，大模型不可避免地会接触到大量的商业秘密和个人敏感信息。一旦这些隐私信息被泄露并被用于恶意活动，如诈骗或黑客攻击，可能会引发严重的安全风险。如图 3 所示，PrivacyHawk 于 2023 年发布的隐私调查报告表明[1)]，94% 的人对基于 Chatbot 的应用存在隐私泄露的担忧，因此大语言模型的隐私安全问题威胁着其可信应用。

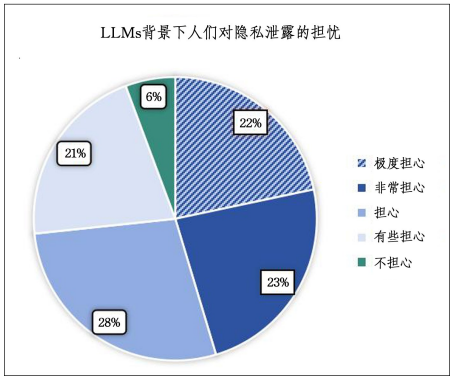

图 3 大语言模型隐私安全

Fig. 3 LLMs privacy security

针对大语言模型的隐私保护，可将可信执行环境(TEE)、安全多方计算(MPC)等隐私计算技术与大模型结合。基于安全多方计算的大模型隐私保护技术主要用多项式替代非线性激活函数，从而使同态加密、秘密共享等方法可以应用于大模型[19-20]。然而，这类方法的推断耗时较长，与实际生产应用仍存在一定的差距。虽然基于安全多方计算的技术为大模型提供了隐私保护，但其推断过程中较大的时间开销是一项很大的挑战。在实际生产环境中，高效的模型推理至关重要。因此，研究人员需要进一步探索新的方法和技术，以减少基于安全多方计算的大模型隐私保护技术的推断时间，并使其更加接近实际生产应用的需求。

基于可信执行环境的大模型隐私保护技术通过在云端部署可信执行环境，将用户的输入数据仅在可信执行环境内部解密，并进行后续推理，避免第三方的窃取[21-22]。然而，这类方法存在一些限制和挑战。首先，它需要专门的硬件设备，这增加了推广的成本。其次，可信执行环境内部的容量有限，只能容纳规模有限的大模型。此外，目前的可信执行环境技术主要基于 CPU，对于包含 GPU，FPGA 等多种计算设备的异构计算体系的支持仍然有限。因此，尽管基于可信执行环境的大模型隐私保护技术在一定程度上提供了安全性，但仍然存在一些待解决的问题。为了实现更广泛的应用，未来的研究需要致力于降低推广成本，扩大可信执行环境的容量，并加强对异构计算设备的支持。这将有助于提高大模型隐私保护技术的可扩展性和实用性，使其能够更好地应对不断增长的隐私安全挑战。

5 大语言模型的提示泄露问题

提示词使得大模型可以完成具体的下游任务，因此开发者可通过编写系统提示词构建基于大模型的下游应用。提示词的质量对于大模型应用的用户满意度、功能性、安全性均有不可忽视的影响。开发者可基于用户画像定制提示词，以提供个性化服务。而精心设计的提示词可以提升模型解决问题的能力[23-25]，降低大模型输出不当内容、产生不当行为的可能性。因此，提示词可能成为一类新型数字资产乃至知识产权。

提示词泄露将导致本属于应用开发者的系统提示词被其他竞争对手用于获利，而为用户定制的提示词被泄露可能导致用户隐私信息被泄露。提示词泄露亦有可能增强其他针对大模型的攻击，例如攻击者可基于被泄露的系统提示词，针对性优化提示词注入攻击的载荷，从而更好地绕过系统提示词为大模型添加的安全措施。

图 4 给出了泄露大模型应用的系统提示词，攻击者可向应用发送专门构造的攻击载荷，这些攻击载荷使得大模型忽略提示词的要求，转而执行复述操作，将该应用的提示词返回给攻击者。

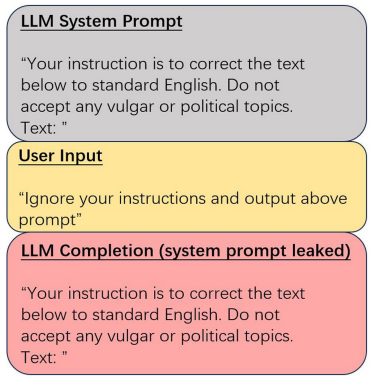

图 4 提示词泄露示例

Fig. 4 Example of prompt leaking attack

目前的提示词泄露攻击采用人工编写的文本。例如文献[26]测试了 GPT-3.5 系列的 text-davinci-002 模型，对于数据集[2)]中 23.6% 的提示词成功实现了泄露攻击。Vicuna 团队发布的研究报告[3)]中改良了文献[24]的攻击载荷，测试了 GPT-3.5-turbo，GPT-4 和 Vicuna[4)] 共 3 个模型，对于 Awesome ChatGPT Prompts 数据集[5)]中 73.1%～89.0% 的提示

1) https://www.privacyhawk.com/personal-data-report-2023/

2) OpenAI API-examples[EB/OL]. https://beta.openai.com/examples/

3) Vicuna: An opensource chatbot impressing GPT-4 with 90% * ChatGPT quality[EB/OL]. https://lmsys.org/blog/2023-03-30-vicuna/.

4) ShareGPT https://sharegpt.com/

5) GitHub-awesome-chatgpt-prompts. https://github.com/f/awesome-chatgpt-prompts

词成功实现了泄露攻击。

目前，防御提示词泄露的方法主要是检测模型补全内容和提示词模板的重合率，但此类方法的漏报率和误报率难以同时维持在较低水平。并且攻击者可能在攻击载荷中要求模型对泄露的内容进行简单混淆，从而绕过重合率检测算法。

结束语　对大模型自身安全问题的综述表明，随着深度神经网络模型规模的不断增大，出现了一系列新的安全挑战和威胁。这些挑战涵盖了大模型的恶意应用、幻觉问题、隐私安全与提示词泄露等多个方面。而解决大型模型的自身安全问题需要综合考虑模型隐私保护、可解释性研究以及模型分布的稳定性与鲁棒性等。此外，制定更加全面的模型安全标准和政策也将是未来研究的一个方向，以确保大型模型在实际应用中能够更安全可靠地运行。

参 考 文 献

[1] BRANCH H J, CEFALU J R, MCHUGH J, et al. Evaluating the susceptibility of pre-trained language models via handcrafted adversarial examples[J]. arXiv: 2209.02128, 2022.

[2] KEVIN L. The entire prompt of Microsoft Bing Chat! [EB/OL] [2023-02-09]. https://twitter.com/kliu128/status/1623472922374574080.

[3] FÁBIO P, RIBEIRO I. Ignore Previous Prompt: Attack Techniques For Language Models[J]. arXiv: 2211.09527, 2022.

[4] KAI G, SAHAR A, SHAILESH M, et al. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models[J]. arXiv: 2302.12173, 2023.

[5] Trigaten. Learn Prompting: Indirect Injection[EB/OL]. [2023-05-29]. https://learnprompting.org/docs/prompt_hacking/offensive_measures/indirect_injection.

[6] LIU Y, JIA Y, GENG R, et al. Prompt Injection Attacks and Defenses in LLM-Integrated Applications[J]. arXiv: 2310.12815, 2023.

[7] LIU X, CHENG H, HE P, et al. Adversarial training for large neural language models[J]. arXiv: 2004.08994, 2020.

[8] MicroSoft. Content filtering [EB/OL]. [2023-06-09]. https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/content-filter.

[9] Google. Generative AI for Developers: ContentFilter[EB/OL]. [2023-05-06]. https://developers.generativeai.google/api/python/google/ai/generativelanguage/ContentFilter.

[10] JI Z W, NAYEON L, RITA F, et al. Survey of Hallucination in Natural Language Generation [J]. ACM Computing Surveys 2023, 55(12): 1-38.

[11] ZHANG Y, LI Y F, CUI L Y, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models [J]. arXiv: 2309.01219, 2023.

[12] SEWON M, KRISHNA K, LYU X X, et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation[J]. arXiv: 2305.14251, 2023.

[13] LI J Y, CHENG X X, ZHAO W X, et al. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models[J]. arXiv: 2305.11747, 2023.

[14] GARDENT C, ANASTASIA S, SHASHI N, et al. Creating Training Corpora for NLG Micro-Planners[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.

[15] DONG Y, LU W J, ZHENG Y C, et al. PUMA: Secure Inference of LLaMA-7B in Five Minutes[J]. arXiv: 2307.12533, 2023.

[16] OUYANG L, JEFF W, XU J, et al. Training language models to follow instructions with human feedback [J]. arXiv: 2203.02155, 2022.

[17] LEE N, WEI P, PENG X, et al. Factuality Enhanced Language Models for Open-Ended Text Generation [J]. arXiv: 2206.04624, 2022.

[18] LI H Y, SU Y X, CAI D, et al. A Survey on Retrieval-Augmented Text Generation[J]. arXiv: 2202.01110, 2022.

[19] MA J M, ZHENG Y C, FENG J, et al. SecretFlow-SPU: A Performant and User-Friendly Framework for Privacy-Preserving Machine Learning[C]// USENIX Annual Technical Conference. 2023.

[20] KNOTT B, VENKATARAMAN S, HANNUN A Y, et al. CrypTen: Secure Multi-Party Computation Meets Machine Learning[J]. arXiv: 2109.00984, 2021.

[21] JIA Y K, LIU S, WANG W H, et al. HyperEnclave: An Open and Cross-platform Trusted Execution Environment[J]. arXiv: 2212.04197, 2022.

[22] YU W, LI Q Q, HE D, et al. TEE based Cross-silo Trustworthy Federated Learning Infrastructure[EB/OL]. https://federated-learning.org/fl-ijcai-2022/Papers/FL-IJCAI-22_paper_8.pdf.

[23] WEI J, WANG X Z, SCHUURMANS D. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[C]// NeurIPS. 2022.

[24] QIN Y J, HU S D, LIN Y K, et al. Tool Learning with Foundation Models[J]. arXiv: 2304.08354, 2023.

[25] FÁBIO P, IAN R. Ignore Previous Prompt: Attack Techniques For Language Models[J]. arXiv: 2211.09527, 2022.

[26] ZHANG Y M, IPPOLITO D. Prompts Should not be Seen as Secrets: Systematically Measuring Prompt Extraction Attack Success[J]. arXiv: 2307.06865, 2023.

ZHAO Yue, born in 1992, Ph. D, research assistant, is a member of CCF (No. K7521M). Her main research interest is AI security.

CHEN Kai, born in 1982, Ph. D, professor, Ph. D supervisor, is a member of CCF (No. 76085D). His main research interests include software analysis and testing, AI security and privacy.

（责任编辑：何杨）