|  |  |  |  |
| --- | --- | --- | --- |
| **专利申请技术交底书** | | | |
| 发明人  基本  资料 | 第一  发明人  (对本发明做出最大贡献的人) | 姓 名 | 陈祺东 |
| 电 话 | 18806186120 |
| 其它联系方式 |  |
| 第二  发明人(对本发明做出次要贡献的人) | 姓 名 | 王永杰 |
| 电 话 | 18438100554 |
| 其它联系方式 |  |
| 其它  发明人(对本发明做出次要贡献的人) | 姓 名 |  |
| 电 话 |  |
| 其它联系方式 |  |
| 申请人  基本  资料 | 第一申请人 | 名 称 | 陈祺东 |
| 地 址 | 江苏省无锡市锡山区安镇街道无锡学院 |
| 邮 编 | 214000 |
| 组织机构代码 |  |
| 第二  申请人  (如果仅有一个此项不填) | 名 称 | 王永杰 |
| 地 址 | 江苏省无锡市锡山区安镇街道无锡学院 |
| 邮 编 | 214000 |
| 其它事项 |  |
| 其它  申请人(如果仅有二个此项不填) | 名 称 |  |
| 地 址 |  |
| 邮 编 |  |
| 其它事项 |  |
| 以上资料在申请专利时具有重要意义，请详细填写。 | | | |

# 一、发明/实用新型名称

一种针对RAG系统的细粒度分层多轮迭代语义优化攻击方法

# 二、技术领域

本发明涉及人工智能安全与大型语言模型（LLM）对抗攻击技术领域，具体是面向检索增强生成（Retrieval-Augmented Generation, RAG）系统的知识库安全与对抗攻击方法，特别是一种基于细粒度分层多轮迭代语义优化的知识破坏攻击方法及系统，用于在不修改LLM权重的情况下，通过向RAG系统知识库注入经过迭代优化的对抗文本，实现高隐蔽、高成功率的定向攻击，从而研究与验证RAG系统在安全防护和鲁棒性方面的潜在缺陷及对策，属于人工智能安全、自然语言处理安全、机器学习安全和信息检索安全技术领域。

# 三、现有技术（背景技术）

近年来，随着GPT-4、LLaMA、Qwen等大型语言模型在医疗问答、金融分析、法律咨询等领域被广泛应用，其响应的准确性、实时性和专业性面临挑战。检索增强生成（RAG）技术应运而生，通过结合外部知识库检索与生成模型回答用户查询以提升响应的准确性和时效性。RAG系统通常包含知识库、检索器和生成模块，其中检索器利用向量相似性（如余弦相似性、点积或近邻检索）从知识库中选取与用户查询最相关的文档片段，再由LLM生成最终回答。典型开源系统包括LangChain、LlamaIndex（GPT Index）、Haystack和HuggingFace RAG等，商用场景包括ChatGPT检索插件、Azure AI Search和Google Vertex AI Search等。

在安全层面，现有的对RAG系统的攻击主要包括提示注入攻击、贪婪坐标梯度（GCG）攻击、数据中毒攻击和知识破坏攻击。提示注入攻击通过在用户查询中嵌入恶意指令诱导LLM执行不符合安全策略的操作，但由于攻击特征明显，易被安全检测。GCG攻击通过迭代生成对抗性字符后缀操纵查询向量分布以绕过过滤器，但通常针对的是提示注入，不适合直接污染知识库。数据中毒攻击主要通过向训练集注入大量恶意或伪造内容，使模型产生错误输出，典型如“Humpty Dumpty”和“Joint-GCG”，但这类攻击往往需要访问训练数据，成本较高且周期长。

针对RAG系统的知识破坏攻击（如PoisonedRAG），攻击者通过向知识库中注入少量精心设计的恶意文本，在不修改LLM权重的情况下，使被污染文档在检索中被优先选取，从而操纵生成模块输出错误答案或攻击者期望的内容。然而，现有知识破坏攻击存在结构性缺陷：

（1）通常采用单轮生成后直接注入的策略，缺乏持续优化，导致注入文本与目标查询语义匹配度不足；

（2）依赖单一相似性指标（如余弦相似性）指导优化，未同时兼顾向量空间的方向敏感性和幅度差异，导致攻击隐蔽性不足，易被基于相似性阈值过滤的防御机制检测；

（3）缺乏对不同语义匹配区间的分层优化，未能根据检索器对不同相似度段落的敏感性，精细化调整文本结构与词汇表达，导致在语义与可检索性之间无法兼顾，影响攻击有效性与隐蔽性。

因此，现有技术在结构上未能实现注入文本的多轮语义优化和分层引导，无法在保持语义一致性的同时显著提升在RAG系统中被检索到并生成目标输出的概率，影响了攻击效果和潜伏性，这限制了对RAG系统安全性和鲁棒性研究的深入开展。

# 四、发明目的

针对现有面向检索增强生成（RAG）系统的知识破坏攻击技术存在的“仅支持单轮注入、缺乏迭代语义优化、依赖单一相似性指标、攻击隐蔽性差、无法在保持高语义一致性的同时提升被检索和被生成概率”等缺陷，本发明提出一种基于细粒度分层多轮迭代语义优化的知识破坏攻击方法与系统，以实现对RAG系统中知识库和检索模块的精准、高效、隐蔽的对抗攻击。通过本发明，能够在注入极少量精心设计的对抗文本情况下，诱导RAG系统在回答特定查询时输出攻击者预期的错误或误导性内容，同时保持注入文本与查询的高语义一致性，从而避免被向量相似度检测机制或语义偏差过滤规则检测，达到在不修改LLM参数情况下完成定向对抗攻击的目的，为RAG系统安全性评估和防御机制设计提供强有力的技术支撑。

# 五、本申请的技术方案

。

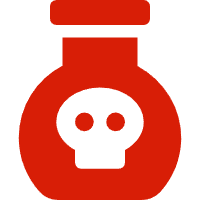
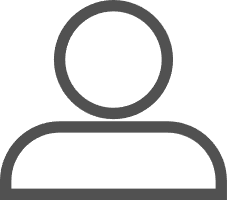
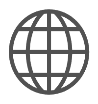
# 六、技术效果

。

# 七、提出要求保护的技术关键

。

# 八、附图



**建议的 FHM-ISO 攻击方法**

**目标问题土拨鼠日是哪一天？**

**目标答案二月二日**

**恶意文本生成器**

**恶意文本优化器**

**恶意文本：[土拨鼠日[...]3 月 15 日。**

**注入**

**外部系统**

**错误答案：3 月 15 日**

**输出**

**用户**

**维基百科**

**土拨鼠日是哪一天？**

**输入**

**收集**

**LLM**

**上下文：[土拨鼠日[...]3 月 15 日。问题土拨鼠日是哪一天？**

**提示：请根据上下文回答问题。**

**土拨鼠日[......]在每年的 2月 2 日。**

**[......]土拨鼠的**

**土拨鼠日[...]3 月 15 日。**

**RAG系统**

**寻回器**

图1 各传感器展示