검색 증강 생성 (RAG) 시스템 공격과 방어 연구 동향

김민석*, 구형준**

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

성균관대학교 SecAI Lab.

2024. 11. 29

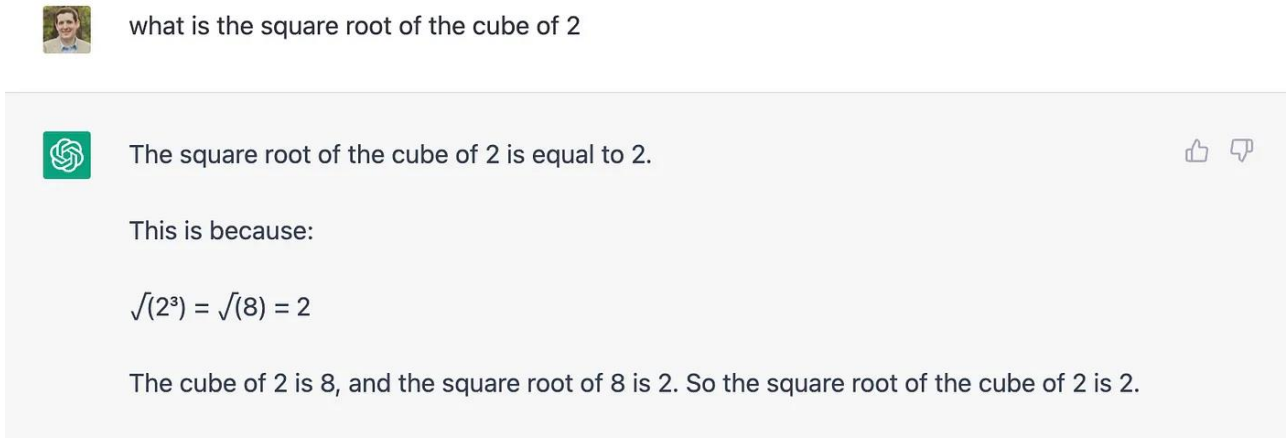
검색 증강 생성 (RAG) 시스템

✓ **검색 증강 생성 시스템** (RAG)

- **검색기**: 외부 데이터베이스에서 정보를 검색
- **생성기**: 검색된 정보를 바탕으로 질의에 맞는 정확한 응답 생성

✓ RAG 시스템의 필요성

- LLM의 한계(i.e. Hallucination) 극복
- 최신 정보 및 전문 지식을 정확히 반영 가능



https://levelup.gitconnected.com/is-chatgpt-testing-its-users-643dff339f9f

RAG 시스템의 취약점과 연구 방향

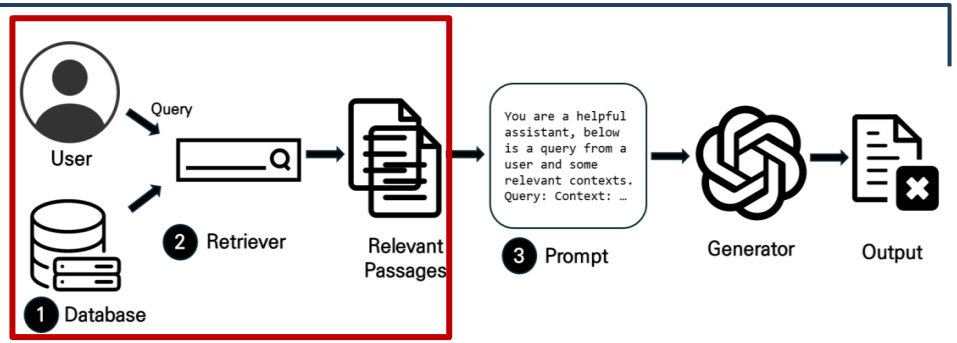
✓ RAG 시스템의 **잠재적 취약점**

- 검색과 생성 기능의 결합으로 적대적 공격에 노출 가능

✓ 연구 방향(Contribution)

- 공격 유형 분석: RAG 시스템이 직면한 다양한 적대적 공격 탐구
- 방어 기법 분석: 적대적 공격으로부터 안정성 확보를 위한 기술적 접근법 평가

검색기 (Retriever)



✓ 역할

- 사용자의 질의를 분석, 데이터베이스에서 **관련 구절을 추출**
- 희소 검색, 밀집 검색 두 가지로 분류

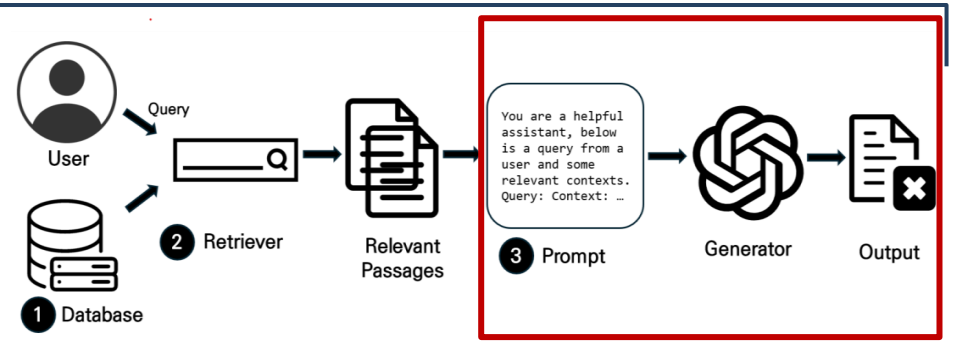
✓ 희소 검색(Sparse Retrieval)

- TF-IDF, BM25와 같은 용어 일치 및 가중치 기반 기술 활용
- 장점: 계산 효율성 높음
- 단점: 의미론적 뉘앙스 반영 한계

✓ 밀집 검색(Dense Retrieval)

- 신경망 임베딩을 활용, 질의(query)와 구절(passage)을 고차원 벡터 공간으로 변환
- 장점: 의미론적 유사성을 우수하게 포착
- 단점: 계산 복잡도 높고 대규모 데이터베이스에서 높은 자원 요구

생성기 (Generator)



✓ 역할

- **검색된 구절과 사용자 질의를 결합**하여 논리적이고 맥락적으로 적절한 최종 응답을 생성

✓ 원리

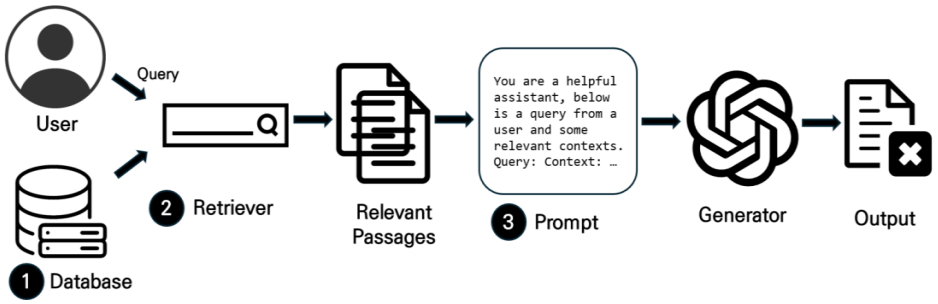
- 입력: 검색기에서 제공된 구절 + 사용자의 질의
- 기반 기술: 트랜스포머(Transformer) 기반 아키텍처
- 학습된 매개변수와 언어 이해 능력을 활용하여 응답 생성

RAG 시스템에 대한 공격 표면

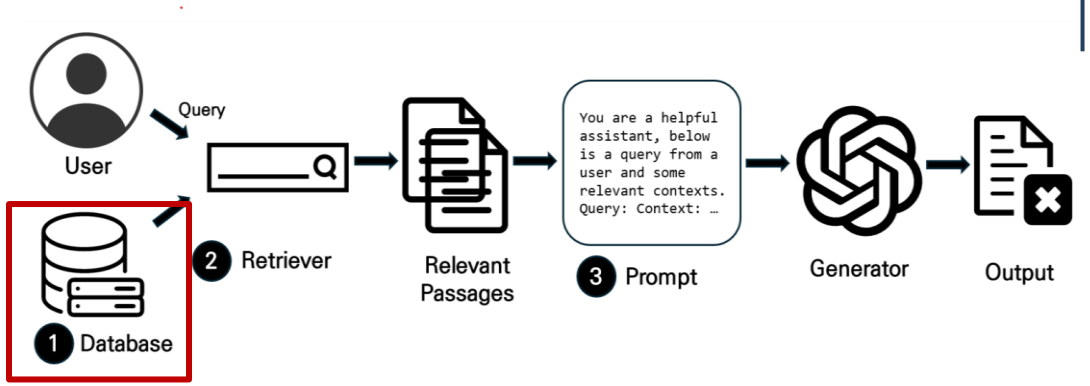
✓ 검색 증강 생성 시스템에 대한 세 가지 주요 공격 경로

- **데이터 오염 공격** (데이터베이스 ❶에 대한 공격)
- **검색 오염 공격** (검색기 ❷에 대한 공격)
- **프롬프트 조작 공격** (프롬프트 ❸에 대한 공격)

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



데이터 오염 공격



✓ 공격 방식:

- 데이터베이스에 악의적 **또는 오류 정보를 주입**하여 데이터베이스(❶)를 조작

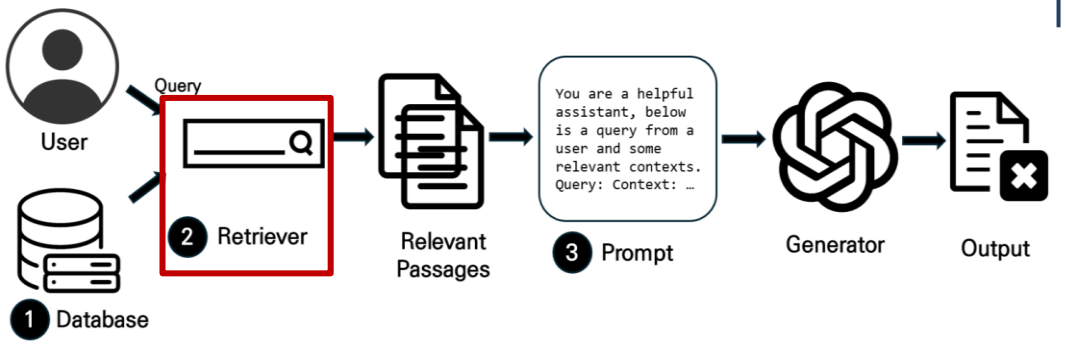
✓ 예시:

- PoisonedRAG: 악의적 텍스트를 주입 → 목표 질문에 대한 목표 답변 생성 유도
- GARAG: 적은 양의 구절(passage)을 수정 → 적은 수정만으로도 큰 영향 유발

✓ 결과:

- **왜곡된 생성 응답**으로 인한 사용자의 시스템 신뢰성 저하

검색 오염 공격



✓ 공격 방식:

- 검색기(❷) 조작 → **특정 조건에서 해당 구절이 항상 검색**되도록 설정

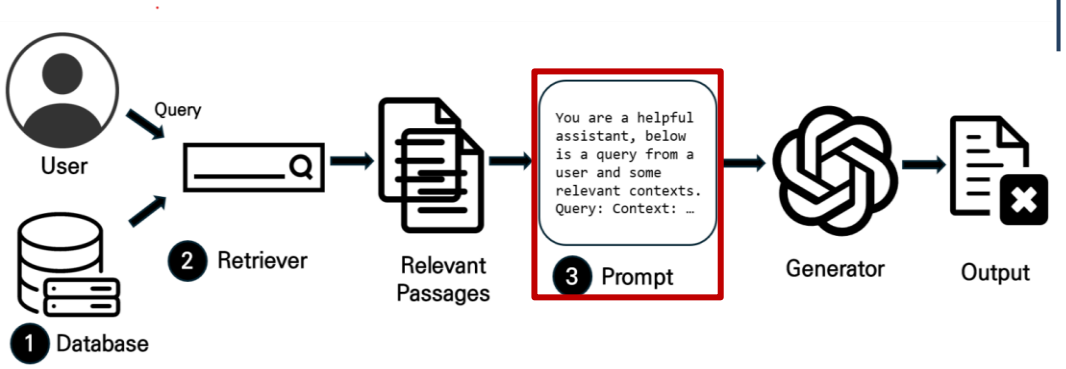
✓ 예시:

- BadRAG: 사전적으로 사용자 정의된 악의적 구절 주입 → 특정 조건에서 악의적 응답 반환
- TrojanRAG: 트리거 집합 구성 → 트리거 조건에서 악의적 목표 구절 출력

✓ 결과:

- 허위 정보 확산, 실제 환경에서 치명적 영향 초래

프롬프트 조작 공격



✓ 공격 방식:

- **프롬프트(❸)에 악의적 변경** 삽입 → 생성기 출력 왜곡

✓ 예시 (Gradient Guided Prompt Perturbation (GGPP)) :

- 짧은 접두사(prefix) 삽입 → 응답 왜곡 유도
- 비관련 구절 무시 프롬프트에도 불구, 출력 왜곡 성공

✓ 결과:

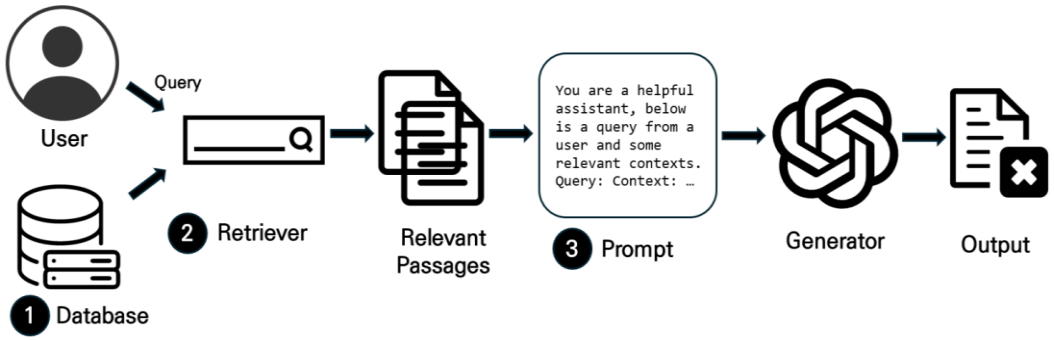
- 적은 수정 거리로도 시스템의 출력을 크게 변경 가능

검색 증강 생성의 방어 메커니즘

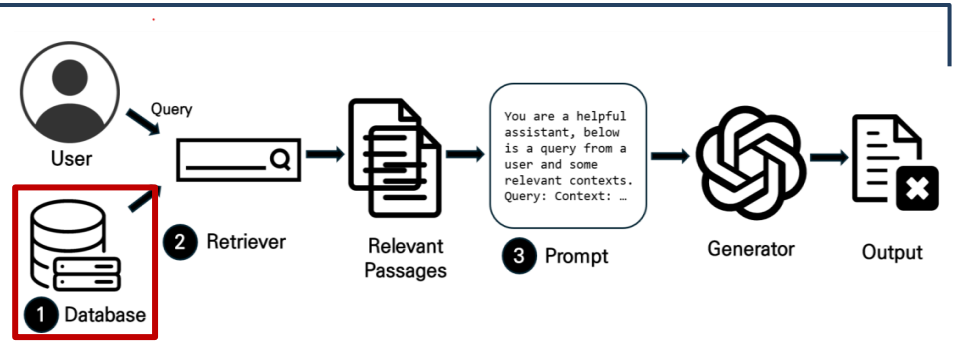
✓ 방어 메커니즘 개요

- 검색 증강 생성 시스템에 대한 공격에 대응하기 위해 다양한 방어 전략이 제안됨
- **주로 데이터 오염 공격에 초점**

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



데이터 오염 공격에 대한 방어 메커니즘



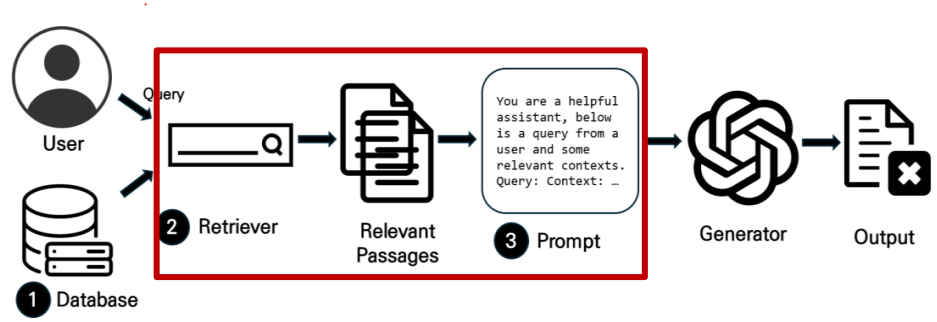
✓ RobustRAG: Isolate-then-Aggregate 전략

- 각 검색된 구절로부터 **독립적으로 LLM 응답 생성**
- 키워드 및 디코딩 기반 알고리즘으로 응답 aggregation
- 일부 구절(소수) 오염에도 전체 응답 정확도 유지 및 certifiable robustness 보장

✓ Discern-and-Answer: **Knowledge Conflict를 해결**

- 판별자(discriminator) 모델 fine-tuning → 검색된 문서 간 충돌 식별 및 해결
- 신중하게 설계된 프롬프트로 LLM에 내재된 판별 능력 활용
- 출력(output)의 반사실적 노이즈 감소 및 응답 신뢰성과 정확성 향상

현 방어 전략의 한계 및 시사점



✓ 현 방어 전략의 한계

- 검색 오염 공격: 구절 검색 프로세스를 미묘하게 조작, 기존 데이터 오염과 다른 특성을 지녀 탐지 및 대응 어려움
- 프롬프트 조작 공격: 입력 프롬프트에 악의적 변형 삽입, 부정확하거나 적대적 응답 생성 유도, 효과적인 대응 부족

✓ 시사점

- **현재 방어 전략은 데이터 오염 공격에 초점** → 검색 오염 및 프롬프트 조작에 대한 대응 부족
- 보안 격차: 공격자들이 보다 다변화된 방식으로 시스템을 공격할 가능성 증가

결론

✓ 검색 증강 생성의 장점

- 사실적 정확성 및 맥락적 관련성 향상
- 외부 지식 통합을 통한 성능 극대화

✓ 새로운 보안 위협

- 데이터 오염, 검색 오염, 프롬프트 조작 공격

✓ 현재의 방어 방식 한계

- 데이터 오염 방어에 국한된 대응
- 검색 및 프롬프트 기반 공격에 취약

✓ 향후 연구 방향

- 다각적 보안 전략 개발: 검색 오염 방지 기술 및 프롬프트 조작 방어 알고리즘 연구 필요

감사합니다

2024. 11. 29