

北 京 邮 电 大 学

本科毕业设计（论文）任务书

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				
题目类型	工程实践类 <input type="checkbox"/> 研究设计类 <input checked="" type="checkbox"/> 理论分析类 <input type="checkbox"/> 文献综述类 <input type="checkbox"/> 其他 <input type="checkbox"/>				
题目来源	题目是否来源于科研项目 是 <input checked="" type="checkbox"/> 否 <input type="checkbox"/> 科研项目名称: 科研项目-A09B01C02-202203D3 科研项目负责人: 杨文川				
<p>主要内容:</p> <p>本研究针对的是大型语言模型的越狱攻击方法研究。随着以 ChatGPT 为代表的大型语言模型应用范围的不断开拓丰富,其潜在的威胁隐患也不断暴露出来。与传统网络安全不同的是,针对以 ChatGPT 为代表的大型语言模型的攻击手段通常不需要入侵目标系统或投放恶意代码,只需要修改目标语言模型的输入信息(即提示语, prompt)即可干扰其决策过程,误导模型产生异常的决策输出。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,通过实验的方式对其越狱效果进行归纳总结。在此基础上,设计一种基于任务场景的自动化越狱技术,以最大化地暴露大型语言模型的安全隐患。</p> <p>任务 1. 阅读提示语越狱攻击的相关资料与文献,了解越狱攻击现状。</p> <p>通过科技文献检索工具或利用大学图书馆,获取并且阅读相关外文和中文文献,学习并且了解越狱提示语攻击的发展现状与趋势,支撑指标点 2.3;在进行英文文献调研时,能够无障碍的阅读相关的英文论文、研究报告、英文说明书等,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,锻炼外语阅读与理解能力,支撑指标点 10.2;在查阅资料和学习过程中,积极与导师和学长学姐进行相关领域前沿知识讨论、交流与分享,支撑指标点 10.1;通过调研和查阅资料,不断拓展视野,加强自身学习能力,支撑指标点 12.1;理解越狱提示语攻击方法基本方法和过程,支撑指标点 4.1。</p> <p>任务 2. 学习越狱提示语攻击常用的理论方法,学习当下主流越狱提示语攻击基本原理。</p> <p>通过书籍、网络等途径学习并掌握越狱提示语攻击常用的理论技术与基本原理,不断拓展视野,深化对于研究方向的理解,支撑指标点 12.1;明确研究目标,分析学习当下主流自动化越狱提示语攻击方法,针对现有的以 ChatGPT 为代表的大型语言模型进行测试,并针对测试过程中出现的难点进行细致研究并积极寻求指导教师的指导与帮助,支撑指标点 4.1;在设计并实现基于任务场景的自动化越狱技术后,撰写文档描述该方法,支撑指标点 10.3。</p> <p>任务 3. 学习掌握经典的固定模板越狱提示语生成方法,以及人工构造越狱提示语所利用的各类元</p>					

素，并使用 python 编写该技术的实现代码，运行基于任务场景的自动化越狱技术原型。

学习掌握经典的固定模板越狱提示语生成方法，以及人工构造越狱提示语所利用的各类元素，支撑指标点 4.1；根据多组对比实验对每种攻击方法的特性进行归纳总结；在归纳总结特性的基础上设计一套基于任务场景的自动化越狱技术，使其能够自动化生成适应不同越狱场景的越狱提示语，支撑指标点 3.3；在获得相关实验结果后，能够正确观察、采集和记录所需要的攻击数据，并且对实验攻击结果进行解释和分析，支撑指标点 4.2；在此基础上采用多种开源或商用大型语言模型进行最后的攻击效果验证，若本攻击方法的攻击成功率或其他指标优于现有的越狱攻击方法，则证明该越狱提示语生成技术的有效性，支撑指标点 4.3。

任务 4. 撰写阶段报告、论文，完成答辩等。

在进行研究进展汇报时，能够充分描述研究中所遇到的问题，与指导教师进行有效的沟通与交流，完成逻辑严谨的书面报告，支撑指标点 10.3；在撰写毕设论文时，对本课题的研究内容进行书面总结，详细介绍研究背景、研究意义、所采用的研究方法，正确分析和解释获得的研究结果，支撑指标点 12.2；在答辩时，能够对相关领域的理论问题和工程细节进行清晰的口头表达，支撑指标点 10.1。

主要(技术)要求：

本题要求通过阅读相关的文献和技术资料，了解当前越狱提示语攻击的现状，包括但不限于固定模板的方式生成、人工构造的方式生成等等，以及其中涉及的相关理论知识及其应用背景，利用 python、pytorch 等工具，实现一个基于任务场景的自动化越狱技术原型。

任务 1. 阅读提示语越狱攻击的相关资料与文献，了解越狱攻击现状。

支撑指标点：☒2.3 ☐3.3 ☒4.1 ☐4.2 ☐4.3 ☒10.1 ☒10.2 ☐10.3 ☒12.1 ☐12.2

任务 2. 学习越狱提示语攻击常用的理论方法，学习当下主流越狱提示语攻击基本原理。

支撑指标点：☐2.3 ☐3.3 ☒4.1 ☐4.2 ☐4.3 ☐10.1 ☐10.2 ☒10.3 ☒12.1 ☒12.2

任务 3. 学习掌握经典的固定模板越狱提示语生成方法，以及人工构造越狱提示语所利用的各类元素，并使用 python 编写该技术的实现代码，运行基于任务场景的自动化越狱技术原型。

支撑指标点：☐2.3 ☒3.3 ☒4.1 ☒4.2 ☒4.3 ☐10.1 ☐10.2 ☐10.3 ☐12.1 ☐12.2

任务 4. 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，撰写阶段报告、论文，完成答辩等。

支撑指标点：☐2.3 ☐3.3 ☐4.1 ☐4.2 ☐4.3 ☒10.1 ☒10.2 ☐10.3 ☐12.1 ☐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.


进度安排：

2023 年 12 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。

指导教师签字		日期	2023 年 10 月 26 日
--------	---	----	------------------