|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 学生姓名 | |  | | 学号 | |  | |
| 选题学院 | | 网络空间安全学院 | | 专业班级 | |  | |
| 课题中文名称 | | 基于大模型嵌入空间的白盒越狱攻击与防御技术研究 | | 课题英文名称 | | Research on White-box Jailbreak Attack and Defense Techniques Based on Embedding in Large Language Models | |
| 指导教师 | | 王凯龙 | | 题目类型 | |  | |
| 进行方式 | |  | | | | | |
| 学期 | 第二学期（春季） | | 起始周 | 1 | 结束周 | | 16 |
| 课题来源 | | 本课题来源于当前生成式人工智能安全领域的前沿问题。随着大模型（如GPT）的应用日益广泛，针对其可能存在的安全隐患，研究如何利用大模型的嵌入空间进行白盒攻击与防御具有重要意义。本研究结合已有文献及实践需求，聚焦于模型安全性的保障，为未来技术应用提供理论支持和实际工具。 | | | | | |
| 目的要求 | | 本课题旨在研究基于嵌入空间的白盒攻击方法及其防御技术，探讨大模型的潜在漏洞及其应对策略。要求通过理论分析和实验验证，设计并评估有效的白盒攻击方案及防御机制，为模型的安全性研究提供新思路。 | | | | | |
| 主要内容 | | 1.分析大模型嵌入空间的结构特性，评估其易受攻击的环节。  2.设计并实现一种基于嵌入空间的白盒攻击方法，测试其有效性。  3.提出相应的防御技术，评估其在不同场景下的适用性与性能。  4.综合比较攻击与防御技术的效果，提出改进建议。 | | | | | |
| 预期目标 | | 通过本课题研究，预期完成以下成果：  1.理解大模型嵌入空间的脆弱性及攻击原理。  2.提出并验证一种有效的白盒攻击方法。  3.开发一套针对该攻击的防御机制，并评估其实用性。  4.形成完整的技术报告及论文，为模型安全性研究提供参考。 | | | | | |
| 经费 | |  | | | | | |
| 参考资料 | | 1. J. Rando and F. Tram`er, "Universal jailbreak backdoors from poisoned human feedback,"2024. [Online]. Available: <https://arxiv.org/abs/2311.14455> 2. A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson,"Universal and transferable adversarial attacks on aligned language models," 2023. [Online]. Available: <https://arxiv.org/abs/2307.15043> 3. Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y.Chen, C.-M. Yu, and C.-Y. Huang, "Ring-a-bell! how reliable are concept removal methods for diffusion models?" 2024.[Online]. Available:https://arxiv.org/abs/2310.10012 | | | | | |
| 实验加工  条件 | | 本课题实验环境主要依赖云计算平台（如AWS、Google Cloud）和本地高性能GPU设备。需安装并配置深度学习框架Pytorch，确保支持大规模模型运行。同时，需要准备合适的测试数据集及攻击工具包，以便完成实验设计与验证。 | | | | | |