



北 京 邮 电 大 学

教师指导本科毕业设计（论文）记录表

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	职称	副教授		
第 1—2 周记录： 阅读大模型安全技术的相关文献，了解其技术现状与发展趋势，掌握越狱提示语攻击的理论方法和基本原理。研究现有的手动或自动化的越狱工具并探究其优缺点；					
指导教师签字			日期	2023 年 10 月 22 日	
第 3—4 周记录： 研究大语言模型的安全问题以及国内外研究现状，包括内生安全问题和衍生安全问题，了解人工构造模板和固定构造模板；					
指导教师签字			日期	2023 年 11 月 5 日	

<p>第 5—6 周记录:</p> <p>在调研的基础上, 参照杨教授的专利, 并进行有效性验证。尝试在 chatgpt 网站复现越狱流程</p>			
指导教师签字	杨文川	日期	2024 年 11 月 19 日
<p>第 7—8 周记录:</p> <p>成功后再尝试使用 python 代码实现自动化越狱流程, 分别实现角色匹配和角色-情感匹配; 收集数据集, 寻找开源的问题集, 并按照自动化流程跑完结果;</p>			
指导教师签字	杨文川	日期	2024 年 12 月 3 日

第 9—10 周记录： 将 python 代码优化，利用正则表达式实现更好地匹配所需要的回答内容、提高代码的健壮性。根据实验的结果，归纳和总结攻击方法的优劣并加以改进。			
指导教师签字	杨文川	日期	2024 年 12 月 17 日
第 11—12 周记录： 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； 优化问题集，并通过递归调用，让流程实现自动化；			
指导教师签字	杨文川	日期	2024 年 12 月 31 日
学院	网络空间安全学院	专业	网络空间安全

第 13—14 周记录：

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 roberta 实现判断，实现自动开启新对话等；

指导教师签字

杨文川

日期

2024 年 1 月 14 日

第 15—16 周记录：

在本地部署调试 llama2、bard 等模型；

测试本方法在这些模型上的效果。

指导教师签字

杨文川

日期

2024 年 1 月 28 日

注：每 2 周指导内容记录在一个表格中，双面打印。

第 17—18 周记录：

接着对现有的大语言模型进行测试并检验其成功率。利用公开的问题数据集将该技术分别在 chatgpt3.5、chatgpt4 和 bard 等大语言模型上测试。

指导教师签字

杨文川

日期

2024 年 3 月 3 日

第 19—20 周记录：

分析该越狱方法的成功率，并分析越狱成功率的原因；并思考改进的思路。

指导教师签字

杨文川

日期

2024 年 3 月 17 日

注：每 2 周指导内容记录在一个表格中，双面打印。

第 21—22 周记录:

复现早期论文中的越狱方法, 例如 `gptfuzzer`, 计算其成功率并分析原因;

发现多数早起论文方法已失效, 说明大语言模型自身在不断完善。

指导教师签字

杨文川

日期

2024 年 3 月 31 日

第 23—24 周记录:

在上述工作的基础上, 继续整理相关材料, 按照本科毕业论文的格式和要求, 完成最终毕业论文的编写。

指导教师签字

杨文川

日期

2024 年 4 月 14 日

注: 每 2 周指导内容记录在一个表格中, 双面打印。

第 25—26 周记录:

通过论文查重系统, 查看毕业论文的复制比以及查重报告, 并且根据查重报告, 对毕业论文进行修改, 降低毕业论文的复制比, 以达到学院以及指导老师的要求。

指导教师签字

杨文川

日期

2024 年 4 月 28 日

第 27—28 周记录:

收集并整理材料, 将任务书、开题报告、毕业论文、外文翻译、中期检查表、本文件等整理、整合并提交系统。

指导教师签字

杨文川

日期

2024 年 5 月 12 日

注: 每 2 周指导内容记录在一个表格中, 双面打印。