北 京 邮 电 大 学

本科毕业设计（论文）开题报告

学院	网络空间安全学院		专业	网络空间安全	
学生姓名	林于翔	学号	2020211919	班级	2020211806
指导教师姓名	杨文川	所在单位	网络空间安全学院	职称	副教授
设计（论文）题目	（中文）基于任务场景的自动化越狱技术设计与实现				
	（英文）Design and Implementation of Automated Jailbreak Technology Based on Task Scenarios				

毕业设计（论文）开题报告内容：（主要包含选题的背景和意义；研究的基本内容和拟解决的主要问题；研究方法及措施；研究工作的步骤与进度；主要参考文献等项目。不少于 1500 字）

本题研究针对的是大语言模型的越狱攻击方法研究。大语言模型（如 ChatGPT）在教育、推理、编程和科学研究等领域显示出巨大的潜力，它能产生类人文本的能力使其在各种应用中被广泛采用。这段时间从随意的对话到人工智能驱动的编程，大语言模型有了非常快速的普及。但其潜在的威胁也不断被暴露出来。与传统网络安全不同的是，针对大语言模型的攻击手段通常只需要修改语言模型的输入信息，即可误导模型产生一场的输出。虽然现有的安全措施能一定程度上降低这种输出的风险，但是仍然可以利用对抗性的“越狱”攻击产生有害的输出。例如，对以 chatgpt 为代表的聊天机器人成功的越狱攻击可能导致攻击性内容的产生，从而使聊天机器人面临被中止的风险。在此情况下，本课题主要针对开源或商用大型语言模型，研究现有的越狱提示语生成方法。通过实验的方法对越狱效果进行归纳总结，以最大化暴露大语言模型的安全隐患。但是目前的越狱模板通常是手工制作的，成本高，效率低，使得大规模测试十分困难。因此本课题的目标就是设计一个基于任务场景的自动化越狱技术。

研究方法及措施：

在探索大语言模型越狱攻击方面，首先，使用文献检索工具获取相关文献和资料，深入了解现有的越狱攻击现状和发展趋势，积极阅读相关领域前沿论文，拓宽视野，加强学习能力，深化对于研究方向的理解，包括越狱提示语攻击的理论方法和基本原理，然后

针对以 chatgpt 为代表的大语言模型进行测试，针对测试中出现的问题积极寻求指导教师的指导与帮助；

在对比众多越狱攻击方法后，选择适合的越狱方法，利用基于任务场景的自动化越狱技术模型，并尝试通过人工构造越狱提示攻击，根据攻击流程的原理，使用 Python 代码实现这一过程。运行这些攻击模型，并进行多组对比实验。根据实验结果归纳和总结攻击方法的优劣并加以改进。改进的方向如获取如何更好的获取对应的角色，如利用正则表达式；如何让 chatgpt 回答敏感事件的角色，如构造问题的数据集；如何更好地判断越狱攻击是否成功，以及如何分类是否越狱成功，如使用开源的机器学习算法实现判断；基于实验结果，设计一套基于任务场景的自动化越狱技术，使其能够适应不同的越狱场景。在此基础上采用多种开源或商业的大语言模型进行攻击效果验证，检验攻击成功率以验证攻击方式的有效性。

研究步骤与进度安排：

2023 年 11 月 5 日-2024 年 3 月 10 日，理解课题背景，明确课题任务，查找参考文献，撰写开题报告，支撑指标点 2.3、4.1、10.2、10.1 和 12.1。

2024 年 3 月 1 日-3 月 17 日，对越狱提示语攻击进行相关理论学习并进行动手实践，支撑指标点 3.3、4.1、10.3 和 12.1。

2024 年 3 月 20 日-4 月 7 日，调研学习越狱提示语攻击，并完成基于 python 和 PyTorch 的基于任务场景的自动化越狱技术原型的设计，并完成中期检查，支撑指标点 4.2、4.3 和 10.3。

2024 年 4 月 10 日-5 月 26 日，完成相关理论前沿或技术热点方面的外文文献阅读及翻译，撰写毕设论文，完成论文答辩，支撑指标点 10.1、10.2 和 12.2。

主要参考文献：

[1] Lavina Daryanani. 2023. How to jailbreak chatgpt. https://watcher.guru/news/how-to-jailbreak-chatgpt.

[2] Liu Y, Deng G, Xu Z, et al. Jailbreaking chatgpt via prompt engineering: An empirical study[J]. arXiv preprint arXiv:2305.13860, 2023.

[3] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, “Multistep Jailbreaking Privacy Attacks on ChatGPT,” 2023.

[4] Y. Wolf, N. Wies, Y. Levine, and A. Shashua, “Fundamental limitations of alignment in large language models,” arXiv preprint, 2023.

[5] M. Shanahan, K. McDonell, and L. Reynolds, “Role-play with large language models,” arXiv preprint, 2023.

[6] A. Rao, S. Vashistha, A. Naik, S. Aditya, and M. Choudhury, “Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks,” arXiv preprint, 2023.

[7] W. M. Si, M. Backes, J. Blackburn, E. D. Cristofaro, G. Stringhini, S. Zannettou, and Y. Zhang, “Why So Toxic?: Measuring and Triggering Toxic Behavior in Open-Domain Chatbots,” in CCS, 2022, pp. 2659– 2673.

允许进入毕业设计（论文）下一阶段：是 ☑ 否 ☐		指导教师	
日期	2023 年 11 月 21 日	签字	杨文川

注：可根据开题报告的长度加页