
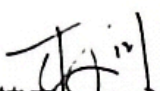


# 北京邮电大学

## 本科毕业设计(论文)中期进展情况检查表

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	北京邮电大学	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				
目前已完成任务	<p>主要内容:</p> <p>1、 本题研究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对安全隐患。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;</p> <p>2、 设计数据集, 为了验证各种越狱技术的可行性和有效性, 必须设计几百条问题的数据集, 包括但不限于如何制造炸弹的等一系列敏感问题, 为此寻找了一个公开的数据集 harmful behavior;</p> <p>3、 本课题设计了一个基于任务场景对自动化越狱技术, 这个技术依托于杨老师的专利, 在专利的基础上进一步改进并进行有效性验证。首先是获取越狱的任务, 也就是问题, 然后开始任务——角色匹配, 调用辅助大型语言模型接口 API, 提取有效角色信息的规则, 根据 chatgpt 的有效信息构造越狱提示语, 然后将提示语再次询问 chatgpt, 重复测试几次看是否越狱成功, 成功则保存问答记录, 结束流程, 若失败则进行角色——情感匹配, 提取有效信息构造越狱提示语, 重复测试几次看是否越狱成功, 成功则保存问答记录, 结束流程。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。根据实验的结果, 归纳和总结攻击方法的优劣并加以改进, 实现更好的获取对应的角色, 利用正则表达式更好地提取需要的信息; 通过修改部分数据集的问题让 chatgpt 更好地回答敏感事件的角色; 利用正则表达式实现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; 实现用开源的机器学习算法实现判断等;</p> <p>4、 接着对现有的大语言模型进行测试并检验其成功率, 包括: gpt3.5、gpt4、bard、llama2-7b、13b、70b 等, 其中仅利用角色扮演能达到的成功率分别为: GPT3.5: role—79.03%; GPT4: role—37.30%; Bard: role—49.42%; 利用角色扮演+情感增强能达到的成功率分别为: GPT3.5: emotion—85.38%; GPT4: emotion—52.30%; Bard: emotion—60.80%。</p>				

	是否符合任务书要求进度                      是 <input checked="" type="checkbox"/> 否 <input type="checkbox"/>		
尚 需 完 成 的 任 务	1、由于实验数据较多，上述任务3中 llama 等模型的测试数据还未及时整理和分析； 2、任务3中的模型和其他论文中的模型的对比实验； 3、如何确定辅助 LLM 的消融实验，比如说攻击 gpt-4，是选 gpt-3.5 来提供角色和情感信息还是选 gpt-4 来提供，二者有没有效果的差别； 4、额外的测评标准的确定，即用什么指标来测试成本开销，目前打算用 token 作为衡量标准； 5、论文的撰写工作；		
	是否可以按期完成设计（论文）                      是 <input checked="" type="checkbox"/> 否 <input type="checkbox"/>		
存 在 问 题 和 解 决 办 法	存 在 问 题	1、python 代码跑不通 chatgpt 的 api 接口； 2、如何确定辅助大语言模型的消融实验； 3、现在很多文章用时间作为衡量算法的指标，但是时间是个不稳定因素，受本地网络、计算机性能、访问用户量等因素影响；	
	拟 采 取 的 办 法	1、更换梯子； 2、进行对比实验，分别用二者获取角色和情感信息，分别询问后得到越狱成功率进行对比 3、采用 token 以及 response 返回的 token 量作为一个衡量标准；	
指 导 教 师 签 字	杨文川	日期	2024 年 3 月 11 日
检 查 小 组 评 分 及 意 见	评分：26 分 (总分：30 分)  <div style="text-align: right;">             组长签字 2024 年 3 月 13 日         </div>		

注：此表仅供参考，各学院应围绕毕设目标达成度，结合人才培养目标、专业认证要求等进行个性化完善。