|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 学生姓名 | |  | | 学号 | |  | |
| 选题学院 | | 网络空间安全学院 | | 专业班级 | |  | |
| 课题中文名称 | | 大语言模型检索增强生成中的幻觉分析与检测系统的实现 | | 课题英文名称 | | Research on Hallucination Analysis and Detection in Retrieval Augmented Generation of Large Model Retrieval | |
| 指导教师 | | 王凯龙 | | 题目类型 | |  | |
| 进行方式 | |  | | | | | |
| 学期 | 第二学期（春季） | | 起始周 | 1 | 结束周 | | 16 |
| 课题来源 | |  | | | | | |
| 目的要求 | | 近年来，随着大规模预训练模型（如GPT、BERT等）的迅速发展，检索增强生成（RAG）成为了一种有效提升生成模型性能的方法。通过结合外部检索的知识，RAG能够在一定程度上解决传统生成模型在知识丰富性和准确性上的不足。然而，这种方法在实际应用中也面临着幻觉问题，即生成的内容可能与真实世界事实不符，甚至与输入信息无关。为了解决这一问题，本课题的目标是：   1. 全面分析大模型检索增强生成过程的幻觉产生机制，深入探讨生成模型在信息检索、生成阶段中可能出现幻觉的原因及其背后的机制； 2. 结合信息论方法，研究大模型检索增强生成场景下的幻觉分布，通过量化输入和生成输出之间的依赖性，揭示幻觉生成的规律和模式； 3. 探索如何检测大模型检索增强生成中的幻觉问题，提出合理的检测方法和缓解策略，减少幻觉输出，提高生成内容的准确性和可靠性。 | | | | | |
| 主要内容 | | 1. **文献调研准备**   查阅关于大模型（LLM）和检索增强生成（RAG）的相关研究论文，了解当前该领域的最新进展。  研究与幻觉问题相关的文献，分析幻觉产生的原因、机制及已有的解决方法。  关注信息论在自然语言处理中的应用，特别是如何衡量模型生成内容与输入信息的依赖关系。   1. **研究问题分析与框架设计**   明确大模型RAG过程中幻觉产生的潜在原因，设计研究框架以全面分析幻觉的机制。  确定分析大模型输出和检索信息依赖关系的度量标准，结合信息论方法为后续研究打下基础。  制定研究方法，包括数据集选择、实验设计及评估标准。   1. **大模型检索增强生成幻觉机制分析**   深入分析RAG流程中的各环节，如输入、检索、生成等，识别可能引发幻觉的因素。  研究RAG模型生成过程中的误差传播，探索可能导致生成幻觉的关键步骤。  结合已有文献和实际案例，验证分析结果。   1. **幻觉分布的定量分析与建模**   结合信息论方法，构建度量模型依赖关系的框架，通过互信息等度量工具分析生成输出和检索信息之间的依赖性。  研究幻觉输出的分布特征，探索其规律性，为后续检测和缓解策略提供理论支持。  实现相关实验，定量分析输入与生成内容之间的相互信息、相依性。   1. **幻觉检测方法的设计与实验**   提出新的检测方法，结合多种数据分析手段，如相似度计算、语义理解等，来检测生成内容中的幻觉。  基于Copula理论、依赖建模等方法，优化检测模型，提高检测的准确性和可操作性。  开展实验，测试提出的检测方法在实际任务中的效果。   1. **结果分析与论文撰写**   总结研究成果，分析实验结果，讨论幻觉问题的本质及缓解策略的可行性。  撰写论文，详细阐述研究过程、方法、结果及结论，并根据反馈进行修改和完善。  完成论文的最终定稿，准备论文答辩。 | | | | | |
| 预期目标 | | 1. 深入分析大模型RAG过程中的幻觉产生机制,系统分析RAG过程中幻觉输出产生的内在原因，识别输入、检索和生成等环节的潜在问题。 2. 定量研究幻觉输出的分布特征,运用信息论方法（如互信息等）分析大模型输出与输入之间的依赖关系，揭示幻觉输出的规律性和分布特征。 3. 设计并验证幻觉检测方法,提出新的检测方法，结合依赖建模等技术，确保能够高效、准确地识别RAG过程中的幻觉输出。 | | | | | |
| 经费 | | 无 | | | | | |
| 参考资料 | | 1. Ji, Ziwei, et al. "Llm internal states reveal hallucination risk faced with a query." arXiv preprint arXiv:2407.03282 (2024). 2. Favero, Alessandro, et al. "Multi-modal hallucination control by visual information grounding." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. 3. Hu, Haichuan, Yuhan Sun, and Quanjun Zhang. "LRP4RAG: Detecting Hallucinations in Retrieval-Augmented Generation via Layer-wise Relevance Propagation." arXiv preprint arXiv:2408.15533 (2024). 4. Sun, Zhongxiang, et al. "ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability." arXiv preprint arXiv:2410.11414 (2024). 5. Feldman, Philip, James R. Foulds, and Shimei Pan. "Ragged edges: The double-edged sword of retrieval-augmented chatbots." arXiv preprint arXiv:2403.01193 (2024). | | | | | |
| 实验加工  条件 | | 无 | | | | | |