The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.9, pp.429~436
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.9.429

A Survey on the Latest Research Trends in Retrieval-Augmented Generation

Eunbin Lee† · Ho Bae††

ABSTRACT

As Large Language Models (LLMs) continue to advance, effectively harnessing their potential has become increasingly important. LLMs, trained on vast datasets, are capable of generating text across a wide range of topics, making them useful in applications such as content creation, machine translation, and chatbots. However, they often face challenges in generalization due to gaps in specific or specialized knowledge, and updating these models with the latest information post-training remains a significant hurdle. To address these issues, Retrieval-Augmented Generation (RAG) models have been introduced. These models enhance response generation by retrieving information from continuously updated external databases, thereby reducing the hallucination phenomenon often seen in LLMs while improving efficiency and accuracy. This paper presents the foundational architecture of RAG, reviews recent research trends aimed at enhancing the retrieval capabilities of LLMs through RAG, and discusses evaluation techniques. Additionally, it explores performance optimization and real-world applications of RAG in various industries. Through this analysis, the paper aims to propose future research directions for the continued development of RAG models.

Keywords : LLM, Retrieval-Augmented Generation, Hallucination

검색 증강 생성(RAG) 기술의 최신 연구 동향에 대한 조사

이 은 빈† · 배 호††

요 약

Large Language Model(LLM)의 급격한 발전은 자연어 처리 분야에 혁신을 불러 일으켜 이를 적절하게 활용하는 것이 중요한 주제로 떠오르고 있다. 방대한 데이터로 훈련된 LLM은 다양한 주제에 대한 텍스트 생성이 가능하여 콘텐츠 생성, 기계 번역, 챗봇 등 여러 방식으로 적용이 가능하나 특정 유형이나 전문적 지식이 부족할 수 있어 일반화하기 어렵다는 단점이 존재한다. 모델 훈련이 완료된 이후의 최신 정보로 즉각 업데이트되기도 어려우며, 모델이 실제로 존재하지 않는 정보나 오류에 대해 그럴 듯하게 답변하는 환각 현상(Hallucination) 역시 주요 문제점이다. 이를 극복하기 위해 지속적으로 업데이트되는 최신 정보를 포함한 외부 데이터베이스에서 정보를 검색해 응답을 생성하는 Retrieval-Augmented Generation(RAG, 검색 증강 생성) 모델을 도입하여 LLM의 환각 현상을 최소화하고 효율성과 정확성을 향상하기 위한 연구가 활발히 이루어지고 있다. 본 논문에서는 RAG의 기본 아키텍처를 소개하고, LLM에 RAG를 적용하기 위한 연구 및 최적화의 최신 동향을 분석한다. RAG를 평가하기 위한 다양한 기법들을 소개하고, 실제 산업에서 RAG를 활용하기 위해 성능을 최적화하거나 응용한 사례들을 분석한다. 이를 바탕으로 향후 RAG 모델이 발전할 수 있는 연구 방향성을 제시하고자 한다.

키워드 : 대형 언어 모델, 검색 증강 생성, 환각 현상

1. 서 론

최근 Large Language Model(LLM)의 급격한 발전은 자연어 처리 분야에 혁신을 불러일으켰으며, 실제 산업에서 이를 적절하게 활용하는 것이 중요한 주제로 떠오르고 있다. LLM은 방대한 데이터로 훈련되어 다양한 주제에 대한 텍스트 생성이 가능하다는 장점이 있어 콘텐츠 생성, 기계 번역, 챗봇과 같은 형식으로 산업에 활용되고 있다. 하지만 대규모 데이터셋으로 사전 학습되었음에도 전문적인 지식이나 특정 유형에 대한 정보가 부족할 수 있어 도메인 특화된 작업을 처리하는 데 한계가 있다[1]. 모델의 훈련이 완료된 이후에는 새로운 데이터나 최신 정보로 즉각 업데이트되기 어렵다. 모델이 실제로 존재하지 않는 정보나 오류에 대해 그럴듯하게 답변하는 환각 현상(Hallucination) 역시 주요 문제점으로, 실제 산업에서 LLM의 응답을 신뢰하며 사용하기 어렵게 만드는 요소이다.

이러한 문제를 극복하기 위해 제안된 기술이 Retrieval-Augmented Generation(RAG, 검색 증강 생성)[2]이다. RAG는 지속적으로 업데이트되는 최신 정보를 포함한 외부 데이터베이스에서 정보를 검색하여 LLM의 응답을 생성하는 기술이

※ 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2021-II212068, 인공지능 혁신 허브 연구 개발).
※ 이 논문은 2024년 ACK 2024의 일반논문으로 "검색 증강 생성(RAG) 기술에 대한 최신 연구 동향"의 제목으로 발표된 논문을 확장한 것임.
† 비 회 원 : 이화여자대학교 인공지능융합전공 석사과정
†† 종신회원 : 이화여자대학교 사이버보안학과 교수
Manuscript Received : July 11, 2024
Accepted : August 9, 2024
* Corresponding Author : Ho Bae(hobae@ewha.ac.kr)

다. 모델이 학습 과정에서 얻은 정보를 내부 파라미터에 저장하는 기존의 매개변수 메모리(Parametric Memory)를 사용하던 기존의 방식과 달리 학습이 끝난 이후에도 외부 메모리 소스를 사용하여 지속적으로 정보를 저장하고 검색할 수 있는 비매개변수 메모리(Non-parametric Memory)를 사용하는 방식이다. 이를 통해 LLM은 최신 정보를 신속하고 정확하게 제공할 수 있으며, 외부 데이터베이스에 전문 지식을 추가하면 해당 분야에 특화된 응답을 얻을 수 있다.

RAG는 LLM의 환각 현상을 최소화할 수 있는 효과적인 방법으로 여겨지며, 특히 장문의 문장에서 맥락을 처리할 때 기존의 방법론인 파인튜닝(Fine-Tuning)[3]이나 문맥 창(Contextual Window)[4]보다 더 우수한 성능을 보인다. 파인튜닝을 통해 모델을 업데이트하면 학습에 필요한 데이터를 준비하거나 학습하기 위한 시간과 비용이 많이 소모되고 모델의 범용성이 제한될 수 있다. 반면, RAG를 적용할 경우 LLM 모델과 별개로 외부 데이터베이스만 실시간으로 업데이트하면 되기 때문에 효율성이 높고 응답 품질이 향상된다는 측면에서 더욱 탁월하다. Ovadia, Oded, et al.[5]은 다양한 주제에 걸쳐 파인튜닝과 RAG의 성능을 비교하고 평가한 결과, RAG가 기존 지식과 새로운 지식을 모두 다루는 데 있어 일관되게 더 나은 성능을 발휘함을 발견했다.

본 논문의 섹션 2에서는 RAG의 기본 아키텍처를 소개하고, 섹션 3에서는 LLM의 검색 기능을 강화하기 위한 RAG의 연구들을 소개한다. 섹션 4에서는 RAG를 최적화하기 위한 연구들의 최신 동향을 분석하며, 섹션 5에서는 RAG를 평가하기 위한 다양한 기법들을 소개한다. 섹션 6에서는 실제 산업에서 RAG를 활용하기 위해 성능을 최적화하거나 응용한 사례들을 분석한다. 이를 바탕으로 섹션 7에서는 향후 RAG 모델이 발전할 수 있는 연구 방향성을 제안하고자 한다.

2. RAG의 구조

RAG의 핵심 아이디어는 지속적으로 업데이트되는 최신 정보를 포함한 외부 데이터베이스를 LLM과 분리하여 두고, 사용자로부터 입력 쿼리를 받을 때마다 외부 데이터베이스로부터 정보를 검색해 이를 기반으로 LLM의 응답을 생성하는 것이다. Fig 1은 RAG 모델이 작동하는 전체적인 흐름을 보여준다. 전통적인 RAG 모델의 아키텍처는 크게 Indexing, Retrieval, Generation의 3가지 컴포넌트로 구성된다[6].

2.1 Indexing

Indexing이란 텍스트 형식으로 변환할 수 있는 다양한 유형의 원시 데이터를 검색이 가능하도록 구조화하는 단계를 의미한다. 원시 데이터를 청크(Chunk) 라는 작은 단위 조각으로 분할한 후 임베딩(Embedding) 과정을 통해 텍스트 데이터를 벡터(Vector) 형식으로 전환한다. 텍스트의 의미와 정보를 담은 벡터를 벡터 데이터베이스(Vector Database)에 저장하여 N차원 공간에서 유사한 의미를 가진 다른 텍스트 벡터들이

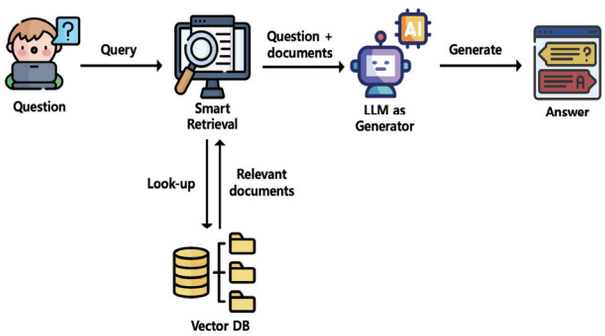

Fig. 1. Architecture of RAG

가까이 위치하게 된다. 벡터 데이터베이스는 쿼리 벡터가 주어졌을 때, 그에 대하여 가장 가까운 이웃을 효율적이고 빠르게 검색할 수 있게 한다.

2.2 Retrieval

Retrieval이란 사용자로부터 입력 쿼리가 들어오면 적합한 문서를 벡터 데이터베이스로부터 검색하는 과정이다. Indexing에 사용된 것과 동일한 임베딩 모델을 활용하여 사용자의 입력 쿼리를 벡터로 변환한다. 변환된 벡터와 관련성이 높은 정보를 가지고 오기 위해 벡터 데이터베이스에서 유사도를 계산하고, 그중 가장 높은 유사도를 가진 상위 K개의 문서를 선택한다.

정보를 검색하는 방식은 크게 2가지로 구분된다. Sparse Vector를 기반으로 일치하는 단어를 고려하여 검색하는 Sparse Retrieval의 대표적인 방법론은 BM25(Best Match)[7]가 있다. 하지만 이 방법은 검색 대상 키워드와 문서에 포함된 단어가 일치하지 않을 경우 검색이 되지 않는 문제가 있다. 이를 극복하기 위해 텍스트를 의미 벡터로 표현하는 Dense Vector를 기반으로 검색하는 Dense Retrieval이 제안되었으며, 대표적인 방법론은 DPR(Dense Passage Retrieval)[8]이 있다.

2.3 Generator

Retrieval에서 검색된 문서와 입력 쿼리를 전달받은 LLM이 텍스트 형식의 응답을 생성하는 단계이다. LLM 모델 내부에 파라미터화된 지식과 Retrieval을 통해 얻은 정보를 통합하여 사실적이면서도 맥락에 적절한 응답을 생성할 수 있다. Generation 모델은 BART[9], FiD[10], Llama[11], GPT-4[12] 등이 사용될 수 있다.

3. LLM 검색 성능 향상을 위한 RAG

사용자로부터 쿼리를 입력받을 때, 사전학습된 LLM에 포함되어 있지 않은 외부 지식을 참고해야 하는 경우가 있다. 모델의 관점에서 사전학습되지 않은 정보는 파라미터화되지 않았으므로 외부 지식으로 간주하는데, 이처럼 외부 지식이 필요한 문제들을 지식 집약적(Knowledge-Intensive) 문제라고 한다. LLM이 지식집약적 문제를 처리하는데 취약하기 때문에 이를 해결하기 위해 별도로 외부 지식으로부터 정보를 참조하

여 보완하기 위한 시도가 Lewis, Patrick, et al.[2], Guu, Kelvin, et al.[13]에서 이루어졌으며, Chen et al[14]은 Wikipedia와 같은 외부 문서로부터 정보를 검색하는 연구를 진행하였다. 이처럼 기존의 Language Model에 Retrieval 개념을 적용하여 모델이 학습한 내부 지식뿐 아니라 외부 지식을 추가로 활용하는 모델을 Retrieval- Augmented Language Models(RALM)라고 부르며, 이는 RAG 연구의 초기 흐름과 연결된다.

Language Model의 성능을 향상시키기 위해 Retrieval을 적용한 대표적 연구는 REPLUG 프레임워크를 제안한 Shi, Weijia, et al.[15]으로, 모델 내부의 아키텍처를 변경하지 않고도 검색된 문서를 Generator LLM의 입력에 추가하여 검색 구성 요소와 언어 모델 간의 시너지를 향상시켰다. Yu, Wenhao, et al.[16]는 외부 지식 데이터베이스에 관련 없는 정보가 포함되어 있을 때에도 RALM 모델의 강건성을 향상시키기 위해 검색된 문서에 대해 순차적인 리딩 노트(Reading Note)를 생성하는 Chain-of-Note(CoN) 방법을 제안한다. 리딩 노트를 통해 사용자 쿼리와 관련성이 낮거나 신뢰할 수 없는 정보를 필터링하여 응답의 정확성을 높일 뿐 아니라 문맥적으로도 일관성 있는 응답을 가능케 한다.

환각 현상을 줄이기 위해 모델 자체를 개선한 방법론으로 Jiang, Zhengbao, et al.[17]은 동적 검색 시스템인 FLARE (Forward-Looking Active REtrieval augmented generation)을 제안하였다. 이 방법은 미래의 요구사항을 예측하여 검색 쿼리를 생성하고, LLM의 출력을 직접 검색 쿼리로 사용하여 관련 정보를 가져오는 방식이다. 특히 장문의 텍스트를 생성할 때, 모델의 유연성과 정확성을 크게 향상시켜 전반적인 환각 현상을 감소시켰다. 그러나 생성과 검색을 번갈아 진행하는 과정에서 오버헤드와 생성 비용이 증가하고, 검색할 때마다 LLM이 여러 번 활성화되어야 한다는 한계가 있다.

이후 Language Model의 응답 품질과 사실적 무결성을 높이기 위해 Asai, Akari, et al.[18]은 검색, 생성 및 자기 비평(Self-Critique) 메커니즘을 통합한 Self-RAG을 제안하였다. Self-RAG는 검색한 내용을 비평할 수 있는 리플렉션 토큰(Reflection Token)을 사용하여 검색된 정보와 생성된 응답의 관련성과 유용성을 스스로 평가한다. 이 방법은 원래 모델의 창의성과 다양성을 유지하면서도 응답의 품질과 사실성을 향상시키는 모델이 자가 조절할 수 있는 메커니즘이라는 장점이 있다. 그러나 생성된 결과물 중 일부 내용이 인용된 자료와 완전히 일치하지 않거나 인용된 자료로 완전히 뒷받침되지 않는 정보를 포함할 수 있다는 한계가 있다.

모델이 스스로 평가한 내용을 학습에 반영하는 자가 평가의 개념을 결합한 연구는 Wang, Yile, et al.[19]에서도 수행되었다. 이들은 모델이 스스로 알고 있는 정보와 모르는 정보를 평가하고 지식 공백에 대해서만 외부 정보를 검색하도록 하는 Self-Knowledge Guided Retrieval Augmentation(SKR) 방법을 제안하였다. 이 방법은 직접적인 질문과 문맥 학습을 통해 필요한 경우에만 검색을 수행함으로써, 모델의 효율성을 높이고 정확성과 관련성을 향상시켰다.

이 외에도 기존의 인공지능 방법론에 RAG 메커니즘을 결합하여 모델의 성능을 향상시키기 위한 여러 연구가 진행되었다. Zhang, Jianyi, et al.[20]은 Language Model의 Retrieval을 지식 증류(Knowledge Distillation) 메커니즘과 통합한 ReAugKD 모델을 제안하였다. 이는 교사 모델의 소프트 라벨과 예측값을 포함하는 지식 베이스(Knowledge Base) 형태로 구성된 비매개변수 메모리에 학생 모델이 접근하여 지식 베이스에서 효과적으로 정보를 검색할 수 있도록 한다. Yao, Shunyu, et al.[21]은 Chain-of-Thought(CoT)와 Retrieval을 결합하여 내부 지식만 사용하던 기존의 CoT와 달리 외부 지식을 계속 검색하여 추론하는 프로세스인 ReACT를 제안하였다. 이를 통해 CoT의 주요 문제점이었던 빈번한 환각 현상을 감소시키고, 추론 작업에서 우수한 성능을 보였다.

그러나 이와 같은 다양한 시도에도 여전히 외부 정보에서 정보를 검색할 때 관련성이 낮거나 정확하지 않은 정보를 검색할 가능성이 있고, 맥락에 적절하지 않은 응답을 생성할 수 있는 단점이 존재한다. 검색된 정보 자체에 대한 신뢰성이 보장되지 않거나 불필요한 데이터에 지나치게 의존할 경우 외부 정보에서 지식을 검색하는 것이 오히려 생성 품질을 저해할 수 있기 때문이다. 이러한 문제를 극복하기 위해 다양한 최적화 연구가 진행되고 있다.

4. RAG 최적화

RAG 최적화를 위한 연구는 크게 3가지 측면에서 전개된다. 검색된 문서의 사실적 정확성 여부(정확성), 사용자의 쿼리와 관련된 문서의 검색 여부(관련성), 검색 과정의 효율성 달성(효율성)이 주요 목표이며, Table 1에서 각 목표와 그에 따른 방법론을 확인할 수 있다. 본 절에서는 각 측면의 연구를 소개하고 특징을 분석한다.

4.1 정확성

검색된 문서가 사실적으로 정확하고, 오류나 잘못된 정보가 포함되지 않도록 하는 것이 목표이다. 이는 문서의 내용이 사실에 기반하고 있으며, 정보를 제공하고 있는 출처가 신뢰할 수 있는지 확인하는 것을 포함한다.

Table 1. Optimization for RAG

	Goal	Methods
Correctness	The retrieved documents are factually accurate and include no errors or incorrect information	[22-23]
Context Relevence	Search for documents related to the query and provide appropriate and useful answers to question	[24-25]
Efficiency	Efficienty in search speed and resource utilization, along with optimization of system performance	[26-27]

복잡하거나 추상적인 작업을 처리할 때 LLM에 불필요한 데이터를 입력하면 정확한 응답을 출력하는 성능이 저하될 수 있다. 특히 Shi, Freda, et al.[22]은 수리 추론이나 논리적 판단처럼 높은 정밀도가 요구되는 작업에서 모델이 불필요한 정보를 필터링하고 주어진 과제에 집중할 수 있는지 검증하고 평가했다. [22]은 프롬프트에 불필요한 정보를 무시하도록 명시적 내용의 지시문을 추가하여 모델의 강건성을 유지하는 방법을 제안했다. 그러나 이러한 명시적 지시가 모든 상황에 적용되지는 않으며, 모델 자체의 변경 없이 추가적인 프롬프트에만 의존해야 하므로 모델 자체의 개선이 필요하다는 단점이 있다.

이 외에도 Asai, Akari, et al.[23]는 관련성이 있는 증거 기반의 정보를 평가하고 선택하는 과정을 생성 프로세스에 추가하여 RAG 모델의 정확성을 높이는 방법을 제안했다. 이는 단편적인 출력을 생성할 때 증거에 기반한 응답을 생성하여 정확도를 향상시키는 데 효과적이지만 전체적인 문맥에서는 여전히 관련성이 낮은 응답을 생성할 수 있다는 한계가 있다.

4.2 관련성

사용자의 쿼리와 관련된 문서가 검색되었는지 평가하고, 검색된 문서가 사용자 질문의 정보 요청이나 문맥과 밀접하게 연관되는 것을 목표로 한다. 사용자가 찾고자 하는 정보에 대해 적절하고 유용한 답변을 제공할 수 있어야 한다.

Wang, Zhiruo, et al.[24]는 타당성 추론, 어휘 중첩, 교차 상호 정보를 포함한 세 가지 필터링 전략을 사용하여 유용한 문맥과 그렇지 않은 문맥을 효과적으로 구분하는 방법을 제안했다. 그러나 이러한 필터링 기술은 매우 복잡하거나 추상적인 주제에 대한 검색 결과에서는 개선이 미미하다는 한계점이 있으므로 일반화하기 어렵다는 특징이 있다.

Ma, Xinbei, et al.[25]은 검색 쿼리의 관점에서 효과적인 검색을 위해 사용자의 입력을 기반으로 검색 쿼리를 재작성하여 활용하는 Rewrite-Retrieve-Read 방식을 제안하였다. 이를 통해 사용자의 입력과 실제로 필요한 지식 정보 사이에 발생할 수 있는 간극을 최소화하여 의도에 맞는 답변을 검색할 수 있도록 하였다.

4.3 효율성

검색 속도와 자원 활용 측면에서 효율성을 달성하는 것을 의미한다. 검색 효율성이 높은 모델은 적은 자원으로 빠르게 관련된 정보를 검색할 수 있으며, 동시에 시스템의 성능을 최적화하는 것을 목표로 한다.

RAG 기반 LLM의 효율성을 향상시키기 위해 He, Zhenyu, et al.[26]는 처리 속도에 중점을 둔 REST(Retrieval -Based Speculative Decoding) 모델을 제안했다. 이 모델은 코퍼스에서 구축된 문맥-계속성(Context-Continuation) 쌍이 포함된 외부 데이터베이스를 활용하여 주어진 문맥에서 연결될 가능성이 높은 토큰을 문맥-계속성 쌍으로부터 검색함으로써 전체적으로 쿼리 시 발생하는 추론 과정을 간소화하였다.

Jin, Chao, et al.[27]는 하드웨어 측면에서 검색 효율성을 달성하기 위한 연구를 진행하였다. 검색된 정보의 중간 상태를 Knowledge Tree로 구성하고, GPU와 Host Memory 계층에 캐싱하는 RAGCache를 제안하여 중복 계산을 감소시켰으며, LLM 애플리케이션을 위한 최신 캐싱 시스템과 비교하였을 때도 최대 1.8배 높은 처리량을 달성하였다.

5. RAG 성능 평가

5.1 평가 측면

RAG 프레임워크에서 생성된 응답의 품질을 평가할 때 크게 2가지 관점에서 이루어진다[28-30].

1) 충실도

모델이 응답을 생성할 때 검색한 문맥을 충실하게 반영하여 생성되었는지 평가한다. 검색한 정보가 올바르게 사용되었으며 신뢰할 수 있는지를 평가하며, 응답을 생성하는 과정에서 정보가 왜곡되거나 잘못된 해석이 없는지 점검한다.

2) 관련성

모델이 생성한 응답이 사용자의 질문과 직접적인 관련이 있는지 평가하여 모델이 질문에 적절하게 답변하고 있는지 확인한다. 생성된 응답이 전체적인 맥락에 맞춰 일관성을 유지하고 있는지를 확인하여 모델이 사용자의 요구를 제대로 이해하고 있는지 평가할 수 있는 기준이 된다.

5.2 평가 과제

RAG의 성능을 평가할 수 있는 다운스트림 작업은 크게 질의응답(QA, Question Answering), 대화문(Dialog), 정보 추출(IE, Information Extraction), 추론(Reasoning)으로 구분할 수 있다. Table 2은 RAG 모델의 다운스트림 작업 전체 개요를 보여준다.

질의응답 작업의 대표적인 데이터셋으로는 Natural Questions(NQ)[31], PopQA[32], HotpotQA[33], ELI5[34], TriviaQA(TQA)[35] 등이 있다. 이 외에도 대화문 작업은 Wizard of Wikipedia(WoW)[36], KBP[37], 정보 추출 작업은 WikiEvent

Table 2. Evaluation Tasks of RAG

Tasks	Datasets
QA	Natural Question(QA) [31] PopQA [32] HotpotQA [33] ELI5 [34] TriviaQA(TQA) [35]
Dialog	Wizard of Wikipedia(WoW) [36] KBP [37]
Information Extraction	WikiEvent [38] RAMS [49]
Reasoning	HellaSwang [40] CoT Reasoning [41]

[38], RAMS[39], 추론 작업은 HellaSwag[40], CoT Reasoning [41] 등이 있다.

5.3 평가 지표

RAG의 다운스트림 작업을 평가할 때, 각 작업에 적합한 기존의 지표를 사용한다. 질의응답 작업에서 모델이 정확하게 응답을 제공하는지 평가하기 위해 모델의 응답이 정답과 완전히 일치하는지 측정하는 EM(Exact Match)와 부분적인 일치를 평가하는 F1 지표를 사용한다[25, 41-43]. 사실성에 초점을 맞출 때는 Accuracy를 사용한다[2, 23, 41, 44]. 응답의 유창성을 평가할 때는 MAUVE 지표를 사용하며[18, 30], 응답의 품질을 평가할 때는 BLEU(Bilingual Evaluation Understudy Score) 및 ROUGE(Recall-Oriented Understudy for Gisting Evaluation), ROUGE-L 지표를 사용한다[17-18].

5.4 평가 프레임워크

RAG 기반 시스템을 평가하기 위해 Saad-Falcon, Jon, et al.[28]은 표준화된 방법론인 ARES(Automated RAG Evaluation System)을 제안하였다. 이 방법은 주어진 테스트 세트 안에서 응답을 자동으로 평가하고, 성능 메트릭을 계산하는 방식으로, 정확도, 문서 유사도, 문서 재구성 등 다양한 메트릭을 통해 RAG 시스템의 성능을 평가한다. Es, Shahul, et al.[29]의 RAGAS는 세 가지 메트릭을 사용하여 RAG 시스템을 평가한다. 충실도, 답변 관련성, 문맥 관련성 등의 측면에 대하여 인간의 판단을 포함한 WikiEval 데이터셋을 도입하였고, 이를 평가에 활용하였다. Hoshi, Yasuto, et al.[45]은 사용자가 직접 Retrieval 기반 LLM을 개발하고 성능 측정 및 평가까지 가능하게 하는 프레임워크를 제공하여 편의성을 제공하고 있다. 이 외에도 RGB, CRUD와 같은 벤치마크 또한 RAG의 성능을 평가할 수 있다[46-47].

Gao, Tianyu, et al.[30]는 모델의 환각을 줄이기 위해 거대 텍스트 말뭉치로부터 정보를 검색하여 생성된 텍스트에 인용을 추가하고, 생성된 인용의 정확도나 생성 품질 등을 체계적으로 평가하기 위한 벤치마크 ALICE를 도입하였다. 이를 통해 사실적 정확성과 검증 가능성을 개선하였다.

이처럼 전체적인 개발 및 평가를 가능케 하는 프레임워크 연구는 이루어져 있지만 특정 도메인이나 언어에 특화된 평가를 위한 연구는 부족한 실정이다. 향후 다층적인 관점에서 RAG를 평가할 수 있는 기법들에 대한 연구가 이루어져야 할 것이다.

6. RAG의 응용 분야

RAG 모델은 빠르게 변화하는 최신 정보를 즉각 반영할 수 있으며, 외부 데이터베이스에서 실시간으로 검색한 정보를 기반으로 응답을 생성할 수 있다. 이와 같은 장점을 활용하여 전문적인 지식이 필요한 분야나 산업에 특화시키기 위한 응용 연구들이 제안되고 있다.

Lozano, Alejandro, et al.[48]은 임상 질문에 답할 수 있도록 의료과학 문헌을 활용한 RAG와 오픈소스 웹 애플리케이션을 통합한 시스템을 제안했다. 이 시스템은 RAG를 적용하여 의료 분야 연구자들이 최신 연구 결과를 효율적으로 파악하고 의료 조언의 정확성을 높일 수 있도록 함으로써 RAG의 응용 분야를 확장하였다. 그러나 잠재적으로 RAG에서 잘못된 정보나 오해의 소지가 있는 정보를 생성할 가능성이 있어 이를 맹목적으로 신뢰하는 것은 위험할 수 있다는 한계가 있다.

Kang, Haoqiang, et al.[49]은 금융 분야에서 발생할 수 있는 환각 현상의 범위와 조건을 식별하였다. 이를 통해 전문 분야에서 RAG를 사용할 때 발생할 수 있는 환각 현상을 이해할 수 있는 실증적 증거를 제공하였다.

이처럼 금융뿐 아니라 의료, 자율주행, 법률 등 다양한 전문 분야에 특화해 발생할 수 있는 환각 현상의 유형을 파악하고 완화하기 위한 연구가 필요하다. Jin, Jiajie, et al.[50]는 RAG 연구를 보다 용이하게 구현하고 평가할 수 있는 표준화된 프레임워크를 제안하여 연구자들의 요구사항을 효율적으로 충족시키고 있다.

7. 향후 발전 방향

최근 생성형 AI에서는 텍스트뿐 아니라 이미지, 오디오 등 여러 타입의 멀티모달 데이터를 통합하여 활용할 수 있는 LMM(Large Multimodal Models)이 등장하였다. 다양한 모달리티의 데이터를 처리할 수 있는 확장된 언어 모델의 중요성이 강조되고 있으므로 이를 처리할 수 있는 RAG의 연구도 필요하다[51].

RAG를 사용하더라도 여전히 관련성이 낮거나 부정확한 정보를 얻을 수 있는 문제가 존재한다. 따라서 중요한 결정을 할 때 RAG의 응답을 전적으로 신뢰하는 것은 위험할 수 있다. Yan, Shi-Qi, et al.[52]은 검색 프로세스 중 검색된 문서의 신뢰도 점수 평가하고 이를 생성 과정에 반영하여 RAG 모델의 강건성을 강화하는 모델을 제안하였다. 이와 같이 RAG 모델의 성능을 향상시키고, 실용성을 높일 수 있는 신뢰도 평가 및 개선 등의 연구가 추가로 필요하다.

검색과 생성을 번갈아 수행하는 과정에서 발생하는 오버헤드와 높은 생성 비용 역시 해결해야 할 과제이다. 이를 극복하기 위해 정보를 더욱 효율적으로 저장하고 검색할 수 있도록 개선이 필요하다. RAG 모델의 소프트웨어 최적화뿐 아니라 하드웨어 최적화를 통해 처리 능력을 향상시키는 연구가 필요하다[53].

이 외에도 LLM 모델 자체에 존재하는 성별, 지역, 인종 등 여러 편향 문제를 RAG를 통해 해결할 수 있다. Shrestha, Robik, et al.[54]는 텍스트에서 이미지로 생성하는 Diffusion Model에서 종종 발생하는 사회적 편견을 해결하기 위해 외부 이미지 데이터베이스에서 참조 이미지를 검색하여 이를 생성 모델에 조건을 거는 방식으로 편향을 완화하기 위한 전략을 제안하였다. 이처럼 RAG는 모델의 편향 문제를 해결하기 위한 수단으로 유용하게 활용될 수 있을 것이다. 반면, RAG 모

델을 사용하는 것이 오히려 편향 문제를 증폭시킬 수 있다. 명백한 지표에 기반하여 편향을 줄이기 위한 노력이 오히려 쉽게 정량화하기 어려운 미묘한 편향을 간과할 수 있기 때문이다. RAG 모델이 특정 문화, 지역 등 일부 특성만 포함하고 있을 경우, 다양한 관점을 제공하지 못해 반향실 효과(Echo Chamber) 현상을 초래할 수 있다[55]. 따라서 RAG 모델의 잠재적인 위험을 인지하여 모델의 성능을 다각적으로 평가하고, 편향을 완화할 수 있는 후속 연구를 진행하는 것이 필요하다.

8. 결 론

RAG의 등장은 전통적인 LLM에서 발생할 수 있는 여러 한계를 해결할 수 있는 중요한 열쇠가 되었다. RAG를 사용하면 외부 데이터베이스로부터 정보를 검색하고 이를 기반으로 응답을 생성하는 과정을 통해 사용자의 요구에 적절한 정확한 응답을 제공할 수 있다. LLM의 성능을 향상시키기 위해 정확하고 관련성 높은 응답을 생성할 수 있는 RAG 연구가 활발히 진행되고 있다.

본 논문은 RAG 기술의 기본 개요를 소개하였으며, LLM과 결합하여 성능을 향상시키기 위한 다양한 연구의 특징들을 분석하였다. 최적화 연구의 속성을 크게 정확성, 관련성, 효율성으로 구분하고 속성별 발전된 기술과 RAG를 평가할 수 있는 다양한 평가 지표들을 살펴보았다. 다양한 도메인에 특화하거나 실제 산업에 응용하는 연구는 아직 초기 단계에 불과하며 추가적인 연구가 필요하다.

RAG는 LLM이 가지고 있는 잠재력을 끌어낼 수 있는 도구로서 앞으로 생성형 AI 분야에 혁신을 불러일으킬 수 있는 기술이다. RAG의 장점을 극대화하고, 취약점을 극복할 수 있도록 지속적인 이론적 연구 및 응용 분야 적용이 필요하다.

References

[1] M. Alex, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, “When not to trust language models: Investigating effectiveness of parametric and non-parametric memories,” *in Proceedings of the ArXiv Preprint*, arXiv: 2212.10511, 2022.

[2] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems 33*, Online, pp.9459-9474, 2020.

[3] K. Tian, E. Mitchell, H. Yao, C. D. Manning, and C. Finn, “Fine-tuning language models for factuality,” *in Proceedings of the ArXiv Preprint*, arXiv:2311.08401, 2023.

[4] S. Chen, S. Wong, L. Chen, and Y. Tian, “Extending context window of large language models via positional interpolation,” *in Proceedings of the ArXiv Preprint*, arXiv:2306.15595, 2023.

[5] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, “Fine-tuning or retrieval? comparing knowledge injection in llms,” *in Proceedings of the ArXiv Preprint*, arXiv:2312.05934, 2023.

[6] Y. Gao et al., “Retrieval-augmented generation for large language models: A survey,” *in Proceedings of the ArXiv Preprint,* 2023, arXiv:2312.10997.

[7] S. Robertson, and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” *Foundations and Trends® in Information Retrieval*, Vol.3. No.4, pp.333-389, 2009.

[8] V. Karpukhin et al., “Dense passage retrieval for open-domain question answering,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, pp.6769-6781, 2020.

[9] M. Lewis et al., “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp.7871-7880, 2020.

[10] G. Izacard, and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume,* Online, pp.874-880, 2021.

[11] H. Touvron et al., “Llama: Open and efficient foundation language models,” *in Proceedings of the ArXiv Preprint*, arXiv:2302.13971, 2023.

[12] J. Achiam et al., “Gpt-4 technical report,” *in Proceedings of the ArXiv Preprint*, arXiv:2303.08774, 2023.

[13] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” In *Proceedings of the 37th International Conference on Machine Learning*, Online, pp.3929-3938, 2020.

[14] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading wikipedia to answer open-domain questions,” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, pp.1870-1879, 2017.

[15] W. Shi et al., “REPLUG: Retrieval-Augmented Black-Box Language Models,” In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, pp.8371-8384, 2024.

[16] W. Yu, H. Zhang, X. Pan, K. Ma, H. Wang, and D. Yu, “Chain-of-note: Enhancing robustness in retrieval-augmented language models,” *in Proceedings of the ArXiv Preprint*, arXiv:2311.09210, 2023.

[17] Z. Jiang et al., “Active retrieval augmented generation,” In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp.7969-

7992, 2023.

[18] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," *in Proceedings of the ArXiv Preprint*, arXiv:2310.11511, 2023.

[19] Y. Wang, P. Li, M. Sun, and Y. Liu, "Self-knowledge guided retrieval augmentation for large language models," *In Findings of the Association for Computational Linguistics: EMNLP*, Singapore, pp.10303-10315, 2023.

[20] J. Zhang et al., "ReAugKD: Retrieval-Augmented Knowledge Distillation For Pre-trained Language Models," In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, pp.1128-1136, 2023.

[21] S. Yao et al., "React: Synergizing reasoning and acting in language models," *in Proceedings of the ArXiv Preprint*, arXiv:2210.03629, 2022.

[22] F. Shi et al., "Large language models can be easily distracted by irrelevant context," in *Proceedings of the 40th International Conference on Machine Learning*, Hawai, pp.31210-31227, 2023.

[23] A. Asai, M. Gardner, and H. Hajishirzi, "Evidentiality-guided Generation for Knowledge-Intensive NLP Tasks," *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, pp.2226-2243, 2022.

[24] Z. Wang, J. Araki, Z. Jiang, M. R. Parves, and G. Neubig, "Learning to filter context for retrieval-augmented generation," *in Proceedings of the ArXiv Preprint*, arXiv: 2311.08377, 2023.

[25] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query Rewriting in Retrieval-Augmented Large Language Models," *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp.5303-5315, 2023.

[26] Z. He, Z. Zhong, T. Cai, J. D. Lee, and D. He, "Rest: Retrieval-based speculative decoding," *in Proceedings of the ArXiv Preprint*, arXiv:2311.08252, 2023.

[27] C. Jin et al., "RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation," *in Proceedings of the ArXiv Preprint*, arXiv:2404.12457, 2024.

[28] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," *In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, pp.338-354, 2024.

[29] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *in Proceedings of the ArXiv Preprint*, arXiv:2309.15217, 2023.

[30] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp.6465-6488, 2023.

[31] T. Kwiatkowski et al., "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, pp.452-466, 2019.

[32] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories," *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, pp.9802-9822, 2023.

[33] Z. Yang et al., "HotpotQA: A dataset for diverse, explainable multi-hop question answering," *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, pp.2369-2380, 2018.

[34] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "ELI5: Long form question answering," *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, pp.3558-3567, 2019.

[35] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, pp.1601-1611, 2017.

[36] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *in Proceedings of the ArXiv Preprint,* 2018, arXiv:1811.01241.

[37] H. Wang et al., "Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues," *In Findings of the Association for Computational Linguistics*, Singapore, pp.9556-9569, 2023.

[38] S. Li, H. Ji, and J. Han, "Document-Level Event Argument Extraction by Conditional Generation," I*n Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, pp.894-908, 2021.

[39] S. Ebner, P. Xia, R. Culkin, K. Rawlins, and B. V. Durme, "Multi-Sentence Argument Linking," *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp.8057-8077, 2020.

[40] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "HellaSwag: Can a Machine Really Finish Your Sentence?,"

In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, pp.4791-4800, 2019.

[41] S. Kim et al., "The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning," *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp.12685-12708, 2023.

[42] B. Wang et al., "InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining," *in Proceedings of the ArXiv Preprint*, arXiv:2310.07713, 2023.

[43] Z. Feng, X. Feng, D. Zhao, M. Yang, and B. Qin, "Retrieval-Generation Synergy Augmented Large Language Models," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, pp.11661-11665, 2024.

[44] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy," *In Findings of the Association for Computational Linguistics*, Singapore, pp.9248-9274, 2023.

[45] Y. Hoshi et al., "RaLLe: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models," *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* Singapore, pp.52-69, 2023.

[46] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," *In Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, pp.17754-17762, 2024.

[47] Y. Lyu et al., "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models," *in Proceedings of the ArXiv Preprint*, arXiv:2401.17043, 2024.

[48] A. Lozano, S. L. Fleming, C. Chiang, and N. Shah, "Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature," *In Pacific Symposium On Biocomputing*, pp.8-23, 2024.

[49] H. Kang and X. Liu, "Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination," *in Proceedings of the ArXiv Preprint*, arXiv:2311.15548, 2023.

[50] J. Jin, Y. Zhu, X. Yang, C. Zhang, and Z. Dou, "FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research," *in Proceedings of the ArXiv Preprint*, arXiv:2405.13576, 2024.

[51] W. Chen, H. Hu, X. Chen, P. Verga, and W. Cohen, "MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text," *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, pp.5558-5570, 2022.

[52] S. Yan, J. Gu, Y. Zhu, Z. Ling, "Corrective retrieval augmented generation," *in Proceedings of the ArXiv Preprint*, arXiv:2401.15884, 2024.

[53] Z. Wang, S. Teo, J. Ouyang, Y. Xu, and W. Shi, "M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions," *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, pp.1966-1978, 2024.

[54] R. Shrestha, Y. Zou, Q. Chen, Z. Li, Y. Xie, and S. Deng, "FairRAG: Fair Human Generation via Fair Retrieval Augmentation," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* Seattle, pp.11996-12005, 2024.

[55] M. Narayan, J. Pasmore, E. Sampaio, V. Raghavan, and G. Waters, "Bias Neutralization Framework: Measuring Fairness in Large Language Models with Bias Intelligence Quotient (BiQ)," *in Proceedings of the ArXiv Preprint*, arXiv:2404.18276, 2024.

[56] E. Lee, and H. Bae, "A Survey on Retrieval-Augmented Generation", *Proceedings of the Annual Symposium of Korea Information Processing Society Conference (KIPS)* 2024, Vol.31-1, pp.745-748, 2024.

이 은 빈

https://orcid.org/0009-0004-9319-4648
e-mail : eunbinlee@ewha.ac.kr
2023년 이화여자대학교 사이버보안전공 (학사)
2023년 ~ 현 재 이화여자대학교 인공지능융합전공 석사과정
관심분야 : LLM, Retrieval Augmented Generation, Synthetic Data

배 호

https://orcid.org/0000-0002-5238-3547
e-mail : hobae@ewha.ac.kr
2007년 런던대학교(UCL) 컴퓨터과학(학사)
2009년 런던대학교(UCL) 정보보안(석사)
2021년 서울대학교 자연과학대학(박사)
2021년 ~ 현 재 이화여자대학교 사이버보안학과 교수
관심분야 : AI Security and Privacy, Synthetic Data