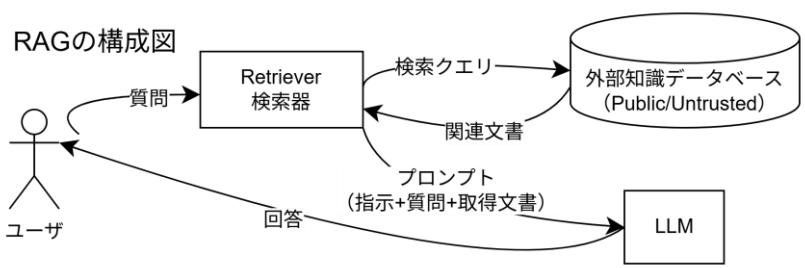
公開RAGのセキュリティに関する研究

A Study on the Security of Public RAG

宇治川ひかる・ネットワーク分科会・情報セキュリティ大学院大学

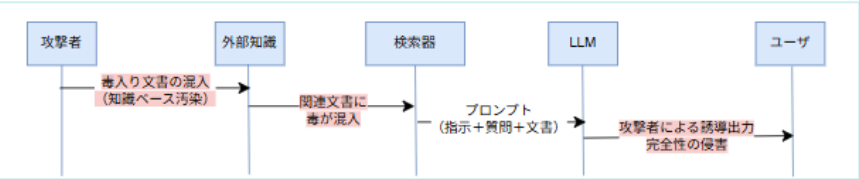
研究背景

RAG（Retrieval-Augmented Generation）とは



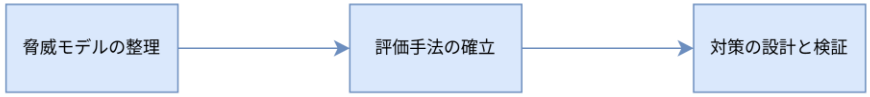
・RAGは、外部知識ベースから取得した文書を根拠としてLLMが回答を生成する枠組みである。[1]
・知識ベースが公開・非信頼コンテンツを含む場合、少量の汚染文書の混入で回答が誘導され得る。

攻撃手法: 知識ベース汚染と間接プロンプトインジェクション



・知識ベース汚染:
　攻撃者が、RAGが参照する外部知識ベースに、悪意のある情報や偽の情報を意図的に混入させる攻撃。[2]
・間接プロンプトインジェクション:
　汚染文書に含まれる指示を、RAGが取得文書としてプロンプトに含めた結果、LLMが命令として解釈して従ってしまう現象。

本研究のアプローチ手法



1. 公開RAGの攻撃面を明確化
　知識ベースが攻撃面になり得る点を脅威モデルとして整理する。
2. 影響を完全性として定量化
　プログラムで判断可能な出力制約を設け、制約に違反した回答を測定する。
3. 軽量な緩和策の効果を評価可能に
　取得文書の出所を明示し、悪意のある指示の実行を抑制する。

評価手法

制約違反率（ASR）の導入

制約違反率（Attack Success Rate: ASR）という指標を導入する。
安全か危険かといった曖昧な基準ではなく、プログラムで判断可能な明確な出力制限を設ける。攻撃を受けた際に、制約に違反した回答が生成された割合をASRとして測定する。
ASR（制約違反率）＝（制約に違反した回答数）/（全回答数）

今後の展望

評価手法を基盤とし、以下の課題に取り組む。
・設定依存性の体系的評価:
　知識ベースの汚染率、取得文書数（top-k）など、RAGの設定がASRに与える影響を調査する。
・防御アーキテクチャの検討:
　取得文書の出所の明示だけでなく、複数の防御策を検討する。

[1] P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS 2020.
[2] W. Zou et al., *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models*, arXiv:2402.07867, 2024.