RAG의 내재적 한계 분석 및 고도화 방향성 연구

정세연, 김건민, 김경백
전남대학교 인공지능융합학과
320eyeonsm@gmail.com, geonminkim@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

A Study on the Intrinsic Limitations and Advancement Directions of RAG

Seyeon Jeong, Geonmin Kim, Kyungbaek Kim
Dept. of AI Convergence, Chonnam National Univ.

요 약

거대 언어 모델(Large Language Models, LLMs)은 광범위한 분야에서 인간 수준의 텍스트 생성을 선보였으나, 사실과 다른 정보를 생성하는 환각(Hallucination) 문제는 여전히 신뢰성 확보의 주요 걸림돌로 남아있다. 이러한 문제를 완화하기 위해 외부 지식 소스를 참조하여 응답의 사실적 일관성을 높이는 RAG(Retrieval-Augmented Generation)가 핵심 기술로 부상했다. 하지만 RAG 시스템 역시 정보 검색, 컨텍스트 통합, 최종 생성 단계에 걸쳐 내재적인 한계를 명확히 드러내고 있다. 본 논문은 RAG 파이프라인을 Retreival, Integration, Generation의 세 단계로 나누어 각 단계에서 발생하는 핵심적 한계를 분석한다. 구체적으로 Lost in the Middle 현상과 같은 Retreival 실패, 상충되는 정보의 비효율적 통합, Generation 단계의 환각 문제를 지적한다. 이를 바탕을 Query 최적화, Hybrid Search, Re-Ranking, Self-Correction 등 각 한계를 극복하기 위한 최신 연구 동향을 종합하여 RAG 시스템의 고도화를 위한 구체적인 방향성을 제시한다.

Ⅰ. 서 론

최근 Transformer 아키텍처를 기반으로 한 거대 언어 모델(LLMs)은 자연어 처리(NLP) 분야에서 전례 없는 발전을 이끌었다[1-2]. LLM은 방대한 텍스트 데이터로부터 학습한 파라미터 내 지식(Parametric Knowledge)을 활용하여 질의응답, 요약, 번역 등 다양한 태스크에서 뛰어난 성능을 보였다. 하지만 이러한 모델들은 학습 데이터에 존재하지 않는 최신 정보에 취약하며, 사실과 어긋난 정보를 그럴듯하게 생성하는 환각 문제를 야기한다. 이는 금융, 의료, 법률 등 정확성이 필수적인 도메인에서 LLM의 실용적 적용을 저해하는 치명적인 단점으로 작용한다. 이러한 한계를 극복하기 위한 대안으로, LLM의 내부 지식에만 의존하지 않고 외부 지식 소스를 동적으로 참조하여 답변을 생성하는 RAG가 주목받고 있다[3]. RAG는 정보 검색을 통해 질문과 관련된 외부 지식을 확보하고, 이를 바탕으로 언어 모델이 답변을 생성하도록 유도함으로써 환각을 억제하고 정보의 최신성을 보장한다. 그러나 RAG가 많은 가능성을 제시했음에도 불구하고, 실제 시스템을 구축하고 운영하는 과정에서 여러 내재적 한계가 발견되고 있다. 정보 검색의 정확성 문제부터 검색된 여러 문서를 효과적으로 통합하는 능력의 부재, 그리고 외부 지식이 주어졌음에도 불구하고 발생하는 환각 문제에 이르기까지 여러 한계가 존재한다[4]. 이에 본 논문은 RAG 시스템의 파이프라인을 단계적으로 분석하여 각 과정에서 발생하는 한계를 유형별로 분석하는 것을 목표로 한다. 나아가, 이러한 한계를 극복하기 위해 제안된 연구 동향을 바탕으로 RAG 시스템의 강건성과 신뢰도를 높이기 위한 구체적인 방향성을 제시하고자 한다.

Ⅱ. RAG의 작동 원리 및 구성 요소

RAG 시스템은 기본적으로 검색(Retrieve)와 생성(Generate)이라는 두 가지 핵심 모듈의 조합으로 구성된다. 전체 파이프라인은 일반적으로 Indexing, Retrieval, Generation의 3단계로 나누어 설명할 수 있다[5-6]. Indexing 과정에서는 외부 지식 소스의 텍스트를 고정된 크기의 청크로 분할하고, 각 청크를 임베딩 모델을 사용하여 고차원의 벡터로 변환한다. 이 벡터들은 Vector DB에 저장되어 유사도 기반 검색이 가능하도록 인덱싱된다. 인덱싱 후, Retreival 과정을 통해 입력된 사용자 쿼리에 대해 동일한 임베딩 모델을 사용하여 쿼리를 벡터로 변환한다. 이후 Vector DB 내에서 쿼리 벡터와 가장 유사한 K개의 청크 벡터를 검색하여 관련성이 높은 문서를 추출한다. 이 과정에서 BM25와 같은 전통적 키워드 기반 검색과 임베딩 기반의 의미 검색(Dense Retrieval)이 단독 또는 혼합되어 사용된다. 최종적으로 Generation 과정을 위해 검색된 K개의 문서는 사용자 쿼리와 함께 LLM의 프롬프트에 주입된다. LLM은 주입된 컨텍스트 정보를 바탕으로 최종 답변을 생성한다. 이를 통해 모델은 자신의 파라미터 내 지식에만 의존하지 않고 제공된 외부 지식을 근거로 답변하게 된다.

Ⅲ. RAG의 한계 분석

RAG 시스템의 전체 성능은 정보 검색(Retrieval)과 생성(Generation)이라는 유기적인 상호작용에 의해 결정되지만, 각 단계는 명확한 실패 지점(Failure Point)을 내포하고 있다. 가장 근본적인 문제는 정보 검색 단계에서 발생한다. Retriever가 사용자 쿼리의 미묘한 의미나 복잡한 의도를 완벽하게 포착하지 못해 관련성이 낮은 문서를 검색하면 RAG 시스템의 신뢰도를 저하된다. 이는 잘못된 정보를 후속 생성 단계에 전달하여 전체 응답의 질을 저하시키는 Garbage In, Garbage Out 현상을 야기한다. 설령 관련성 높은 문서를 성공적으로 검색했더라도 문제는 지속된다. LLM이 긴 컨텍스트 내에서 처음과 마지막 부분의 정보는 비교적 잘 활용하지만, 중간에 위치한 정보는 놓치는 경향을 보이는 Lost in the Middle 현상[7]은 검색된 정보의 가치를 반감시킨다. 즉, 핵심 정보가 검색되었더라도 프롬프트 내 위치에 따라 LLM에게 무시될 수 있다. 더 나아가, 원본 지식 소스에 서로 모순되거나 상충되는 정보가 존재할 경우, Retriever는 이를 구분하지 못하고 모두 가져올 수 있으며, 이는 Generator에게 논리적 혼

란을 야기하여 잘못된 답변을 생성하는 직접적인 원인이 된다.

이러한 검색 단계의 한계는 컨텍스트 통합의 비효율성으로 이어진다. Retriever가 여러 개의 문서 청크를 가져왔을 때, LLM이 이 정보들을 효과적으로 종합·추론·요약하는 데 실패하는 것이다. 특히 검색된 다수의 정보 중 어떤 내용이 질문의 핵심에 더 가까운지 우선순위를 판단하지 못하고, 일부 지엽적인 내용에만 편향되어 답변을 생성하는 문제가 발생한다. 또한 여러 문서에 흩어져 있는 단편적인 정보들을 논리적으로 연결하여 새로운 사실을 추론해야 하는 복잡한 질문에 대해, 단순히 특정 문서를 요약하는 수준에 그치는 한계를 보이기도 한다. 궁극적으로, RAG의 가장 큰 도입 목표인 환각 억제도 완벽히 보장되지 않는다. 제공된 컨텍스트가 모델 내부의 파라미터 지식과 충돌할 경우, 모델은 외부 정보를 무시하고 내부 지식을 기반으로 답변을 생성하려는 경향을 보인다. 또한, 컨텍스트 내의 특정 내용을 인용하거나 출처를 제시하도록 요구받았을 때, 존재하지 않는 내용이나 잘못된 출처를 제시하여 응답의 신뢰도를 스스로 훼손하기도 한다.

Ⅳ. RAG 시스템 개선을 위한 방향성 제언

앞서 분석한 RAG의 다층적인 한계를 극복하기 위해, 학계와 산업계에서는 파이프라인의 각 단계를 보강하고 전체 프로세스를 지능화하려는 다양한 연구가 진행되고 있다. 이는 검색 이전 단계에서부터 시작된다. Retriever의 성능을 극대화하기 위해 입력되는 쿼리 자체를 개선하는 Pre-Retrieval 접근법이 그 예시이다. 사용자의 초기 쿼리를 LLM을 사용하여 더 명확한 하위 질문들로 분해하거나(Query Decomposition), 검색에 용이한 키워드를 자동으로 추가하여 확장(Query Expansion)함으로써 검색의 시작점부터 정확도를 높이는 방식이 효과를 보이고 있다. 검색 단계 자체를 고도화하는 것 또한 핵심적인 연구 방향이다. 전통적인 키워드 기반의 BM25와 의미론적 유사도 기반의 Dense Retriever를 결합하는 Hybrid Search는 이제 단일 검색 방식의 한계를 보완하는 표준적인 접근법으로 자리 잡고 있다[8]. 여기서 더 나아가, 1차 검색된 문서들을 경량의 LLM을 사용하여 재평가하고 질문과의 관련성에 따라 순위를 정밀하게 조정하는 Re-ranking 과정을 추가함으로써, 최종적으로 Generator에게 전달될 문서의 질을 한 차원 높일 수 있다. 검색이 완료된 후에도 개선의 여지는 남아있다. Post-Retrieval 단계에서는 검색된 문서에서 노이즈를 제거하고 핵심 정보만을 추출하여 LLM의 부담을 줄이는 데 초점을 맞춘다. LLM을 활용해 검색된 청크들 중 질문과 관련 없는 정보를 필터링하거나, 긴 문서를 핵심만 남기고 요약하여 압축(Information Compression)한 후 Generator에 전달하는 방식은 앞서 언급된 Lost in the Middle 문제를 완화하고 모델이 중요한 정보에 집중하도록 돕는다. 마지막으로, 이러한 개별 모듈의 개선을 넘어 RAG 파이프라인 전체를 유기적인 시스템으로 발전시키려는 시도가 주목받고 있다. 단방향 프로세스의 한계를 극복하기 위해, 생성된 답변을 시스템 스스로 평가하고 필요에 따라 파이프라인을 반복하는 순환적, 자기 교정적 구조[9]가 제안되고 있다. 예를 들어, 생성된 답변의 근거가 부족하다고 판단되면 LLM 에이전트가 새로운 검색 쿼리를 생성하여 추가 정보를 탐색하고, 이를 바탕으로 기존 답변을 수정 및 개선하는 Self-RAG [9]와 같은 프레임워크는, RAG의 강건성과 자율성을 획기적으로 높일 수 있는 차세대 기술로서 그 잠재력을 보여준다.

Ⅴ. 결론

본 논문은 RAG 시스템을 구성하는 각 단계에서 발생하는 내재적 한계를 심도 있게 분석하고, 이를 극복하기 위한 최신 연구 동향 기반의 고도화 방향성을 제시했다. Retrieval 단계의 정확성 문제, Context 통합의 비효율성, 그리고 Generation 단계에서 여전히 발생하는 환각 문제는 RAG가 극복해야 할 핵심 과제이다. 이를 해결하기 위해 Query 최적화, Hybrid Search와 Re-ranking을 통한 Retrieval 고도화, Post-Retrieval 처리, 그리고 Self-Correction 메커니즘 도입과 같은 다각적인 접근이 필수적이다. 향후 RAG 시스템은 단순히 텍스트를 넘어 이미지, 코드 등 멀티모달 정보를 검색하고, 복잡한 태스크를 수행하는 LLM 에이전트의 핵심 구성 요소로 발전하며 그 적용 범위를 더욱 확장해 나갈 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-핵심전략 R&D사업 지원을 받아 수행된 연구임 (IITP-2025-RS-2025-02219190, 34%). 본 연구성과는 2025년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (RS-2025-25398164, 33%). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음 (IITP-2025-RS-2023-00256629, 33%).

참 고 문 헌

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

[2] Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 2.1 (2023).

[3] S. Bag, A. Gupta, R. Kaushik and C. Jain, "RAG Beyond Text: Enhancing Image Retrieval in RAG Systems," 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET, Sydney, Australia, 2024, pp. 1-6

[4] S. AboulEla, P. Zabihitari, N. Ibrahim, M. Afshar and R. Kashef, "Exploring RAG Solutions to Reduce Hallucinations in LLMs," 2025 IEEE International systems Conference (SysCon), Montreal, QC, Canada, 2025, pp. 1-8

[5] Gupta, Shailja, Rajesh Ranjan, and Surya Narayan Singh. "A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions." arXiv preprint arXiv:2410.12837 (2024).

[6] Oche, Agada Joseph, et al. "A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions." arXiv preprint arXiv:2507.18910 (2025).

[7] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." arXiv preprint arXiv:2307.03172 (2023).

[8] P. He, S. Wang, S. Chowdhury and T. -. H. Chen, "Evaluating the Effectiveness and Efficiency of Demonstration Retrievers in RAG for Coding Tasks," 2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), Montreal, QC, Canada, 2025, pp. 500-510

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