검색 증강 생성 (RAG) 시스템 공격과 방어 연구 동향

김민석*, 구형준**

성균관대학교 AI시스템공학과 (대학원생*), 성균관대학교 소프트웨어학과 (교수**)

Trends in Attacks and Defenses against Retrieval-Augmented Generation (RAG) Systems

Minseok Kim*, Hyungjoon Koo**
Sungkyunkwan University

요 약

검색 증강 생성(Retrieval-Augmented Generation, RAG)은 질의에 대한 외부 지식을 검색한 결과를 대규모 언어 모델 (LLM)과 통합함으로써 인공지능 모델 성능을 향상시킬 수 있다. 그러나 RAG 시스템은 데이터베이스를 손상하는 데이터 오염 (data poisoning), 검색 과정을 조작하는 검색 오염 (retrieval poisoning), 그리고 입력 프롬프트를 변경하여 출력을 왜곡하는 프롬프트 조작 (prompt manipulation)과 같은 적대적 공격에 취약하다. 본 논문에서는 이러한 공격방식을 RAG 시스템 구성요소 기반으로 분류하고, 이를 보호하기 위한 최근 방어 전략을 살펴본다. 본 연구는 적대적 위협에 대응하여 RAG 시스템의 보안성과 신뢰성을 강화할 수 있는 방안을 제시한다.

I. 서론

검색 증강 생성 (Retrieval-Augmented Generation, RAG) [1] 은 외부 지식 검색과 생성 기능을 결합하여 기존 대규모 언어 모델 (LLM)의 한계점(예: Hallucination)을 극복함으로써 최근 들어서 활용도가 크게 증가하고 있다. LLM은 일관성 있고 맥락적으로 적절한 텍스트 생성에 뛰어나지만, 광범위한 재학습 없이 최신 정보나 전문적인 지식을 정확하게 반영하기 어렵다. 검색 증강 생성은 대규모 지식 기반에서 적절한 구절을 검색하는 검색기 [2 -5]와 이 정보를 바탕으로 정확하고 맥락에 맞는 응답을 생성하는 생성기 (LLM) [14, 15]를 결합하여 이러한 한계를 해결한다. 이를 통해 외부 지식을 폭넓게 활용하면서도 생성 모델의 언어적 능력을 유지함으로써, 검색 증강 생성 시스템은 생성된 응답의 정확성을 향상한다.

그러나 이러한 장점에도 불구하고, 검색 증강 생성 시스템의 검색과 생성 기능의 결합은 새로운 취약점을 초래하여 다양한 형태의 적대적 공격에 노출될 수 있다. 본 논문에서는 RAG 시스템 대상의 공격과 방어기법을 살펴본다.

본 논문은 2024년도 정부(과학기술정보통신부) 정보통신기획평가원 (No. 2022-0-01199, 융합보안대학원(성균관대학교); No. 2024-00337414, SW공급망 운영환경에서 역공학 한계를 넘어서는 자동화된 마이크로 보안 패치 기술 개발; No.RS-2024-00337414, 메모리 안전 언어의 적용 확대 및 안전 적용을 위한 통합 플랫폼 기술 개발)과 한국연구재단 기본연구 (바이너리 코드 문맥 추론을 위한 연속 학습이 가능한 범용 딥러닝 모델에 관한 연구; 2022R1F1A1074373) 지원으로 수행한 연구임.

II. 검색 증강 생성

그림 1과 같이 검색 증강 생성 시스템은 검색기와 생성기라는 두 가지 주요 요소로 이루어진다. 검색기는 사용자의 질의를 분석한 후, 지식 기반에서 관련 구절을 추출한다. 추출된

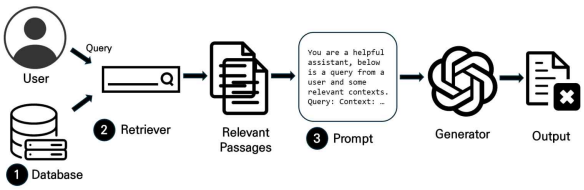

그림 1: 검색 증강 생성(RAG)의 개요와 적대적 공격에 대한 세 가지 공격 표면 (데이터베이스(❶), 검색기(❷), 프롬프트(❸))

구절은 원래의 질의와 결합하여 생성기에 대한 입력 프롬프트를 구성하며, 이를 통해 최종 응답을 생성한다.

검색기(Retriever). 검색 증강 생성 시스템은 검색 단계에서 희소 검색과 밀집 검색의 두 가지 방식을 활용한다. 희소 검색은 TF-IDF [2]와 BM25 [3] 와 같은 용어 일치 및 가중치 기반 기술을 사용하여 계산 효율성을 제공하지만, 의미적 뉘앙스를 완벽히 반영하는 데는 한계가 있다. 반면, 밀집 검색 [4, 5]은 신경망 임베딩 [16]을 활용해 질의와 구절을 고차원 벡터 공간에서 표현하여 의미적 유사성을 효과적으로 포착한다. 그러나 밀집 검색은 계산 복잡도가 높고, 대규모 데이터베이스에 적용할시 높은 계산 자원이 요구되는 단점이 있다.

생성기(Generator). 검색 증강 생성 시스템의 생성기는 검색된 구절에서 추출한 정보를 바탕으로 논리적이고 맥락적으로 적절한 응답을 생성하는 핵심 구성 요소다. 이 생성기는 검색된 구절과 사용자의 질의를 결합한 프롬프트를 입력으로 받아, 사전 학습된 매개변수와 고급 언어 이해 능력을 활용해 응답을 생성한다. 생성기는 트랜스포머 [13] 기반 아키텍처에 바탕을 두고 있으며, 맥락에 맞춘 응답을 생성하는 능력이 뛰어나다. 또한, 생성기는 단순히 명시된 정보에 그치지 않고, 암시적 의미를 추론해 확장된 해석을 제공할 수 있다.

종합적으로, 검색 증강 생성 시스템의 검색기와 생성기는 각각의 한계점을 보완하며 정확한 정보를 전달함에 있어 중요한 역할을 수행한다.

III. 검색 증강 생성 시스템 공격 표면

그림 1 및 표 1에서 제시된 바와 같이, 검색

기법	공격 표면	연도
PoisonedRAG [6]	데이터베이스(❶)	2024
GARAG [7]	데이터베이스(❶)	2024
BadRAG [8]	검색기(❷)	2024
TrojanRAG [9]	검색기(❷)	2024
GGPP [10]	프롬프트(❸)	2024

표 1. 검색 증강 생성에 대한 적대적 공격

증강 생성 시스템은 다음과 같은 세 가지 주요 공격 경로를 통해 취약점을 노출한다. 이는 첫째, 지식 기반의 무결성을 훼손하기 위해 왜곡된 정보를 주입하는 데이터 오염 공격, 둘째, 시스템 출력을 조작하기 위해 구문 검색 프로세스를 방해하는 검색 오염 공격, 그리고 마지막으로 부정확하거나 적대적인 응답을 유도하는 프롬프트 조작 공격으로 구분할 수 있다.

데이터 오염 공격(Data Poisoning Attack). 데이터 오염 공격은 지식 기반에 악의적이거나 오류가 포함된 정보를 주입하는 방식으로 이루어진다. 이를 통해 검색 증강 생성 시스템의 데이터베이스(❶)를 조작해 생성된 응답을 왜곡할 수 있다. 예를 들어, PoisonedRAG [6]는 LLM의 생성 능력에 기반하여, 외부 지식을 활용하는 검색 증강 생성 시스템의 지식 데이터베이스에 몇 개의 악의적인 텍스트를 주입함으로써 공격자가 선택한 목표 질문에 대해 공격자가 선택한 목표 답변을 생성하도록 유도한다. 또한, GARAG [7]는 단순한 오타와 같은 미세한 구절 수정을 도입하여 시스템의 취약점을 공격한다. 이러한 공격은 데이터베이스 내부에 저장된 구절에 적은 양의 악의적 수정으로도 시스템의 신뢰성에 큰 영향을 미칠 수 있음을 보여준다.

검색 오염 공격 (Retrieval Poisoning Attack). 검색 오염은 구절 검색기(❷)를 조작하여 어떤 구절이 검색되는지를 제어함으로써,

최종 응답의 정확성과 무결성에 간접적으로 영향을 미친다. BadRAG [8]는 검색 증강 생성 시스템 내부의 데이터베이스에 휴리스틱 기반의 악의적 구절을 주입하여, 정상적 질의에 대해서는 정상적으로 동작하지만, 특정 조건에서 항상 악의적인 구절을 반환하도록 한다. 또한, TrojanRAG [9]는 검색기가 악의적인 답변을 도출할 수 있도록 사전 정의한 여러 개의 트리거 집합과 목표 구절를 구성하여, 대조 학습을 통해 특정 트리거 조건에만 검색 증강 생성 시스템이 악의적인 목표 구절을 출력한다. 이러한 공격은 검색 증강 생성 시스템이 허위 정보를 확산할 가능성을 높이고, 실제 환경에서 치명적인 영향을 초래할 수 있음을 시사한다.

프롬프트 조작 공격 (Prompt Manuplation Attack). 프롬프트 조작 공격은 입력 프롬프트(❸)를 악의적으로 변경해 LLM이 부정확하거나 적대적인 응답을 생성하도록 유도하는 방식이다. Gradient Guided Prompt Perturbation (GGPP) [10] 는 프롬프트에 짧은 접두사를 삽입하여, 비관련 구절을 무시하라는 명령이 포함된 경우에도 출력이 사실에서 크게 벗어날 수 있음을 보여준다. GGPP는 검색 증강 생성 시스템의 출력을 목표로 하는 잘못된 답변으로 유도하는 최적화 기법을 도입하였으며, 이러한 접두사 삽입은 높은 성공률로 검색 증강 생성 시스템의 출력을 조작할 수 있음을 보였다. 또한, 프롬프트의 명령을 무시하도록 요청하는 지침이 포함된 경우에도 GGPP는 효과적으로 출력 변경이 가능하다. 이러한 공격은 공격자가 검색 증강 생성 시스템의 프롬프트에 접근이 가능할 시 적은 양의 수정 거리 (edit distance)로도 시스템의 결과물을 크게 변경할 수 있음을 함의한다.

IV. 검색 증강 생성의 방어 메커니즘

검색 증강 생성 시스템에 대한 위협이 증가함에 따라, 연구자들은 주로 데이터 오염 공격에 대한 방어 메커니즘을 소개했다. 표 2에서 제시된 바와 같이, RobustRAG [11] 와 Discern-and-Answer [12] 는 각각 검색 증강 생성 파이프라인의 특정 취약점을 해결하기 위해 제안된 두 가지의 접근법이다.

RobustRAG [11]는 데이터 오염 공격에 대응하기 위해 격리-집계(Isolate-then-Aggregate) 전략을 도입하였다. 이 방법은 먼저 각 검색된 구절로부터 독립적으로 LLM 응답을 생성한 후, 키워드 및 디코딩 기반 알고리즘을 사용하여 이러한 응답들을 집계한다. 이를 통해 일부 구절이 오염되더라도 전체 시스템의 응답 정확도를 유지할 수 있으며, 특정 질의에 대해 공격자가 소수의 악의적 구절을 삽입하더라도 정확한 응답을 반환할 수 있음을 보장하는 인증 가능한 강건성 (certifiable robustness)을 보장한다. Discern-and-Answer [12]는 데이터 오염으로 인한 지식 충돌을 해결하기 위해 두 가지 주요 전략을 사용한다. 첫째, 검색된 문서 간의 충돌을 식별하고 해결하기 위해 판별자 모델을 미세 조정 (fine-tuning)한다. 이 판별자 모델은 문서들 간의 일관성을 평가하여 오도된 정보를 포함한 문서를 식별하고 필터링한다. 둘째, 신중하게 설계된 프롬프트를 통해 LLM의 판별 능력을 활용하여 문서 간의 불일치를 인지하고 신뢰할 수 있는 응답을 생성하도록 유도한다. 이러한 접근법은 검색된 구절에 포함된 반사실적 노이즈를 줄여 응답의 신뢰성과 정확성을 향상하며, 다양한 학습 시나리오에서 모델의 강건성을 크게 강화한다.

기법	방어 표면	연도
RobustRAG [11]	데이터베이스(❶)	2024
Discern-and-Answer [12]	데이터베이스(❶)	2024

표 2. 검색 증강 생성에 대한 적대적 공격에 대한 방어 메커니즘

그러나 현재의 방어 전략은 검색 오염과 프롬프트 조작과 같은 다양한 위협에 대해 충분한 대응책을 제공하지 못하고 있다. 예를 들어, 검색 오염 공격은 악의적인 구절을 삽입함으로써 탐지를 회피하면서도 구절 검색 프로세스를 미묘하게 조작할 수 있는 위험을 내포하고 있으며, 이는 기존 방어 메커니즘이 설계된 데이터 오염 공격의 형태와는 다른 특성을 지니고

있다. 또한, 프롬프트 조작 공격은 사용자가 입력하는 프롬프트에 악의적인 변형을 가하여 LLM이 부정확하거나 적대적인 응답을 생성하도록 유도할 수 있는데, 이러한 공격 또한 효과적으로 대응되지 못하고 있다. 이러한 취약점은 검색 증강 생성 시스템 내에서 중요한 보안 격차를 드러내며, 공격자들이 보다 다변화된 방법으로 시스템을 공격할 수 있는 가능성을 시사한다. 따라서, 검색 오염과 프롬프트 조작과 같은 위협에 대해 보다 포괄적이고 효과적인 방어 전략을 개발하는 것이 시급하다.

V. 결론

검색 증강 생성은 외부 지식 검색을 통합함으로써 대규모 언어 모델의 성능을 한층 강화하여 사실적 정확성과 맥락적 관련성을 크게 개선하였다. 그러나 이러한 통합은 데이터 오염, 검색 오염, 프롬프트 조작과 같은 새로운 공격에 대한 취약성도 동시에 초래하였다. 이러한 보안 문제를 해결하기 위해 데이터 오염 공격에 대한 방어 방식이 일부 존재하지만, 검색 증강 생성 시스템의 신뢰성과 안정성 확보를 위해 검색 오염이나 프롬프트 조작과 같은 다양한 공격에도 대응할 수 있는 전략이 필요하다.

[참고문헌]

[1] Patrick Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Proceedings of 34th Conference on Neural Information Processing Systems, 2020.

[2] Juan Ramos et al. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. Proceedings of the First Instructional Conference on Machine Learning, 2003

[3] Stephen Robertsonet al., Simple BM25 Extension to Multiple Weighted Fields. In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. 2004.

[4] Lee Xiong et al., Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In International Conference on Learning Representations. 2021.

[5] Vladimir Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020.

[6] Wei Zou et al., PoisonedRAG: Knowledge Poisoning Attacks to Retrieval-Augmented Generation of Large Language Models. https://arxiv.org/abs/2402.07867 arXiv:2402.07867. 2024.

[7] Sukmin Cho et al., Typos that Broke the RAG's Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations. https://arxiv.org/abs/2404.13948 arXiv:2404.13948. 2024.

[8] Jiaqi Xue et al., BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models. https://arxiv.org/abs/2406.00083 arXiv:2406.00083. 2024.

[9] Pengzhou Cheng et al., TrojanRAG: Retrieval-Augmented Generation Can Be Backdoor Driver in Large Language Models. https://arxiv.org/abs/2405.13401 arXiv:2405.13401 2024.

[10] Zhibo Hu, Chen Wang, Yanfeng Shu, Helen Paik, and Liming Zhu. Prompt Perturbation in Retrieval-Augmented Generation based Large Language Models. https://arxiv.org/abs/2402.07179 arXiv:2402.07179. 2024.

[11] Chong Xiang et al., Certifiably Robust RAG against Retrieval Corruption. https://arxiv.org/abs/2405.15556 arXiv:2405.15556 2024.

[12] Giwon Hong, et al., Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise. In Findings of the Association or Computational Linguistics: NAACL 2024,

[13] Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017.

[14] Gemini Team. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. https://arxiv.org/abs/2403.05530. 2024.

[15] OpenAI. 2023. GPT-4 Technical Report. CoRR abs/2303.08774 https://doi.org/10.48550/ARXIV.2303.08774 arXiv:2303.08774. 2023.

[16] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.