# 확증편향 환각 방지를 위한 AQ-RAG

양진환[1], 최혁순[1], 문남미[2]
[1]호서대학교 컴퓨터공학부 석사과정
[2]호서대학교 컴퓨터공학부 교수
yjh970706@naver.com, hucksoon2001@gmail.com, nammee.moon@gmail.com

# AQ-RAG for Preventing Confirmation-Bias-Induced Hallucinations

Jin-Hwan Yang[1], Hyuk-Soon Choi[1], Nammee Moon[1]
[1]Dept. of Computer Engineering, Hoseo University

요 약

본 연구는 RAG 파이프라인의 확증편향 환각을 완화하기 위해, 원 질문으로부터 지원·반증 적대적 질의 쌍을 의도적으로 생성하고 대칭적으로 검색·평가하는 AQ-RAG를 제안한다. AQ-RAG는 LLaMA 3.1 8B를 활용해 두 질의를 생성한 뒤 FAISS + Qwen3-Embedding-0.6B 기반 MMR로 지원·반증 증거를 회수하고, LLaMA 3.1 8B 기반 신뢰도 점수를 산출해 최종 답변을 생성한다. BoolQ 데이터셋으로 진행한 실험에서 AQ-RAG가 Accuracy 84.19%를 기록하여 BERT-Large+Pretraining 기반 기존 연구 대비 1.99% 향상되었고, 소거 실험에서는 LLaMA 3.1 8B 단일 모델 62.12%, RAG 79.92%로 각각 22.07%, 4.27%의 성능 향상을 보였다. 이는 RAG가 기본 환각을 크게 줄이고, 추가적으로 적대적 질의를 적용한 AQ-RAG가 잔여 오류를 감소시켜 정확한 판단을 가능케 함을 시사한다.

## 1. 서론

LLM(Large Language Model)은 다양한 지식 집약형 과제에서 인상적인 성능을 보이지만, 그럴듯하나 사실과 다른 내용을 출력하는 환각 문제가 존재한다[1]. 이를 완화하기 위해 외부 근거를 결합하는 RAG(Retrieval Augmented Generation)를 활용하여 매개변수화된 지식과 비매개변수화 정보를 결합해 정답의 사실성을 높인다[1]. 그러나 RAG 역시 검색 실패, 증거 선택 편향 등으로 인해 환각을 완전히 제거하지 못한다[1].

최근에는 모델이 스스로 근거를 점검, 비판하도록 설계해 환각을 줄이는 시도들이 등장했다[2,3]. Self-RAG는 필요할 때만 검색하고, 생성물과 근거에 대해 자기 비평을 수행해 사실성을 개선한다[2]. 또 CoVe(Chain of Verification)은 초안 답변을 낸 뒤, 초안을 검증하기 위한 하위 질문을 계획, 독립적으로 답한 후 최종 응답을 산출함으로써 환각을 유의하게 감소시킨다[3]. 이러한 자기 점검 계열 접근은 검색과 생성의 균형을 개선하지만, 본질적으로 초안, 검증의 단일 시각에 치우치기 쉬우며, 반대 근거를 의도적으로 찾도록 강제하지는 않는다[2,3].

한편, 실제 정보 환경에서는 상충하는 증거가 공존하여 사용자의 질문 자체가 모호하다면 확증편향에 노출될 수 있다[4,5]. 최근 연구들은 LLM이 다양한 인지편향을 보일 수 있음을 체계적으로 보고했고, RAG 자체가 특정 데이터 분포나 검색 파이프라인에 의해 편향을 증폭할 위험성도 지적된다[4,5]. 즉, 가설을 지지하는 문서만 더 잘 끌어오는 검색-증강 구조는 반증 근거를 체계적으로 누락시켜 최종 출력을 왜곡할 수 있다[4,5].

따라서, 본 연구는 이러한 문제들을 해결하기 위해 AQ-RAG(Adversarial Question RAG)를 제안한다. AQ-RAG는 사용자의 질문으로부터 지원과 반증을 겨냥한 상반된 질의를 의도적으로 생성하고, 두 질의로 대칭적 검색을 수행한 뒤, 상충하는 증거 묶음의 일관성과 충돌 구조를 평가하여 최종 답변과 신뢰도를 산출하도록 설계한다. 또한, 기존 RAG와의 비교를 통해 AQ-RAG의 성능 향상을 정량적으로 분석한다.

## 2. 관련 연구

### 2.1 LLM 환각과 확증편향

LLM의 환각 문제는 학습 데이터 분포 불일치, 매개변수화된 지식의 한계, 디코딩 및 보정 과정의 문제 등 복합 요인에서 기인한다[6,7]. 이러한 취약

성을 줄이기 위해 외부 지식을 결합하는 RAG가 널리 채택되었으나, 검색-랭킹 파이프라인의 설정에 따라 특정 관점의 근거만 과도하게 노출될 위험이 남아있다[1].

이때 인간 인지에서 잘 알려진 확증편향이 대화형 검색-생성 환경에서도 재현 및 증폭될 수 있음이 최근 연구에서 관찰되었고, 질의가 제시되는 방법에 따른 응답 치우침 역시 보고되었다[8,9].

### 2.2 RAG의 기본 구조와 한계

RAG는 매개변수화된 언어모델의 생성기에 비매개변수형 외부 지식베이스를 결합하여, 질의에 맞는 문서를 검색기가 회수하고 이를 조건으로 답변을 생성하는 표준 파이프라인이다[1].

RAG는 실제 환경에서 관련 문서 회수 실패, 랭킹 및 필터링 오류, 생성 단계의 근거 미반영 등 다단계 실패 지점이 누적되면서 환각이 재발할 수 있음이 보고되었다[10,11].

## 3. AQ-RAG

AQ-RAG는 질의 $Q$에 대하여, 확증편향으로 인한 환각을 구조적으로 완화하기 위해 지원 질의 $Q^+$와 반증 질의 $Q^-$를 의도적으로 생성하고, 각 질의로 회수한 증거 $E^+$, $E^-$를 바탕으로 신뢰성 점수 $S^+ \in [0,1]$과 $S^- \in [0,1]$를 산출한다. 최종 출력 결과 $Answer$는 (수식 1)과 같이 정의한다.

$$Answer = \begin{cases} TRUE & \text{if } s^+ > s^- \\ FALSE & otherwise \end{cases} \quad \text{(수식 1)}$$

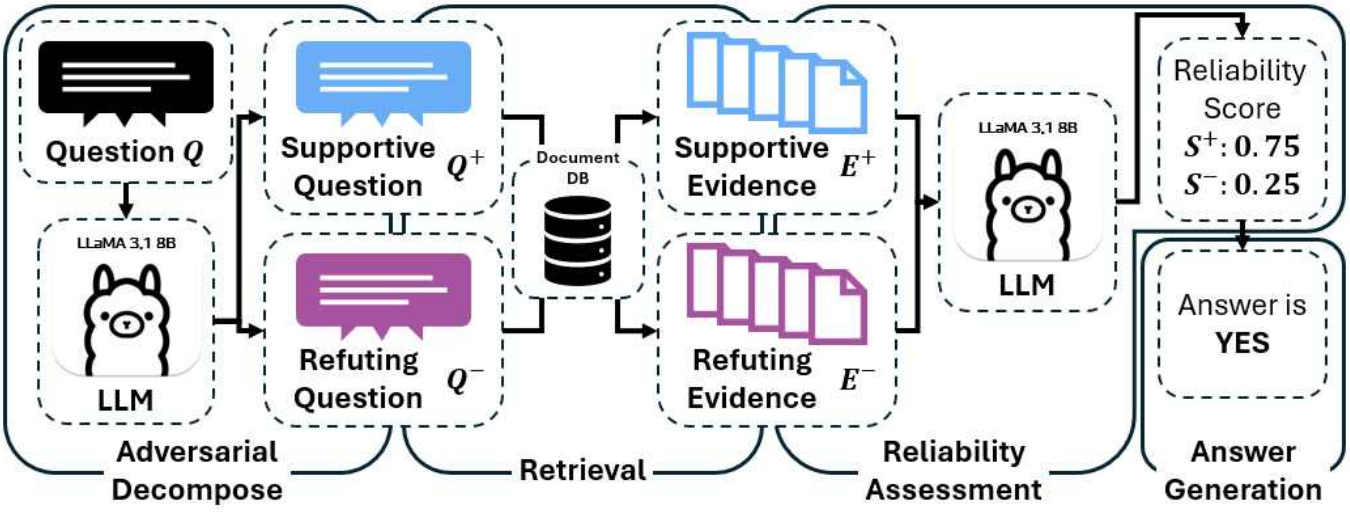
AQ-RAG의 핵심은 질의 단계에서 대립 관점을 강제하는 것이다. 이를 위해 $Q^+$는 원문 질문을 그대로 활용하되 지원 증거를 더욱 잘 회수할 수 있도록, $Q^-$는 문장 내 동사를 최소 수정으로 부정화 하도록 적대 질의 생성 프롬프트를 규정한다. 적대 질의 생성은 이를 적용한 LLaMA 3.1 8B 모델을 활용한다[12].

각 질의는 동일한 검색기로 처리되어 대칭적 회수 조건을 보장한다. 구현은 FAISS 기반 벡터스토어에 Qwen3-Embedding-0.6B 임베딩을 적용하고 MMR(Maximal Marginal Relevance) 회수를 사용한다.

다음으로 원 질의 $Q$와 지원 질의 $Q^+$, 반증 질의 $Q^-$, 각 증거 묶음인 $E^+$, $E^-$를 신뢰성 평가 프롬프트에 입력해 신뢰성 점수 $S^+$와 $S^-$를 산출한다. 신뢰성 평가 프롬프트는 $S^+$과 $S^-$가 각각 $E^+$가 $Q$를 직접 지지하는 강도와 $E^-$가 $Q$를 반증하는 강도를 기준으로 출력하도록 설정한다. 적대 질의 생성과 마찬가지로 신뢰성 점수 산출 또한 LLaMA 3.1 8B 모델을 활용한다[12].

최종 답변 생성 단계에서 산출된 $S^+$와 $S^-$를 통해 이진 답변을 생성한다. AQ-RAG는 질의 생성 단계에서 관점 대립을 강제하고, 회수, 점수화, 판정이 모두 양면 구조로 진행되며, 신뢰성 평가 점수 기반 결정을 통해 "지지 근거만 모이는 검색-그 근거만 반영되는 생성"의 확증편향 경로 차단의 효과를 기대할 수 있다.

AQ-RAG의 전체 개념도는 (그림 1)과 같다.




(그림 1) AQ-RAG 전체 개념도

## 4. 실험

### 4.1 데이터셋

본 연구는 AQ-RAG의 성능 평가를 위해 BoolQ(Boolean Questions) 데이터셋을 사용한다[12]. BoolQ는 실제 사용자 질의에서 발생한 Yes/No 질문을 대상으로, 각 샘플이 질의, 문서, 답변의 구조를 이루는 독해형 QA 데이터셋이다. 데이터 규모는 총 15,942개이며, 공식 배포 분할은 Train 9,427 / Dev 3,270 / Test 3,245로 제공된다.

본 연구는 답변 라벨이 비공개 되어있는 Test 데이터를 제외하고 Train과 Dev 데이터를 병합해 활용한다.

본 연구는 질의와 답변을 성능 평가용 데이터로 구성하고 문서를 Qwen3-Embedding-0.6B 임베딩 모델을 활용하여 FAISS 기반 벡터스토어를 생성하여 활용한다.

### 4.2 실험 환경

실험 환경의 세부사항은 아래 <표 1>과 같다.

<표 1> 실험 환경 세부사항

| Category | Specification |
|---|---|
| CPU | Intel i7-11700 |
| GPU | NVIDIA GeForce RTX 3090 |
| CUDA | 12.6 |
| Python | 3.10 |
| Langchain | 0.3.27 |
| Faiss | 1.9.0 |
| LLM | LLaMA 3.1 8B |

또한 모델의 실험 성능 비교를 위해 검색기의 하이퍼 파라미터를 고정한다. 검색기 하이퍼 파라미터 세부 사항은 <표 2>와 같다.

<표 2> 검색기 하이퍼 파라미터 세부사항

| Category | Value |
|---|---|
| k | 5 |
| fetch_k | 50 |
| lambda_mult | 0.75 |

### 4.3 실험 결과

BoolQ 데이터셋을 활용한 기존 연구들과 본 연구에서 제안한 AQ-RAG의 성능을 비교한 실험 결과표는 <표 3>과 같다.

<표 3> 실험 결과표

| Sources | Model | Accuracy |
|---|---|---|
| Clark et al., 2019[13] | BERT-Large | 78.09 |
| Clark et al., 2019[13] | BERT-Large + Pretraining | 82.20 |
| Sanagavarapu et al., 2022[14] | RoBERTa | 73.00 |
| Dimitriadis et al. 2023[15] | BERT + Weak Supervision | 77.62 |
| Dimitriadis et al. 2023[15] | RoBERTAa + Weak Supervision | 80.59 |
| w/o RAG, AQ | LLaMA 3.1 8B | 62.12 |
| w/o AQ | LLaMA 3.1 8B + RAG | 79.92 |
| **Ours** | **LLaMA 3.1 8B + AQ-RAG** | **84.19** |

실험 결과 기존 미세 조정 및 소거 실험에서 AQ-RAG가 가장 높은 성능을 보였다.

AQ-RAG는 기존 연구 가운데 가장 높은 성능을 보인 BERT-Large + Pretraining 모델에 비해 1.99% 높은 성능을 달성하였다.

소거 관점에서, LLaMA-3.1 8B 단일 모델 대비 RAG 적용 시 성능이 22.07% 상승하였다. 기존 모델이 학습되지 않은 비매개변수화 정보에 대한 답변에 환각 문제가 많이 발생하며 RAG 도입 시 답변의 사실성을 높인다는 것을 알 수 있다.

또한, AQ-RAG를 결합한 모델은 RAG 모델에 비해 4..27% 개선되었다. 이는 AQ-RAG가 기존 RAG에서 발생하는 잔여 오류를 보완하고 확증편향 환각에 대한 효과적인 방법임을 뜻한다.

## 5. 결론

본 연구는 기존 RAG 파이프라인에서 빈번히 관찰되는 확증편향 환각 문제를 구조적으로 완화하기 위해 원 질문으로부터 지원, 반증의 적대적 질의를 생성하고 대칭적 검색-증거 집합의 신뢰성 평가-최종 응답 생성으로 이어지는 AQ-RAG를 제안하였다.

BoolQ 데이터셋에서의 실험 결과 LLaMA 3.1 8B기반 AQ-RAG는 기존 연구보다 높은 정확도를 달성하였고, 동일 LLM 하에서 진행된 소거 실험에서 가장 높은 성능을 보였다. 이는 외부 근거 결합

이 대규모 언어모델의 환각을 실질적으로 감소시키며, 반증 관점의 강제 노출이 RAG가 남기는 잔여 오류를 추가로 줄인다는 점을 시사한다.

향후 연구로는 도메인 일반화 검증을 위해 BoolQ 외의 다양한 벤치마크와 도메인(사실 검증·다중 홉·웹 질의 등)에서 제안 방법론을 체계적으로 평가할 예정이다. 또한, True/False 이진 판단에 국한되지 않도록 단답형 및 장문형으로 확장하고 “지원/반대” 구조의 적대 질의를 “지원/반대/중립/대안” 등 다분기 대립 질의로 일반화하여 증거 집계·판정 모듈을 함께 고도화하는 방향을 모색하고자 한다.

## ACKNOWLEDGEMENT

본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음"(2025-0-00040)

## 참고문헌

[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks.", Advances in neural information processing systems, 33, 9459-9474, 2020

[2] Asai, Akari, et al. "Self-rag: Learning to retrieve, generate, and critique through self-reflection.", 2024.

[3] Dhuliawala, Shehzaad, et al. "Chain-of-verification reduces hallucination in large language models.", arXiv preprint arXiv:2309.11495, 2023.

[4] Malberg, Simon, et al. "A comprehensive evaluation of cognitive biases in LLMs.", arXiv preprint arXiv:2410.15413, 2024.

[5] Hu, Mengxuan, et al. "No free lunch: Retrieval-augmented generation undermines fairness in llms, even for vigilant users.", arXiv preprint arXiv:2410.07589, 2024.

[6] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.", ACM Transactions on Information Systems 43.2, 1-55, 2025.

[7] Farquhar, Sebastian, et al. "Detecting hallucinations in large language models using semantic entropy.", Nature 630.8017, 625-630, 2024.

[8] Shi, Li, et al. "Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates.", arXiv preprint arXiv:2412.04629, 2024.

[9] Li, Alice, and Luanne Sinnamon. "Examining query sentiment bias effects on search results in large language models.", The Symposium on Future Directions in Information Access (FDIA) co-located with the 2023 European Summer School on Information Retrieval (ESSIR), 2023.

[10] Barnett, Scott, et al. "Seven failure points when engineering a retrieval augmented generation system.", Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI. 2024.

[11] Zhang, Wan, and Jing Zhang. "Hallucination mitigation for retrieval-augmented large language models: a review.", Mathematics 13.5, 856, 2025.

[12] Dubey, Abhimanyu, et al. "The llama 3 herd of models.", arXiv e-prints, arXiv-2407, 2024.

[13] Clark, Christopher, et al. "Boolq: Exploring the surprising difficulty of natural yes/no questions.", arXiv preprint arXiv:1905.10044 , 2019.

[14] Sanagavarapu, Krishna, et al. "Disentangling indirect answers to yes-no questions in real conversations.", Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022.

[15] Dimitriadis, Dimitris, and Grigorios Tsoumakas. "Enhancing yes/no question answering with weak supervision via extractive question answering.", Applied Intelligence 53.22, 27560-27570, 2023.