CC-RAG: 규칙 기반 자기 신뢰도 평가와 동적 검색 정책을 통한 RAG 성능 향상

순동현, 오민석, 이상철†
대구경북과학기술원

dhsoon@dgist.ac.kr, harrymark0@dgist.ac.kr, sangchul.lee@dgist.ac.kr

CC-RAG: Enhancing Retrieval-Augmented Generation with Rule-Based Self-Assessment and Adaptive Top-k Retrieval

Donghyeon Soon, Minseok Oh, SangChul Lee†
DGIST

요 약

Retrieval-Augmented Generation(RAG)은 외부 문서를 검색하여 생성 모델의 응답에 근거를 제공함으로써 오픈 도메인 질의응답에서 사실성과 최신성을 향상시키는 대표적 접근법이다. 그러나 기존 RAG 파이프라인은 질문 난이도와 무관한 고정 top-k 검색, 검색 품질과 무관한 답변 생성, 그리고 생성 결과의 신뢰도 추정 부재로 인해 환각(hallucination) 및 무응답 문제가 발생할 수 있다. 본 연구는 이러한 한계를 완화하기 위해, 추가 학습이나 별도의 critic 모델 없이 추론 단계에서만 동작하는 경량 제어 기반 RAG 인 CC-RAG(Confidence-Controlled RAG)를 제안한다. CC-RAG 는 (1) 생성 토큰의 평균 로그확률, 답변 길이, 정보성 토큰 포함 여부, 문서-답변 간 어휘 중복도 등을 결합한 규칙 기반 자기 신뢰도 평가로 낮은 신뢰도의 응답을 탐지하고, (2) 질문 난이도 및 초기 검색 점수에 따라 검색 깊이 k 를 조절하는 동적 검색 정책을 통해 문맥 효율성과 정보 커버리지를 동시에 확보한다. NQ-Open 벤치마크에서 CC-RAG 는 prompt 기반의 Naive RAG(16.58%) 및 Chat-Template 기반의 Baseline RAG(18.91%) 대비 정확도를 향상시켜 22.01%를 달성하였으며, ablation 분석을 통해 자기평가와 동적 검색이 상호 보완적으로 기여함을 확인하였다. 본 연구는 학습 비용 없이도 RAG 의 신뢰성과 성능을 개선할 수 있는 실용적인 추론 제어 전략을 제시한다.

Ⅰ. 서 론

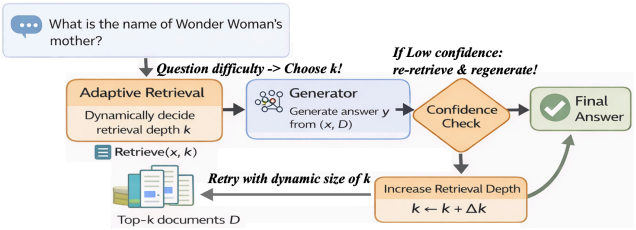
대규모 언어 모델(LLM)은 오픈 도메인 QA 에서 뛰어난 생성 능력을 보이지만, 사전학습에 내재된 파라메트릭 지식에 의존해 최신 정보가 반영되지 않거나 외부 지식이 필요한 질문에서 환각(hallucination)이 발생할 수 있다. 이는 실제 서비스 환경에서 답변의 사실성과 신뢰성을 저해한다. Retrieval-Augmented Generation(RAG)은 외부 문서를 검색해 생성의 근거로 활용함으로써 이러한 한계를 완화하지만, 일반적인 rule-based RAG 는 (1) 질문 난이도와 무관한 고정 top-k 검색, (2) 부정확한 검색 결과에도 답변 생성을 강제, (3) 생성 결과의 신뢰도 추정 부재로 환각을 사전에 억제하기 어렵다는 문제가 있다. 특히 소형 모델에서는 검색 품질에 대한 민감도가 더 크다. 본 연구는 이를 개선하기 위해 CC-RAG(Confidence-Controlled RAG)를 제안한다. CC-RAG 는 추론 단계에서 규칙 기반 신뢰도 점수를 산출해 낮은 신뢰도의 경우 재검색·재생성을 수행하고, 질문 난이도에 따라 검색 깊이 k 를 동적으로 조절한다. 이를 통해 불필요한 문맥 증가를 줄이면서도 정보 누락을 완화해 RAG 의 정확도와 신뢰성을 향상시키는 것을 목표로 한다.

Ⅱ. 본론

2.1 시스템 구성 및 배경

본 연구는 두 단계로 구성된다. 먼저 Transformer 기반 GPT-small 언어 모델을 구현하고 사전학습 및 다운스트림 태스크를 통해 기본 생성 성능을 검증한다. 이후 BM25 검색기와 instruction-tuned LLM을 결합한 프롬프트 기반 RAG를 구축하고 이를 CC-RAG로 확장한다. GPT-small과 Baseline RAG를 직접 구현한 이유는, 프롬프트 구성·검색 깊이·재시도 로직 등 개별 요소의 영향을 분리하여 CC-RAG의 추론 단계 제어 효과를 재현 가능하게 검증하기 위함이다.

Figure 1.CC-RAG 모델의 파이프라인



기본 RAG는 검색된 문서를 프롬프트에 삽입한 뒤 답변을 생성하는 retrieve-then-generate 구조를 따른다. 본 연구에서는 검색 문서를 질문 앞에 단순 연결하는 flat prompt 방식을 Naive RAG로 정의하며, 이는 고정 k 검색과 신뢰도 추정 부재로 인해 검색 품질이 낮을 경우 무응답이나 환각이 발생하기 쉽다. 이를 개선하기 위해, LLaMA의 chat-template과 <docs> 블록을 적용한 Baseline RAG를 구축하고, 여기에 신뢰도 기반 재시도와 동적 검색 깊이 조절을 추가한 CC-RAG를 제안한다. 세 방법의 차이는 프롬프트 구조와 추론 단계 제어 유무로 요약되며, 성능 비교는 **표 1**에 제시한다.

표 1. NQ-Open 에서 RAG 변형 모델들의 정확도 비교

Model	Prompt Format	Accuracy (%)
Naive RAG	Flat prompt	16.58
Baseline RAG	Chat-template	18.91
CC-RAG (*ours*)	Chat-template	**22.01**

2.2 제안기법: CC-RAG(Confidence-Controlled RAG)

본 절에서는 CC-RAG의 핵심 구성요소를 소개한다. 먼저 2.2.1절에서는 추가 학습 없이 계산 가능한 휴리스틱

표 2. Baseline RAG 와 CC-RAG(제안 방법)의 정성적 비교

Question	Baseline RAG	CC-RAG (Ours)	Ground Truth
Who sings "Does He Love Me" with Reba?	*(No answer generated)*	**Linda Davis**	Linda Davis
Where do the Great Lakes meet the ocean?	*(No answer generated)*	**the Saint Lawrence River**	the Saint Lawrence River
Who played Jason in Friday the 13th Part 1?	Tommy Jarvis	**Ari Lehman**	Ari Lehman
Who played Draco Malfoy in the Harry Potter movies?	Tom Felton	Tom Felton	Thomas Andrew Felton
What is the name of Wonder Woman's mother?	Queen Hippolyta	**Queen Hippolyta**	Queen Hippolyta

표 3. Ablation 실험 결과.

Model	Acc (%) ↑	R-1 ↑	R-2 ↑	R-L ↑	R-L_{sum} ↑
Naive RAG	16.60	0.117	0.061	0.116	0.116
Baseline RAG	18.91	0.153	0.098	0.152	0.152
w/o Self-Assessment	21.88	0.254	0.145	0.254	0.254
w/o Dynamic Retrieval	21.79	0.254	**0.146**	0.253	0.254
CC-RAG (Ours)	**22.01**	**0.256**	**0.146**	**0.255**	**0.255**

신호를 결합해 생성 답변의 신뢰도를 추정하고, 신뢰도가 낮을 경우 재검색·재생성을 수행하는 메커니즘을 설명한다. 이어서 2.2.2절에서는 질문 난이도 및 초기 검색 품질에 따라 검색 깊이 k를 동적으로 조절하는 정책을 제시하여, 단순 질의의 불필요한 문맥을 줄이면서도 어려운 질의의 정보 누락을 완화하는 방법을 다룬다.

2.2.1 규칙 기반 자기 신뢰도 평가(Self-Assessment)와 재시도

Self-RAG[1]는 생성 결과를 비평(critic)하여 낮은 신뢰도의 답변에 대해 재검색·재생성을 수행하는 루프 구조를 갖는다. 본 연구는 별도의 학습된 critic 모델 없이 규칙 기반 점수 산출로 대체하여 계산 비용을 크게 줄인다. 답변 y와 문서 집합 D가 주어졌을 때, CC-RAG는 신뢰도 점수 $s_{conf} \in [0,1]$를 다음과 같이 정의한다.

$$s_{\text{conf}}(x) = \sigma\Big(w_1 \cdot \overline{\log p_\theta(y \mid x, D)} + w_2 \cdot f_{\text{len}}(y) + w_3 \cdot f_{\text{NE}}(y) + w_4 \cdot f_{\text{overlap}}(y, D)\Big)$$

여기서 $\overline{\log p_\theta(y \mid x, D)}$는 생성 토큰 평균 로그확률, f_{len}은 답변 길이 휴리스틱, f_{NE}는 숫자/고유명사 등 정보성 토큰 가중, $f_{overlap}$은 답변과 검색 문서 간 어휘 중복(lexical overlap)을 의미한다. 신뢰도 점수가 임계값 τ보다 낮을 경우(τ=0.8), 다음과 같이 재검색 및 재생성을 수행한다.

$$\text{if } s_{\text{conf}}(x) < \tau \Rightarrow k \leftarrow k + \Delta k,\ D \leftarrow \text{Retrieve}(x, k),\ y \leftarrow \text{Generate}(x, D)$$

이 재시도 메커니즘은 검색 근거가 불충분하거나 잘못된 경우 발생하는 환각을 완화하며, 정성 비교에서 무응답/오답이 정답으로 회복되는 사례를 확인할 수 있다(**표 2** 참조).

2.2.2 질문 난이도 기반 동적 검색 깊이(Dynamic Retrieval Depth)

기존 RAG는 모든 질문에 대해 동일한 k개의 문서를 검색하는데, 이는 쉬운 질문에 대해 불필요한 문맥을 증가시키거나 어려운 질문에서 정보 누락을 초래한다. CC-RAG는 질문 길이 및 모호성, 초기 BM25 점수 등을 기반으로 난이도를 추정하여 검색 깊이 k를 동적으로 조절한다. 예를 들어 비교적 단순한 질문에는 k=2 수준의 소량 검색을 적용해 효율성을 확보하고, 정보 요구가 큰 질문에서는 k를 최대 10까지 증가시켜 커버리지를 높인다. 또한 전체 문맥은 예산(예: 3000 토큰) 내에서 절단하여 과도한 토큰 사용을 방지한다.

2.3 실험결과

본 연구는 NQ-Open에서 Accuracy와 ROUGE 계열 지표로 성능을 평가하고, Naive RAG(flat prompt), Baseline RAG(chat-template), CC-RAG를 비교하였다. 표 1의 정량 결과, CC-RAG는 추가 학습 없이 추론 단계 제어만으로 정확도를 개선했다(16.58% → 18.91% → 22.01%).

표 3의 정성 결과에서는 Baseline RAG가 무응답을 출력하거나 잘못된 개체를 생성한 경우, CC-RAG가 재검색을 통해 정답을 회복하는 경향을 보였다. 또한 ablation(표 3)에서 자기 신뢰도 평가와 동적 검색 깊이는 각각 단독 적용 시에도 성능을 높였고, 결합했을 때 최고 성능을 기록해 두 모듈이 상호 보완적으로 작동함을 확인했다.

Ⅲ. 결론

본 연구는 기존의 RAG 파이프라인의 한계인 고정 검색 정책과 신뢰도 추정 부재로 인한 환각 문제를 완화하기 위해, 규칙 기반 자기 신뢰도 평가와 동적 검색 깊이 조절을 결합한 CC-RAG(Confidence-Controlled RAG)를 제안하였다. CC-RAG는 별도의 critic 모델이나 추가 학습 없이 추론 단계에서만 동작하며, 낮은 신뢰도의 답변에 대해 재검색·재생성을 수행함으로써 답변의 정확도와 신뢰성을 향상시킨다. 실험 결과 CC-RAG는 NQ-Open에서 Naive/Baseline RAG 대비 정확도 및 ROUGE 성능을 유의미하게 개선하였고, ablation 분석을 통해 자기평가와 동적 검색이 각각 성능 향상에 기여하며 결합 시 최적의 성능을 보임을 확인하였다. 또한 구현 관점에서 추가 학습 비용이 없고 추론 오버헤드가 매우 작아, 지연 시간이나 계산 자원이 제한된 환경에서도 적용 가능성이 높다. 다만 본 접근법은 검색기가 지속적으로 무관 문서를 상위에 랭크하는 경우 성능 한계가 존재한다. 향후 연구에서는 검색 불확실성 추정, 재랭킹 및 다중 홉 질의응답(multi-hop QA) 확장을 통해 한계를 보완할 수 있으며, 도메인 특화 환경으로의 적용 가능성도 추가 검증이 필요하다.

ACKNOWLEDGMENT

본 연구성과는 2025 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2025-090085).

참 고 문 헌

[1] Asai, A., et al. (2023). Self-RAG: Learning to critique answers for improved retrieval-augmented generation, EMNLP, 2023.