Digital Insight 2024

검색증강생성(RAG) 기술의 등장과 발전 동향

NIA 한국지능정보원
NATIONAL INFORMATION SOCIETY AGENCY

Contents

I 서론

현재 우리는 대규모 언어 모델(Large Language Models, LLM)의 시대에 살고 있다. ChatGPT, GPT-4와 같은 LLM은 자연어 처리 분야에 혁명을 일으키며 다양한 산업과 일상생활에 깊이 침투하고 있다. 그러나 LLM은 몇 가지 한계점을 가지고 있다. 정보의 최신성 부족, 잘못된 정보 제공, 편향된 응답 생성 등이 그것이다.

이러한 한계를 극복하기 위해 등장한 기술이 바로 RAG(Retrieval-Augmented Generation)이다. RAG는 LLM의 강력한 텍스트 생성 능력과 외부 지식 베이스를 결합한 혁신적인 접근 방식이다. 이 기술은 최신 정보를 실시간으로 검색하고 활용할 수 있어, 더 정확하고 신뢰할 수 있는 응답을 생성할 수 있다.

공공 서비스는 시민들의 삶의 질과 직결되는 중요한 영역이다. RAG 기술은 공공 서비스의 행정 효율성을 높이고, 시민들에게 더 정확하고 개인화된 서비스를 제공할 수 있는 잠재력을 가지고 있다. 예를 들어, 복잡한 정책 정보를 쉽게 이해할 수 있도록 설명하거나, 개인의 상황에 맞는 맞춤형 공공 서비스를 추천하는 데 활용될 수 있다.

그러나 RAG 기술의 도입에는 여러 가지 한계점들도 있다. 검색 품질에 대한 의존성, 개인정보 보호, 정보의 편향성 등의 문제를 해결해야 한다. 또한 기술 도입을 위한 제도적, 법적 기반을 마련하는 것도 중요하다.

본 보고서에서는 RAG 기술에 대해 가능한 일반인들이 쉽게 이해할 수 있도록 RAG의 기본 개념, 작동 원리, LLM 대비 장점, 파인튜닝과 RAG, RAG의 한계와 이를 극복하기 위한 진화, 그리고 다양한 적용 사례에 대해 살펴볼 것이다. 특히 공공서비스 혁신을 위한 RAG 기술 도입의 필요성에 대해서도 알아보고, 이를 위해 어떤 준비와 노력이 필요한지에 대한 통찰을 제공하고자 한다.

2 LLM 시대

A LLM 시대의 도래

2024년 겨울, 한국대학교 컴퓨터공학과 2학년 장유빈은 '디지털 디자인' 수업의 과제로 "부동소수점을 표현하기 위한 여러 방법과 그 역사"라는 주제로 10페이지 분량의 보고서를 작성해야 했다. 예전 같았으면 도서관에서 밤을 새워가며 자료를 찾고, 선배들에게 조언을 구하느라 바빴을 것이다. 하지만 이번에는 달랐다. 유빈은 ChatGPT를 열고 이렇게 입력했다. "컴퓨터에서 부동소수점을 표현하기 위한 여러 방법과 각각의 개발 배경 및 역사에 대해서 10페이지 분량의 보고서를 작성해줘". 불과 몇 초 만에 ChatGPT는 상세한 개요와 내용을 제시했고, 유빈은 이를 바탕으로 내용상 부족하다고 생각하는 부분에 대한 추가 정보를 요청했다. 결국, 유빈은 두 시간 만에 보고서를 손에 넣을 수 있었다.

이것이 바로 LLM 시대의 모습이다. LLM은 이미 우리 일상 깊숙이 파고들어 학업, 업무, 그리고 일상생활의 다양한 측면을 변화시키고 있다.

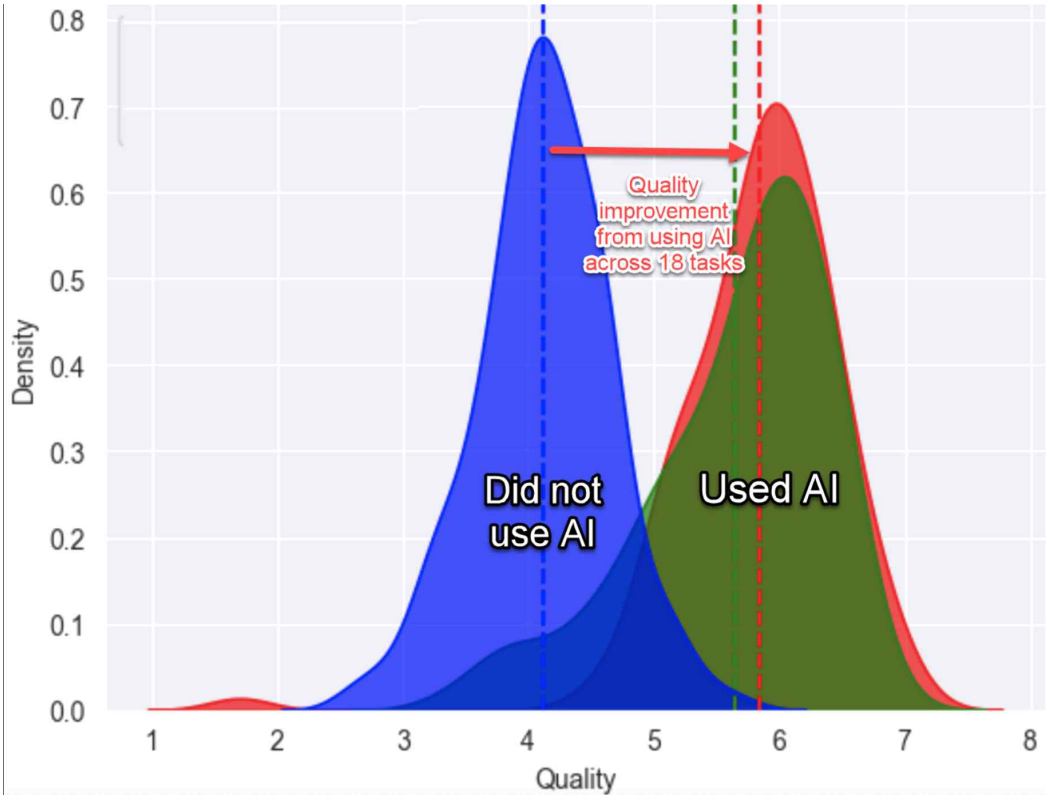
LLM의 활용은 학생들의 과제 수행에만 국한되지 않는다. 직장인들은 보고서 작성과 프레젠테이션 준비에 LLM을 활용하고 있으며, 프로그래머들은 코드 작성과 디버깅에 도움을 받고 있다. 심지어 작가들도 LLM을 통해 새로운 아이디어를 얻거나 작품의 초안을 빠르게 만들어내고 있다.

미국 스탠포드 대학의 인공지능 연구소가 2023년 발표한 보고서에 따르면, 대학생의 78%가 과제 수행에 LLM을 활용한 경험이 있으며, 직장인의 65%가 업무 효율성 향상을 위해 LLM을 사용하고 있다고 한다(Stanford AI Index Report, 2023)[1]. 이는 LLM이 이미 우리 사회에 깊이 뿌리 내렸음을 보여주는 명확한 증거다.

LLM의 활용이 가져온 변화는 매우 놀랍다. 정보 접근성이 크게 향상되었고, 복잡한 문제에 대한 빠른 해결책을 얻을 수 있게 되었다. 또한 창의적 작업에 있어서도 LLM은 새로운 영감의 원천이 되고 있다.

1) https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

하버드 비즈니스 스쿨의 논문 “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality”[2]에 따르면 LLM의 도입으로 인해 전문가들의 생산성이 평균 40% 더 높은 품질의 결과를 생성했다고 한다. 이는 산업혁명 이후 가장 큰 폭의 생산성 향상으로 평가받고 있다.



LLM 활용에 따른 생산품질 향상 [2]

2) https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf

B LLM의 가치

LLM은 방대한 데이터를 학습하여 여러 태스크를 수행할 수 있는 능력을 갖췄으며, 이를 통해 사용자 경험과 업무 효율성을 높이고 있다. 그렇다면 LLM이 왜 이렇게 널리 사용되는지 그 장점에 대해 구체적으로 알아보자.

ⅰ. 방대한 지식 기반을 통한 심층적 응답 가능

LLM은 방대한 텍스트 데이터를 학습하여 여러 분야의 지식과 맥락을 추론하고, 사용자에게 심층적이고 폭넓은 답변을 제공할 수 있다. 예를 들어, GPT-3와 같은 모델은 수십억 개의 매개변수를 통해 다양한 지식과 정보를 학습하여 사용자의 질문에 포괄적이고 정확한 답변을 제공할 수 있다.[3] 이러한 강점은 특히 고객 서비스, 의료 상담 등에서 유용하게 활용된다.

ⅱ. 자연스러운 언어 처리와 생성 능력

LLM의 강점 중 하나는 사람처럼 자연스럽고 문맥을 고려한 언어 생성 능력이다. Microsoft의 연구에 따르면, LLM은 단어 간의 관계와 문맥을 고려하여 문법적으로 정확하고 일관된 응답을 생성하며, 이를 통해 사용자와의 상호작용에서 높은 만족도를 제공한다.[4] 이러한 자연스러운 언어 생성 능력은 가상 어시스턴트, 고객 대화 시스템 등에 큰 기여를 하고 있다.

ⅲ. 다양한 태스크 수행 능력과 높은 범용성

LLM은 텍스트 분류, 번역, 요약, 질의응답 등 다양한 자연어 처리 태스크를 하나의 모델로 수행할 수 있는 다재다능한 특성을 갖추고 있다. Google의 BERT 모델 연구에서도 이러한 다기능성이 언급되었으며, LLM을 통해 기업들이 비용 효율적인 방식으로 여러 기능을 동시에 처리할 수 있다는 점이 강조된다.[5] 이를 통해 기업과 연구 기관은 모델 개발 비용을 절감할 수 있으며, 범용적인 NLP 솔루션으로 활용할 수 있다.

ⅳ. 사용자 친화적인 인터페이스 제공

LLM은 사용자 입력을 자연스럽게 학습하고 응답할 수 있어 접근성과 사용 편의성이 뛰어나다. Meta AI의 연구에 따르면, LLM을 활용한 인터페이스는 비전문가도 쉽게 접근할 수 있는 특징을 가지고 있으며, 이를 통해 사용자 경험을 극대화할 수 있다고 보고하고 있다.[6] 이는 특히 챗봇, 상담 시스템, 가상 어시스턴트 등에서 사용자의 편의성을 높이는 중요한 요인으로 작용한다.

3) Brown, T. B., et al. (2020). “Language Models are Few-Shot Learners.” OpenAI.
4) Radford, A., et al. (2019). “Improving Language Understanding by Generative Pre-Training.” Microsoft.
5) Devlin, J., et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Google.
6) Lewis, M., et al. (2021). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” Meta AI.

v. 다양한 산업 분야에서의 응용 가능성

LLM은 교육, 의료, 고객 서비스, 콘텐츠 제작 등 다양한 산업 분야에서 활용되고 있으며, 그로 인해 산업 전반의 효율성과 생산성을 높이는 데 기여하고 있다. 예를 들어, 의료 분야에서는 환자 상담이나 교육 분야에서의 맞춤형 학습 지원 등 다양한 기능을 수행하여 서비스 품질을 개선할 수 있다.[7] 이처럼 LLM은 다방면에서 활용 가능성이 높아 여러 산업에서 도입이 확대되고 있다.

LLM은 방대한 데이터 학습을 통한 심층적 응답, 자연스러운 언어 처리와 생성 능력, 다양한 태스크 수행, 사용자 친화적인 인터페이스, 그리고 광범위한 산업 응용 가능성 등으로 인해 최근 그 사용이 급격히 확산되고 있다. 이러한 장점들은 LLM이 다양한 응용 분야에서 중요한 도구로 자리잡게 하는 주요 요인이다.

C LLM의 한계

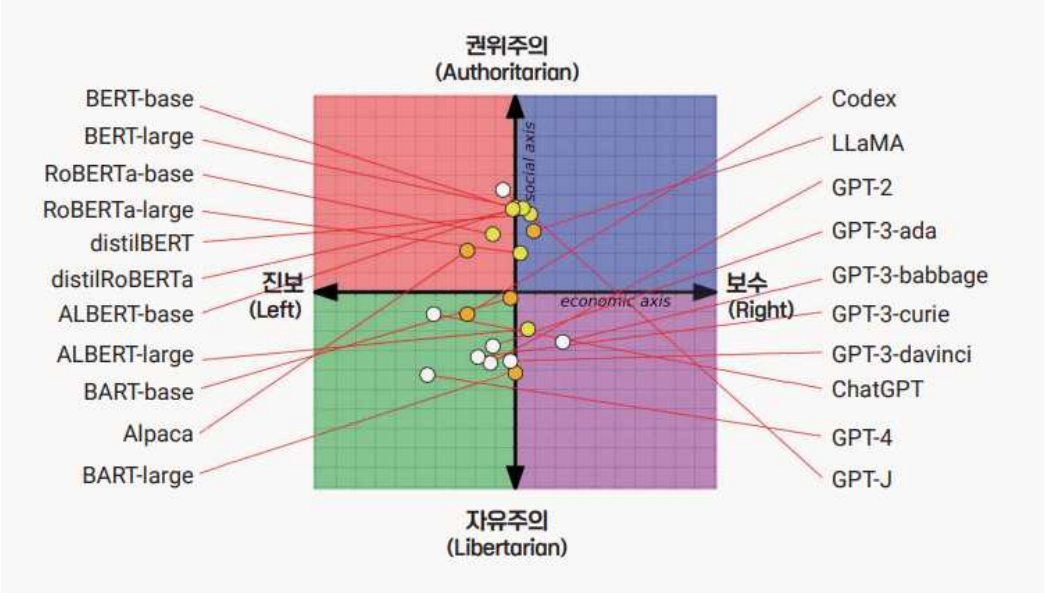
LLM(대규모 언어 모델)은 AI 기술 발전의 선두 주자로 자리 잡고 있지만, 여러 가지의 실용적 문제와 한계를 동반하고 있다. 특히, 편향성과 정확성 문제, 맥락 추론의 제한성, 일관성 문제, 그리고 윤리적 이슈 등이 자주 언급되는 문제들이다. 이 한계점을 살펴보도록 하자.

i. 편향성 문제

대규모 언어 모델은 방대한 양의 텍스트 데이터를 학습하는 과정에서 해당 데이터에 포함된 사회적 편향과 선입견을 그대로 반영하기 쉽다. LLM이 학습하는 데이터는 성별, 인종, 사회적 지위 등에 따라 편향된 표현을 포함할 수 있으며, 모델이 이러한 편향을 반영하여 출력을 생성하게 되면 특정 집단에 대한 차별적 메시지가 강화될 우려가 있다. 이러한 편향은 특정 집단에 대한 차별적 메시지를 강화할 수 있으며, 신뢰성과 공정성을 요구하는 애플리케이션에서 심각한 문제로 작용한다.[8] 이를 해결하기 위해 학습 데이터 전처리나 모델의 출력 제어와 같은 보완책이 연구되고 있다.

다음은 KPF미디어브리프 2024년 1호의 “챗GPT, 생성AI 중 가장 진보적(Liberal)으로 드러나”[9] 중 모델의 편향성을 표현한 그림이다.

7) Lee, S., et al. (2022). “Applications of Large Language Models in Healthcare: An Overview.” Journal of Medical AI Research.
8) Bender, E. M., et al. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.
9) https://www.kpf.or.kr/front/research/selfDetail.do?seq=596057&link_g_homepage=F



ii. 정보 정확성 문제와 Hallucination 현상

LLM은 사용자의 질문에 답변하는 데 뛰어난 능력을 보이지만, 종종 사실이 아닌 정보를 사실처럼 제공하는 할루시네이션(Hallucination) 현상을 일으키기도 한다. 이 현상은 모델이 스스로 지어낸 정보나 근거가 없는 내용을 그럴듯하게 설명하는 경향을 말하며, 특히 의료, 법률 등 고도의 정확성이 필요한 분야에서 큰 단점이 된다.

LLM의 할루시네이션 문제는 크게 두 가지 이유로 발생한다. 첫째, 모델이 방대한 양의 데이터를 기반으로 학습하지만, 독립적인 사실 확인 과정을 거치지 않기 때문에 잘못된 정보를 생성할 수 있다. 둘째, 모델이 학습 과정에서 문맥에 맞는 단어를 확률적으로 예측해 문장을 생성하기 때문에, 사용자의 질문이 기존 학습 데이터와 직접적인 관련이 없거나 불명확할 경우 잘못된 정보가 포함될 가능성이 높다.[10)]

10) Ji, Z., et al. (2023). “Survey of Hallucination in Natural Language Generation.” arXiv preprint arXiv:2303.12013.

예를 들어, LLM은 잘못된 의료 정보를 신뢰성 있는 정보처럼 제공할 수 있으며, 이로 인해 의료 상담과 같은 분야에서 실제 심각한 위험 요소로 작용할 수 있다고 지적된다. LLM이 특정 정보의 사실성을 검증하거나 의심하지 않고 가장 그럴듯한 답변을 생성하기 때문이다.

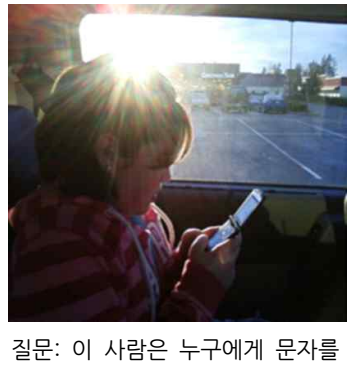
다음은 “Survey of Hallucination in Natural Language Generation.”[9]에서 제시하고 있는 할루시네이션의 예이다.

범주	원문	올바른 번역	할루시네이션 번역
내재형	迈克周四去书店。	마이크는 목요일에 서점에 간다.	제리는 목요일에 서점에 가지 않는다.
외재형	迈克周四去书店。	마이크는 목요일에 서점에 간다.	마이크는 목요일에 친구와 함께 서점에 즐겁게 간다.
분리형	Das kann man nur feststellen, wenn die Kontrollen mit einer großen Intensität durchgeführt werden.	이것은 더 철저한 통제가 이루어질 때만 확인할 수 있다.	피는 유일한 동력이며, 이것을 이해하기 위해 나는 너에게 말한다. 그것은 싸울 특권이다.
진동형	1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.	1995년에 피자 생산량이 3천만 개에 도달했다.	지난 20년 동안 미국은 동일한 입장이었으며, 그 이후에도 그러했다.

할루시네이션의 유형과 예

질문: 전화 화면에 무엇이 있나요?
답변: 친구로부터 온 문자 메시지

질문: 창 밖으로 무엇이 보이나요?
답변: 주차장



질문: 이 사람은 누구에게 문자를 보내고 있나요?
답변: 운전자

이미지 캡션에서의 할루시네이션

ⅲ. 맥락과 장기적인 텍스트 추론의 한계

LLM은 단일 문장이나 짧은 텍스트 추론에는 강점을 보이지만, 긴 텍스트나 복잡한 맥락을 추론하고 적절하게 처리하는 데 어려움을 겪는다. Google의 연구에 따르면, LLM은 문장 간의 연관성과 논리적 연결을 깊이 있게 파악하지 못해 장기적인 맥락을 반영한 응답을 생성하는 데 한계가 있다.[11] 이로 인해 LLM은 예측 가능한 일관된 대화를 제공하기 어려우며, 특히 긴 대화나 논리적 사고를 요구하는 상황에서는 부적합할 수 있다.

ⅳ. 일관성 부족과 확률적 응답의 문제

LLM은 같은 질문에 대해 매번 동일한 답변을 제공하지 않을 수 있다. 이러한 일관성 문제는 모델이 확률 기반의 응답 생성 메커니즘을 사용하기 때문에 발생하며, 이는 특히 신뢰성을 요구하는 응용 프로그램에서 문제로 작용할 수 있다. 예를 들어, 고객 서비스 시스템에서 일관성 없는 답변은 사용자 경험을 저하시킬 수 있으며, OpenAI의 연구에서도 이러한 문제가 보고되고 있다.[12] 이를 해결하기 위해 일부 연구자들은 특정 입력에 대해 일관된 출력을 제공할 수 있는 새로운 학습 구조를 탐색 중이다.

ⅴ. 윤리적 문제와 악용 가능성

LLM은 강력한 언어 생성 능력을 가지고 있지만, 허위 정보 확산, 불법 콘텐츠 작성 등으로 악용될 가능성도 크다. 특히, 사회적 혼란을 조장하는 허위 정보나 차별적 콘텐츠 생성에 LLM이 사용될 수 있다는 점에서 윤리적 문제가 제기되고 있다. 이와 관련하여 연구자들은 LLM의 출력을 제어하고 검증할 수 있는 안전 장치를 마련해야 한다고 주장하며, 이는 AI의 사회적 책임성을 높이는 방향으로 발전해야 한다는 의견이 많다.[13]

이러한 LLM의 한계점들은 특히 정확성과 공정성이 중요한 공공 서비스 분야에서 심각한 문제를 초래할 수 있다. 따라서 이러한 한계를 극복할 수 있는 새로운 기술적 접근이 필요하게 되었고, 이는 RAG(Retrieval-Augmented Generation) 기술의 등장으로 이어졌다.

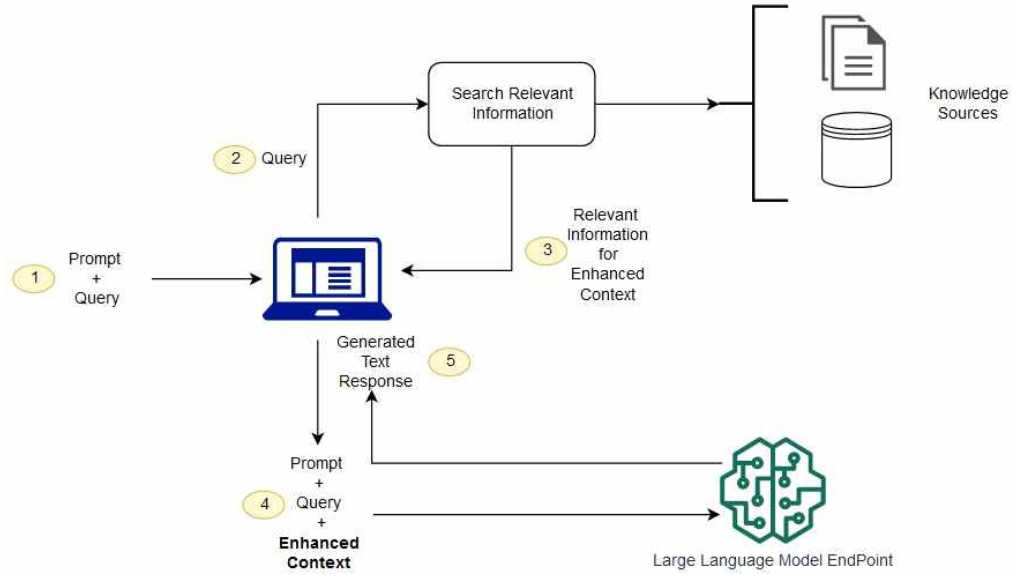
11) Devlin, J., et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Google.
12) Brown, T. B., et al. (2020). “Language Models are Few-Shot Learners.” OpenAI.
13) Bender, E. M., et al. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

3 RAG의 이해

A RAG의 개념

RAG 기술의 개념은 2020년 페이스북 AI 연구팀(현 Meta AI)이 발표한 논문 "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"에서 처음 제안되었다(Lewis et al., 2020). 이 논문에서 연구팀은 RAG가 개방형 도메인 질의응답, 사실 확인, 슬롯 채우기 등 다양한 자연어 처리 작업에서 우수한 성능을 보인다는 것을 입증했다.

RAG 기술은 AI 커뮤니티에서 큰 반향을 일으켰다. 구글 스칼라에 따르면 RAG 관련 연구 논문의 수가 2020년 이후 매년 두 배 이상 증가하고 있으며, 2023년에는 천 편 이상의 논문이 발표되었다. RAG 기술의 등장은 LLM의 한계를 극복할 수 있는 새로운 가능성을 제시했다. 특히 정보의 최신성, 사실 확인의 용이성, 맥락 추론 능력 향상 등의 측면에서 RAG는 LLM을 크게 보완할 수 있는 잠재력을 보여주고 있다.



RAG의 개념도[14]

14) https://aws.amazon.com/ko/what-is/retrieval-augmented-generation/

RAG는 LLM의 한계를 보완하기 위해 고안된 혁신적인 자연어 처리 기술이다. 이 기술은 인덱싱(Indexing), 검색(Retrieval)과 생성(Generation)이라는 세 가지 주요 요소를 결합해서 사용자가 원하는 답변을 생성할 때 학습된 데이터에만 의존하지 않고 외부 데이터 소스를 활용할 수 있도록 한다. RAG는 언어 모델의 생성 능력과 외부 지식 검색의 장점을 결합해서 더 정확하고 상황에 맞는 응답을 제공할 수 있다.[15)]

RAG는 LLM이 가진 다음과 같은 문제점들을 해결하기 위해 개발되었다.

ⅰ. 지식의 한계

LLM은 학습 데이터의 시간적 제약으로 인해 최신 정보를 반영하지 못하는 문제가 있었다. 예를 들어, GPT-4 Turbo의 경우 2023년 12월까지의 데이터만을 학습했기 때문에 그 이후의 정보는 알지 못했다. RAG는 이러한 한계를 극복하기 위해 외부 데이터베이스를 활용하여 실시간으로 최신 정보를 검색하고 활용할 수 있게 하였다.

ⅱ. 할루시네이션 문제

LLM은 때때로 사실이 아닌 정보를 그럴듯하게 생성하는 할루시네이션 현상을 보였다. 이는 모델의 신뢰성을 크게 저하시키는 요인이었다. RAG는 외부 지식 베이스에서 검색한 사실적 정보를 바탕으로 답변을 생성함으로써 이 문제를 완화한다.

ⅲ. 정보의 출처 제공

AI 시스템의 답변에 대한 신뢰성과 투명성 요구가 증가했다. RAG는 검색된 정보의 출처를 명시할 수 있어, 사용자에게 더 신뢰할 수 있는 정보를 제공할 수 있게 되었다.

ⅳ. 도메인 특화 지식의 한계

범용 LLM은 특정 도메인의 심도 있는 전문 지식을 모두 포함하기 어려웠다. RAG를 통해 특정 분야의 전문 데이터베이스를 연결함으로써 도메인 특화된 정확한 정보를 제공할 수 있게 되었다.

15) Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401.

v. 비용 효율성

모델을 지속적으로 재학습시키는 것은 시간과 비용 면에서 비효율적이다. RAG는 모델 자체를 재학습시키지 않고도 외부 데이터베이스를 업데이트함으로써 최신 정보를 반영할 수 있어 비용 효율적이다.

vi. 맥락 추론 능력

단순한 패턴 매칭을 넘어 질문의 맥락을 정확히 추론하고 적절한 답변을 생성해야 한다는 요구가 있었다. RAG는 외부 지식을 활용하여 질문의 배경과 맥락을 더 잘 파악할 수 있게 해준다.

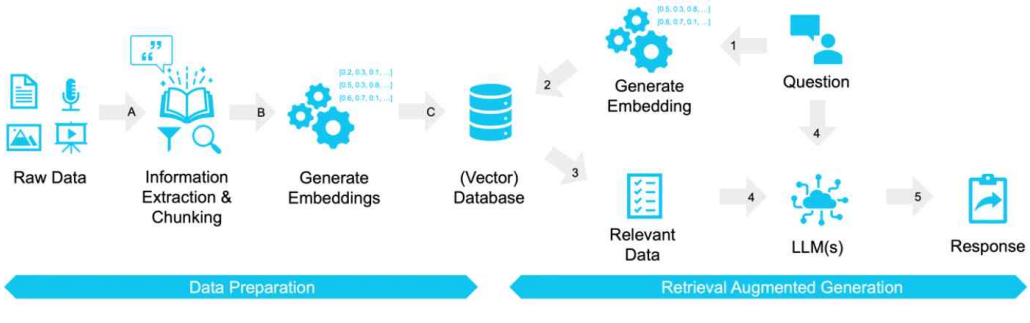
vii. 지식 검색과 언어 생성의 통합

기존의 검색 시스템과 생성 모델을 별도로 운영하는 것보다 이를 통합적으로 수행할 수 있는 프레임워크의 필요성이 제기되었다. RAG는 이 두 가지 기능을 효과적으로 결합한 솔루션을 제공한다.

이러한 배경들로 인해 RAG는 LLM의 한계를 극복하고 더 정확하고 신뢰할 수 있는 AI 시스템을 구축하기 위한 중요한 기술로 부상하게 되었다.

B RAG의 작동 원리

RAG의 작동 프로세스는 1단계: 데이터 준비(Data Preparation), 2단계: 데이터 검색(Data Retrieval Augmented Generation)으로 이루어진다.

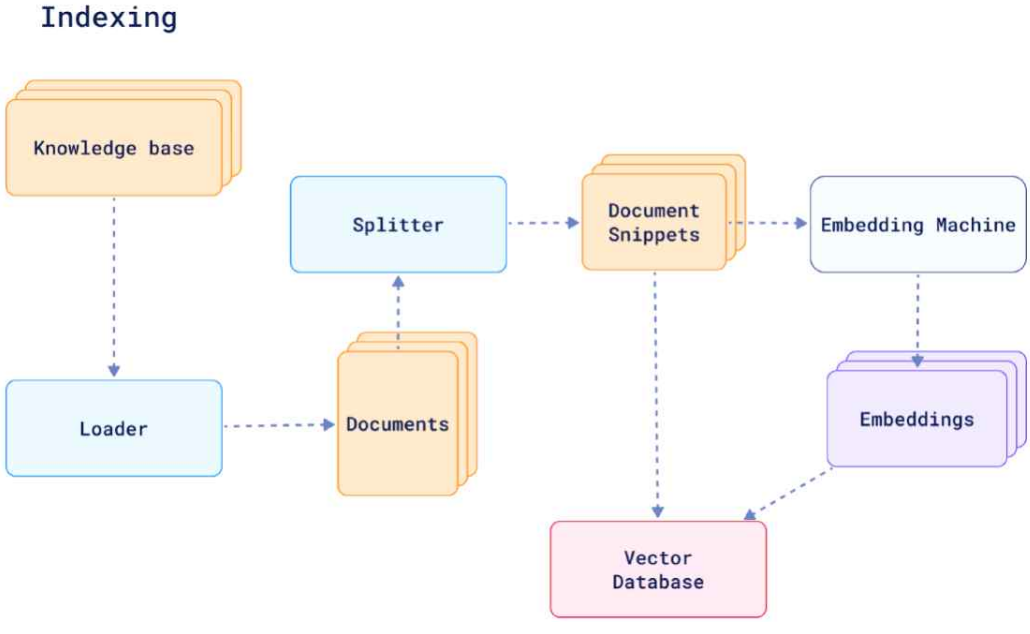


RAG의 작동 프로세스[16)]

16) https://cratedb.com/use-cases/ai-ml-database/retrieval-augmented-generation-pipelines)

● 1단계: 데이터 준비(Data Preparation)

RAG의 데이터 준비는 AI 모델이 효과적으로 활용할 수 있는 고품질 데이터를 준비하는 중요한 과정이며 RAG 시스템의 성능과 정확성에 직접적인 영향을 미치므로 세심한 주의가 필요하다. 이 과정에서는 데이터를 수집하고, 정제하고, 임베딩하는 여러 단계가 포함된다. 아래에서는 데이터 준비의 핵심 과정과 그 기술적 배경을 설명한다.[17]



인덱스 생성(Indexing)을 위한 플로우 차트[18]

ⅰ. 데이터 소스 식별 및 수집

솔루션 또는 사용 사례에 대한 비즈니스 요구 사항을 명확하게 정의한다. 그 후 신뢰할 수 있는 다양한 소스에서 관련 데이터를 수집한다. 데이터는 웹 페이지, 데이터베이스, 문서 저장소 등 여러 형태일 수 있다. 이 단계에서 데이터의 신뢰성과 다양성이 중요한 역할을 하며, 시스템의 성능을 결정짓는다. 데이터 수집 시 웹 크롤링 툴이나 API를 사용하여 데이터베이스와 연결할 수 있다.

17) https://www.abbyy.com/blog/getting-started-with-rag/
18) https://qdrant.tech/articles/what-is-rag-in-ai/

ii. 데이터 추출 및 정제

데이터 정제는 원본 데이터를 사용할 수 있는 형태로 만드는 과정이다. 이 단계에서는 불필요한 데이터를 제거하고 일관된 형식으로 데이터를 표준화하는 등, 데이터의 품질을 높이기 위한 다양한 기법들이 사용된다.

- 중복 제거: 데이터의 중복된 부분을 제거함으로써 시스템의 효율성을 높인다.

- 텍스트 표준화: PDF, HTML, Word, Markdown 등 다양한 파일 형식을 일반 텍스트로 변환하여 일관성을 유지한다. 이를 통해 모델은 다양한 형식의 데이터를 동일한 방법으로 다룰 수 있다.

- 불필요한 정보 제거: 광고, 링크, 헤더와 같은 불필요한 텍스트를 필터링하여 의미 있는 데이터만 남긴다.

iii. 청크 분할 (Chunking)

텍스트 청크 분할은 데이터를 의미 있는 크기로 나누는 과정이다. 언어 모델은 맥락 정보를 처리할 수 있는 양에 제한이 있기 때문에 긴 문서를 작은 단위로 나누는 것이 중요하다.

- 청크 크기 조정: 모델의 입력 크기 제한에 맞추어 텍스트를 나누어야 한다. 일반적으로 언어 모델의 토큰 제한에 맞추어 적절히 텍스트를 분할하는 것이 중요하다.

- 의미 단위로 분할: 청크 분할은 의미를 유지할 수 있도록 문단이나 문장 단위로 이루어지며, 이를 통해 검색 과정에서 유사성을 평가할 때 정보 손실을 줄일 수 있다.

iv. 임베딩 및 벡터화 (Embedding and Vectorization)

임베딩은 텍스트를 고차원 벡터로 변환하여 시스템이 의미론적으로 유사한 데이터를 검색하고 관리할 수 있도록 하는 과정이다. 벡터화는 문서와 질문 간의 유사성을 수치적으로 계산하기 위해 필수적인 단계이다.

- 임베딩 모델 사용: BERT, RoBERTa, DistilBERT와 같은 사전 학습된 모델을 사용하여 텍스트를 벡터화한다. 이 과정은 의미 정보를 보존한 채 텍스트를 수치화하여 저장할 수 있게 한다.

○ 고차원 벡터 공간: 텍스트 간의 의미적 유사성을 계산하기 위해 벡터는 고차원 공간에 위치하게 된다. 이때, 벡터화된 텍스트는 유사도 검색 과정에서의 비교 기준이 된다.

v. 데이터 인덱스 생성 (Index Creation)

데이터 인덱스 생성은 임베딩된 벡터를 빠르게 검색할 수 있도록 데이터베이스에 저장하는 과정이다. 효율적인 인덱스 생성을 통해 대량의 데이터를 실시간으로 검색하는 성능을 높일 수 있다.

○ 벡터 데이터베이스 사용: 벡터화된 데이터를 저장하고 빠르게 검색하기 위해 Faiss와 같은 벡터 데이터베이스를 사용한다.

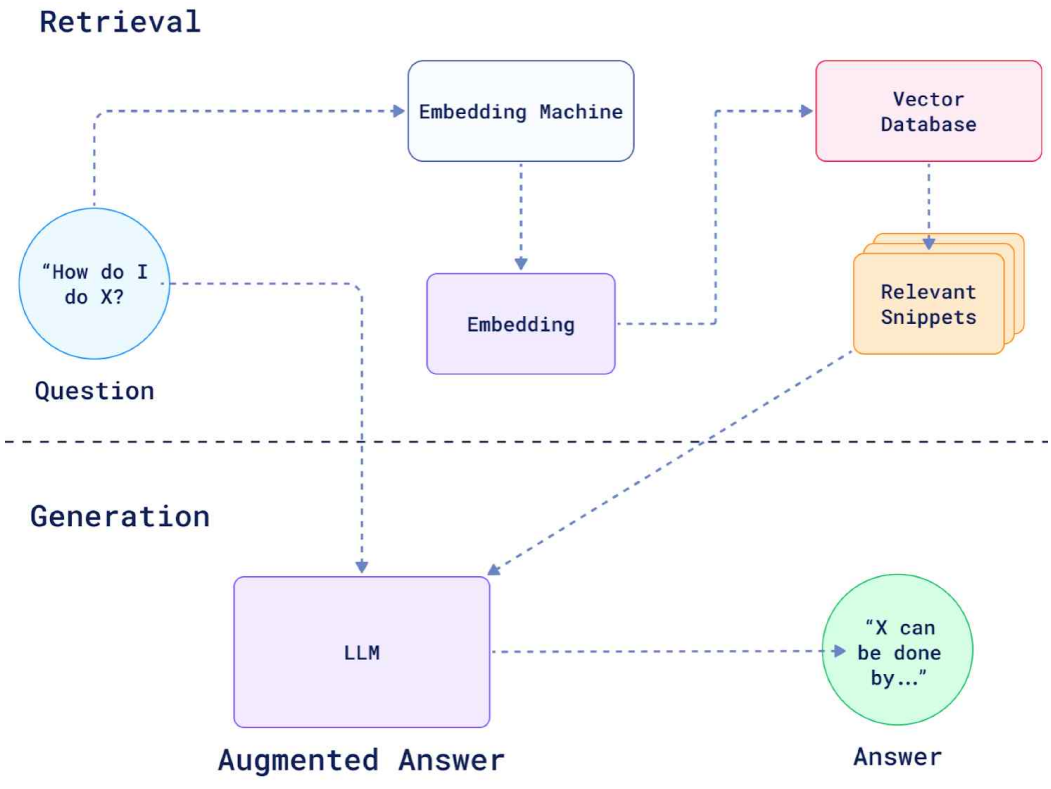
○ Key-Value 저장: 임베딩 벡터는 해당 텍스트와 매칭되는 키와 함께 저장된다. 이를 통해 빠르게 검색 결과를 찾아낼 수 있다.

데이터 준비는 RAG 시스템의 성공적인 구현을 위해 필수적이다. 데이터의 수집, 정제, 청크 분할, 임베딩 및 인덱스 생성의 각 단계는 RAG 시스템이 효율적으로 동작하고 사용자의 질문에 대해 정확한 답변을 제공하는 데 중요한 역할을 한다. 이러한 단계들은 서로 유기적으로 연결되어 있어, 데이터 준비 과정에서의 작은 오류도 시스템의 전체 성능에 영향을 미칠 수 있다. 따라서 데이터 준비 단계의 최적화는 RAG 시스템의 품질을 높이는 핵심적인 요소이다.

● 2단계: 데이터 검색(Data Retrieval Augmented Generation)

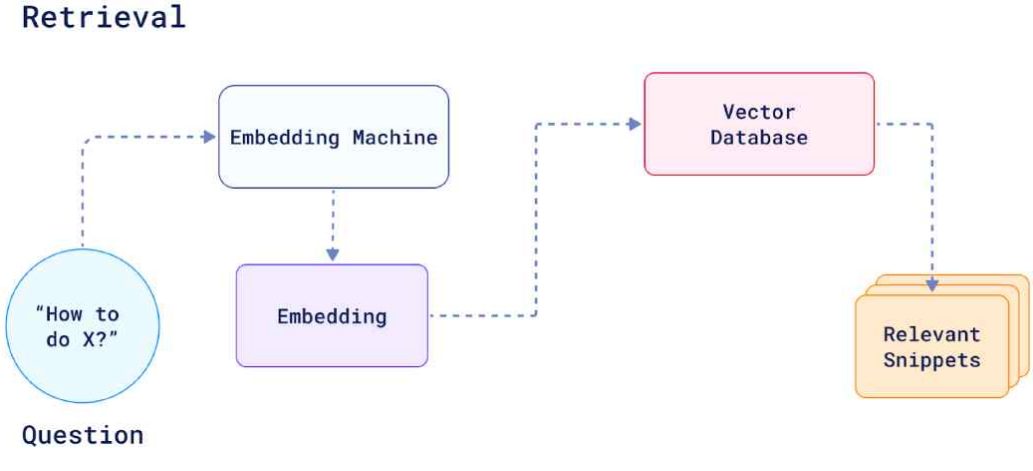
RAG는 인공지능이 외부 지식을 효과적으로 활용하여 질의에 대한 답변을 생성하는 과정으로, 두 가지 주요 단계로 나뉜다. 검색(Retrieval)과 생성(Generation), 이 두 단계는 서로 밀접하게 연계되어 모델이 단순히 학습된 지식을 넘어서 외부의 최신 정보까지 통합적으로 활용할 수 있도록 한다. 아래에서는 데이터 검색의 두 단계와 그 기술적 배경을 설명한다.[19)20)]

19) https://www.abbyy.com/blog/getting-started-with-rag/
20) https://qdrant.tech/articles/what-is-rag-in-ai/



검색(Retrieval)과 생성(Generation)의 플로우[19)]

i. 검색(Retrieval)



검색(Retrieval)을 위한 플로우[19)]

검색 단계는 사용자의 질문이나 프롬프트를 처리하는 것으로 시작된다. 사용자가 입력한 질의는 질의 인코더(Query Encoder)를 통해 임베딩 벡터로 변환된다. 이 임베딩 벡터는 고차원 공간에서 질문을 수치화한 것으로, 컴퓨터가 쉽게 이해하고 처리할 수 있는 형태로 바꿔준다.

이후, 임베딩 벡터는 대규모 벡터 데이터베이스에서 유사성을 기준으로 가장 관련성이 높은 정보들을 찾아낸다. 여기서 사용하는 기법은 주로 유사성 검색(similarity search)으로, 질문과 데이터베이스 내 각 문서 블록 간의 벡터 거리를 계산하여 가장 가까운 상위 K개의 벡터를 선택한다.

검색 단계에서 중요한 점은 검색 모델의 성능이다. 검색 모델은 사용자가 입력한 질문을 얼마나 잘 임베딩하여 벡터 형태로 변환하는지가 성능에 큰 영향을 미친다. 이를 위해 BERT, DPR(Dense Passage Retrieval), Sentence-BERT와 같은 다양한 모델들이 사용된다. 특히 BERT 기반의 모델들은 고차원 공간에서의 벡터 간 거리를 계산하는 데 있어 높은 정확도를 자랑한다.

검색 단계에서 사용하는 벡터 데이터베이스는 대규모의 문서들이 저장된 곳으로, FAISS(Facebook AI Similarity Search)나 Annoy(Approximate Nearest Neighbors Oh Yeah)와 같은 라이브러리를 활용해 빠른 검색이 가능하다. 이러한 라이브러리들은 수백만 개 이상의 벡터를 빠르게 검색할 수 있도록 설계되어 있으며, 검색 효율성을 높이기 위한 다양한 최적화 기법들이 적용되어 있다. 벡터 데이터베이스 내에서는 벡터 간의 코사인 유사도(cosine similarity)나 유클리드 거리(Euclidean distance)를 사용하여 유사성을 계산하며, 이 과정을 통해 가장 관련성 높은 정보를 찾아내게 된다.

검색 단계에서는 또한 LLM과의 적절한 연계를 위해 검색 매커니즘(Retrieval Mechanism)은 중요한 역할을 한다. 여기서는 ElasticSearch, Pinecone, Weaviate와 같은 검색 엔진들이 사용되어 데이터베이스에 저장된 벡터들을 효과적으로 검색한다. 이러한 검색 엔진은 높은 처리량을 유지하면서도 검색의 정확도를 극대화하기 위한 다양한 최적화 기능들을 제공한다.

예를 들어, 사용자가 "블록체인 기술의 장점은 무엇인가요?"라는 질문을 던진다면, 이 질문은 먼저 벡터 형태로 변환된 뒤, 데이터베이스에서 블록체인 기술에 대한 정보가 포함된 벡터들과 비교되어 높은 유사도를 가진 텍스트 덩어리들이 반환된다. 이때 검색된 텍스트는 블록체인의 투명성, 보안성, 탈중앙화 등 주요 키워드가 포함된 정보들로 구성된다.

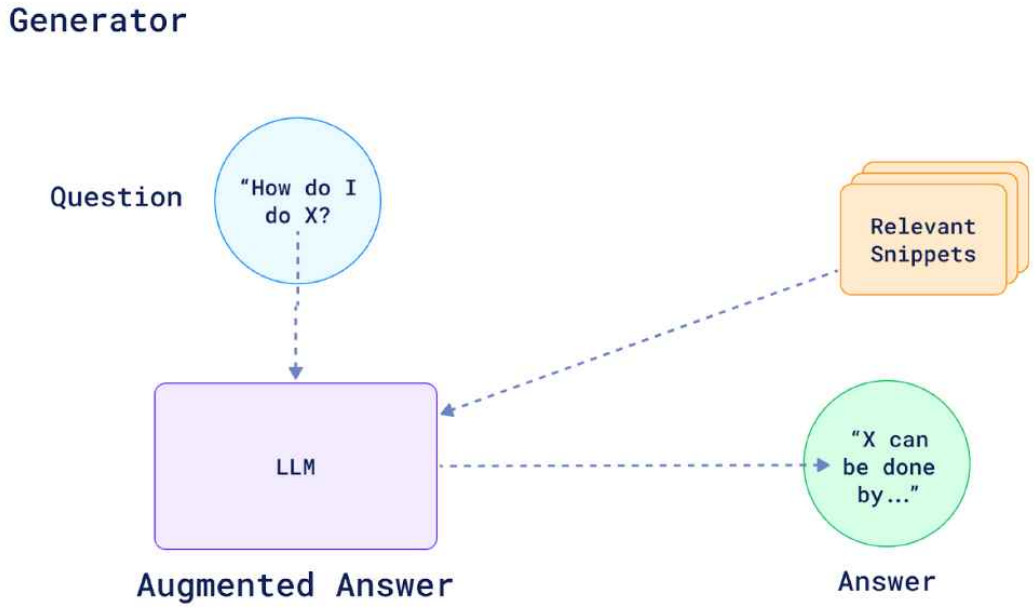
앞의 그림은 검색 단계의 전반적인 흐름을 보여준다. 사용자의 질문이 임베딩 벡터로 변환된 후, 벡터 데이터베이스에서 유사성을 기준으로 정보를 검색하는 과정이 시각적으로 설명되어 있다.

또한, 검색 단계에서 반환되는 정보들은 모델이 답변을 생성하는 데에 직접적으로 사용되기 때문에, 검색의 정확도가 최종 응답의 질을 결정하는 중요한 요소이다. 검색 모델은 여러 개의 후보 텍스트를 반환하며, 이를 통해 모델이 최종 응답을 생성할 때 다양한 정보를 참조할 수 있도록 한다.

검색 단계에서의 기술적 과제는 대규모 벡터 데이터베이스에서 빠르고 정확한 검색을 가능하게 하는 것이다. 이를 위해 HNSW(Hierarchical Navigable Small World)와 같은 알고리즘들이 사용되며, 이 알고리즘들은 높은 효율성과 정확도를 제공하여 실시간 검색에 적합하다. HNSW는 대규모 데이터셋에서도 검색 시간을 최소화하면서 높은 유사성을 유지할 수 있도록 설계된 그래프 기반의 접근 방식이다.

검색 단계에서의 성공적인 수행은 생성 단계의 질에 직접적인 영향을 미치며, 이를 통해 사용자에게 더욱 신뢰성 있고 풍부한 정보를 제공할 수 있다.

ii. 생성(Generation)



생성(Generator)을 위한 플로우[19]

생성 단계는 검색 단계에서 찾은 관련 텍스트들과 사용자의 질문을 결합하여 LLM에 입력으로 제공하는 과정이다. 여기서 사용되는 LLM은 대표적으로 GPT 시리즈와 같은 대형 언어 모델들이며, 이 모델들은 주어진 컨텍스트를 바탕으로 자연스럽고 구체적인 응답을 생성한다.

생성 단계에서 중요한 것은 모델이 검색된 정보들을 어떻게 효과적으로 활용하느냐이다. 검색 단계에서 수집된 정보는 주로 짧은 텍스트 덩어리(chunk)들로 이루어져 있으며, 이러한 정보들을 LLM이 어떻게 결합하여 의미 있는 응답을 만들어내는지가 핵심이다. 예를 들어, GPT-3나 GPT-4와 같은 모델들은 수백억 개 이상의 파라미터를 활용해 검색된 정보를 바탕으로 자연스러운 문장을 생성한다.

생성 단계에서 LLM은 입력된 컨텍스트를 바탕으로 정확한 문맥 파악과 의미 추론을 수행한다. 이 과정에서 모델은 검색된 정보들 사이의 상관관계를 이해하고, 이를 바탕으로 사용자 질문에 대한 최적의 답변을 생성하게 된다. 예를 들어, "블록체인 기술의 장점은 무엇인가요?"라는 질문에 대해, 검색 단계에서 블록체인 기술의 투명성, 보안성, 탈중앙화 등의 정보가 수집되었다면, 생성 단계에서는 이 정보들을 적절히 조합하여 "블록체인 기술의 주요 장점은 투명성, 보안 강화, 그리고 탈중앙화입니다"와 같은 응답을 생성한다.

생성 단계에서는 기술적인 과제도 존재한다. 첫 번째는 모델이 검색된 정보를 정확하게 참조하면서도 할루시네이션 문제를 최소화하는 것이다. LLM은 종종 존재하지 않는 사실을 생성하는 경향이 있기 때문에, 검색된 정보와의 일관성을 유지하는 것이 매우 중요하다.

생성 단계에서는 또한 지식 증강(knowledge augmentation)의 역할이 중요하다. 검색 단계에서 제공된 정보들은 모델이 응답을 생성할 때 참조되는 증강된 정보로 활용된다. 이 증강된 정보는 사용자가 원하는 답변의 신뢰성과 정확성을 높이는 데에 기여한다. 예를 들어, 의료 분야에서 "특정 질병의 최신 치료법은 무엇인가요?"라는 질문이 주어진다면, 검색 단계에서 최신 연구 논문이나 임상 데이터를 수집하고, 생성 단계에서는 이를 바탕으로 구체적인 치료법을 제안하는 응답을 만들어낼 수 있다.

위 그림은 생성 단계에서의 프로세스를 시각적으로 설명하고 있다. 검색된 정보들이 LLM에 입력으로 제공되어 최종적으로 사용자에게 응답이 생성되는 과정을 보여준다.

C LLM 대비 장점

RAG와 LLM은 모두 자연어 처리 기술이지만, 그 접근 방식과 특성에 있어 중요한 차이점이 있다. 이러한 차이점을 이해하는 것은 각 기술의 장단점을 파악하고 적절한 활용 방안을 모색하는 데 중요하다. LLM과 RAG의 차이점을 체계적으로 분석하고, 각 기술의 장단점과 함께 적절한 활용 방안을 알아보자.

ⅰ. 실시간 데이터

LLM은 대량의 텍스트 데이터를 사전에 학습한 후, 파인튜닝을 통해 특정 도메인에 맞는 정보를 습득한다. 그러나 파인튜닝 과정에서 정적인 데이터에 의존하기 때문에 학습이 완료된 이후에는 새로운 지식을 자동으로 업데이트할 수 없다. 예를 들어, GPT-3는 2022년까지의 데이터로 학습되었기 때문에 이후의 정보는 포함되지 않으며, 최신 정보를 제공하기 어렵다는 한계를 가진다.

이와 달리, RAG는 외부 데이터베이스나 검색 엔진을 실시간으로 검색하여 최신 정보를 응답 생성에 활용한다. 이는 RAG가 최신성 유지와 관련해 큰 강점을 가지며, 실시간으로 업데이트된 정보를 기반으로 사용자의 질문에 응답할 수 있음을 의미한다. 특히, 외부 데이터베이스나 검색 엔진의 업데이트만으로 최신 정보를 유지할 수 있어 실시간 정보가 중요한 분야에서 활용도가 높다. 이러한 특징 덕분에 RAG는 최신성과 신뢰성이 요구되는 환경에서 기존 LLM의 한계를 보완하며 더욱 실용적인 솔루션으로 자리 잡고 있다.

ⅱ. 신뢰성과 투명성

LLM은 때때로 할루시네이션이라 불리는 오류를 발생시키며, 정보의 출처를 명확히 제시하지 못하는 한계를 가지고 있다. 또한, LLM은 "블랙박스" 모델로 작동하기 때문에 특정 응답이 생성된 이유를 명확히 설명하기 어렵다. 이러한 특성은 특히 중요한 의사결정에 활용될 때 신뢰성과 투명성을 저하시킬 수 있다.

반면, RAG는 신뢰할 수 있는 외부 소스를 활용해 응답을 생성함으로써 더 높은 신뢰성과 정확성을 제공한다. 뿐만 아니라, 검색된 정보의 출처를 명확히 제시할 수 있어 사용자가 직접 확인하거나 응답의 근거를 검증할 수 있는 장점을 가진다. 이는 RAG가 AI 시스템의 투명성을 높이고, 중요한 의사결정에서 신뢰를 확보하는 데 크게 기여함을 의미한다. RAG는 단순히 정보를 생성하는 데 그치지 않고, 응답의 이유를 명확히 설명할 수 있는 구조를 제공하여 LLM의 한계를 효과적으로 보완한다.

iii. 맥락 이해 능력

LLM은 광범위한 주제에 대해 응답할 수 있지만, 특정 상황이나 사용자의 개인적 맥락을 완전히 이해하지 못할 수 있다. 이는 모델이 사전 학습된 패턴에 기반하여 응답을 생성하기 때문이다. 하지만 RAG는 사용자의 질문과 직접적으로 관련된 정보를 검색하여 활용함으로써, 더 정확한 맥락 이해가 가능하다. 이는 특정 상황이나 도메인에 특화된 정보를 제공할 수 있다는 점에서 LLM보다 유리하다.

iv. 확장성

LLM은 새로운 정보나 지식을 추가하려면 전체 모델을 재학습해야 하며, 이는 시간과 비용이 많이 든다. 반면, RAG는 지식 베이스에 새로운 정보를 추가하는 것만으로도 시스템의 지식을 확장할 수 있어 특정 도메인에 대한 전문성을 쉽게 향상시킬 수 있다.

v. 리소스

LLM은 대규모 모델을 실행하기 위해 상당한 컴퓨팅 리소스가 필요하며, 모델 크기가 커질수록 더 많은 메모리와 처리 능력이 요구된다. RAG는 검색 과정이 추가되므로 LLM보다 더 많은 리소스를 요구할 수 있지만, 더 작은 언어 모델을 사용할 수 있어 전체적인 리소스 요구사항을 줄일 수 있다.

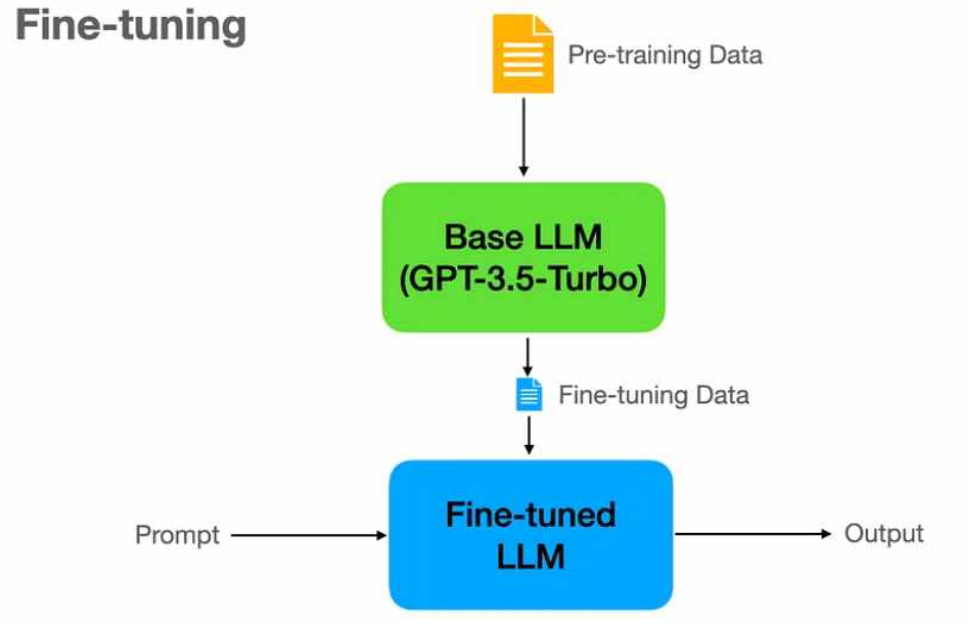
vi. 속도

LLM은 모델 내부의 정보만을 사용하므로 비교적 빠른 응답 생성을 할 수 있지만, 모델 크기가 커질수록 처리 시간이 증가할 수 있다. RAG는 외부 검색 과정이 포함되어 있어 응답 생성에 시간이 더 걸릴 수 있지만, 효율적인 검색 알고리즘과 캐싱 기법을 통해 이를 최소화할 수 있다.

vii. 편향성

LLM은 학습 데이터에 내재된 편향성을 그대로 반영할 수 있으며, 이는 성별, 인종, 문화적 편견 등 다양한 형태로 나타날 수 있다. 반면, RAG는 다양한 소스에서 정보를 검색함으로써 단일 데이터셋에 의존할 때보다 편향성을 조금 줄일 수 있지만, 검색 알고리즘이나 지식 베이스 자체의 편향성에 주의해야 한다.

D 파인튜닝과 RAG



파인튜닝 플로우[21]

ⅰ. 파인튜닝이란?

파인튜닝(fine-tuning)은 대규모 데이터로 사전 학습된(pre-trained) 딥러닝 모델을 특정 작업에 맞게 추가로 학습시켜 성능을 향상시키는 과정이다. 새로운 task에 따라 가중치를 조정(Fine tuning)하는 작업인데, 이를 통해 이미 방대한 데이터로 학습된 모델의 일반화된 지식을 새로운 데이터셋으로 재학습하여 특정 작업에 최적화하는 과정이다.

ⅱ. 파인튜닝의 과정

1) 사전 학습된 모델 선택: 목표 작업과 관련성 높은 모델을 선택한다.
2) 데이터셋 준비: 특정 작업에 맞는 데이터를 수집하고 전처리한다.
3) 모델 조정: 준비된 데이터로 모델을 재학습시킨다. 이때 학습률을 신중하게 설정하여 기존 지식을 유지하면서 새로운 작업에 적응시킨다.

21) https://medium.com/@gurpartap.sandhu3/fine-tuning-llms-using-openais-gpt3-5-to-build-a-hyper-focussed-grumpy-greg-6ecddceedd09

ⅲ. 파인튜닝의 장점

파인튜닝은 적은 양의 데이터로 좋은 성능을 얻을 수 있으며, 학습 시간과 비용을 절감할 수 있다. 그리고 이미 방대한 데이터로 학습한 일반화된 지식을 특정 도메인이나 작업에 맞게 훈련시켰으므로 해당 분야에서는 더 높은 성능을 발휘한다.

ⅳ. RAG VS 파인튜닝[22)]

RAG와 파인 튜닝은 대규모 언어 모델(LLM)의 성능을 개선하기 위한 두 가지 주요 접근 방식으로, 각각 고유한 장점과 적합한 활용 상황을 가지고 있다. RAG는 모델이 외부 데이터베이스나 최신 정보를 실시간으로 참조하여 답변을 생성하도록 돕는 방식을 통해, 새로운 지식을 통합하는 데 매우 효과적이다. 이 방법은 특히 변화가 빠른 정보나 최신 지식이 필요한 상황에서 유리하다. 예를 들어, 최근의 뉴스나 기술 트렌드와 같은 최신 정보를 바탕으로 한 응답을 요구하는 응용 프로그램에서는 RAG의 강점을 극대화할 수 있다.

반면, 파인 튜닝은 모델이 특정 데이터셋을 바탕으로 내부 파라미터를 조정하여, 특정 도메인이나 스타일에 맞는 응답을 일관성 있게 생성하도록 돕는다. 이를 통해 모델은 일정한 형식과 어조를 유지하며 복잡한 요구사항을 충족할 수 있게 된다. 예를 들어, 특정 기업의 고유한 브랜드 어조로 일관된 답변을 생성해야 하거나, 특정한 양식에 맞춘 답변이 요구되는 상황에서는 파인 튜닝이 큰 효과를 발휘한다.

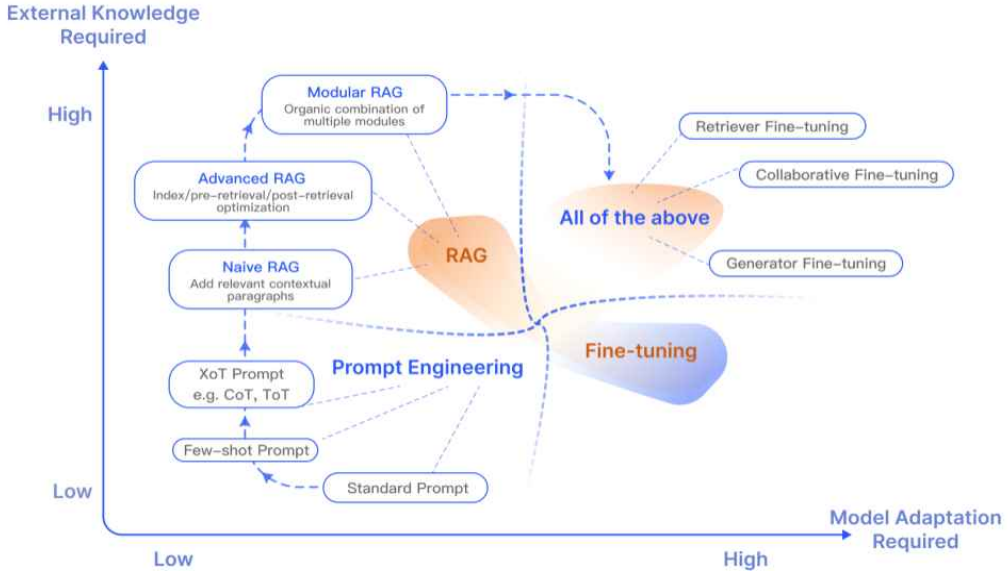
흥미로운 점은, 이 두 방법이 서로 배타적인 것이 아니라는 점이다. 오히려 RAG와 파인 튜닝은 지식 집약적이면서도 확장 가능한 응용 프로그램에서 서로를 보완하는 관계로 작용할 수 있다. 빠르게 변화하는 지식에 대한 접근이 필요하면서도, 특정 형식이나 스타일로 맞춤형 응답을 제공해야 하는 복합적인 상황에서는 RAG로 새로운 지식을 끌어오고, 파인 튜닝으로 모델의 일관성을 유지하면서 최적의 성능을 달성할 수 있다.

다음 표는 RAG와 파인 튜닝된 모델들 간의 특징을 비교한 “Which is the best tool to boost your LLM Application?”라는 글에서 가져온 표이며. [23)] 다음 그림은 “Retrieval-Augmented Generation for Large Language Models: A Survey”[24)] 논문에서 제시하는 외부 검색과 모델 맞춤화의 정도에 따른 최적화 모델 선택 방법이다.

22) https://ahha.ai/2024/07/24/rag/
23) https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7
24) Yunfan Gao, et al. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey

기능	RAG	파인튜닝
지식 업데이트	직접 검색 지식을 업데이트하여 정보가 최신 상태를 유지하도록 하며, 빈번한 재훈련이 필요하지 않으며, 동적인 데이터 환경에 적합	정적인 데이터를 저장하며, 지식 및 데이터 업데이트를 위해 재훈련이 필요
외부 지식	외부 리소스를 능숙하게 활용하며, 문서나 기타 구조화된/비구조화된 데이터베이스에 특히 적합	대형 언어 모델과 사전 학습된 외부 지식을 조정하는 데 사용할 수 있지만, 자주 변경되는 데이터 소스에는 덜 실용적일 수 있음
데이터 처리	최소한의 데이터 처리와 핸들링이 필요	고품질 데이터셋에 의존하며, 제한된 데이터셋은 성능 향상에 도움이 되지 않을 수 있음
모델 맞춤화	정보 검색 및 외부 지식 통합에 중점을 두며, 모델 행동이나 작성 스타일은 완전히 맞춤화하기 어려울 수 있음	특정 톤이나 용어 기반의 특정 도메인 지식을 바탕으로 LLM의 행동과 작성 스타일을 조정할 수 있음
해석 가능성	답변은 특정 데이터 소스로 추적 가능하기 때문에, 해석 가능성과 추적성이 더 향상됨	블랙박스처럼 작동하여, 모델이 특정 방식으로 반응하는 이유를 항상 명확히 알 수 없으며, 해석 가능성이 상대적으로 낮음
계산 자원	데이터베이스 관련 검색을 지원하기 위해 계산 자원이 필요. 외부 데이터 소스 통합 및 업데이트가 유지되어야 함	고품질 훈련 데이터셋의 준비 및 관리, 파인튜닝 목표의 정의, 계산 자원이 필요
지연 요구 사항	데이터 검색이 포함되어 지연이 더 길어질 수 있음	파인튜닝 후 LLM은 검색 없이 응답할 수 있어 지연이 더 짧아짐
할루시네이션 감소	검색된 증거에 기반하여 답변을 생성하기 때문에 할루시네이션이 적음	모델을 특정 도메인 데이터로 훈련시켜 할루시네이션을 줄일 수 있지만, 익숙하지 않은 입력에 직면했을 때 여전히 할루시네이션을 나타낼 수 있음
윤리 및 프라이버시 문제	외부 데이터베이스에서 텍스트를 저장 및 검색함에 따라 윤리 및 프라이버시 문제가 발생할 수 있음	훈련 데이터의 민감한 내용으로 인해 윤리 및 프라이버시 문제가 발생할 수 있음

RAG와 파인튜닝의 특징 비교



외부 검색과 모델맞춤화의 정도에 따른 최적화 모델 선택 방법

E RAG의 한계

RAG(Retrieval-Augmented Generation)는 뛰어난 정보 검색과 생성 능력을 갖춘 기술이지만, 여러 한계와 해결해야 할 과제들도 분명히 존재한다. 이를 명확히 인식하고 적절히 대응해야 RAG를 더 효과적으로 활용할 수 있다. 여기서는 RAG의 대표적인 한계들을 알아보자.

ⅰ. 검색 품질에 대한 의존성

RAG의 가장 큰 한계 중 하나는 검색 단계의 품질에 크게 의존한다는 점이다. 검색이 잘못되면 결국 생성된 응답의 질도 떨어질 수밖에 없다. 이는 검색 알고리즘이 얼마나 정교하고, 얼마나 관련성 높은 정보를 잘 뽑아내는지에 달려 있다. 고급 검색 알고리즘 도입, 검색 결과의 다양성 확보, 그리고 지속적인 모니터링을 통해 검색 품질을 높이는 노력이 중요하다.

ⅱ. 높은 계산 복잡도

RAG는 검색과 생성 단계를 모두 거치기 때문에 계산 복잡도가 높고, 처리 시간과 자원 소모가 증가할 가능성이 크다. 대규모 데이터베이스에서 정보를 검색하고 이를 기반으로 답변을 생성하는 과정은 단순한 생성 모델보다 더 많은 계산 자원과 시간이 요구된다. 이러한 문제는 RAG의 실시간 활용이나 대규모 데이터 처리에서 주요한 한계로 작용할 수 있다.

이를 해결하기 위해 여러 최적화 전략이 제안된다. 예를 들어, 고도화된 인덱싱 기법을 도입해 검색 속도를 향상시키거나, 분산 처리 시스템을 활용해 연산 부담을 여러 시스템에 분산시킬 수 있다. 또한, 캐싱 전략을 적용하면 반복적인 계산을 줄일 수 있어 효율성을 높일 수 있다. 더불어, 하드웨어 자원을 효율적으로 사용하는 방법과 소프트웨어 최적화 기법을 병행함으로써 RAG의 연산 성능을 개선할 수 있다. 이러한 접근법은 RAG의 성능 한계를 극복하고 보다 실용적이고 확장 가능한 시스템으로 발전시키는 데 기여한다.

iii. 개인정보 문제

RAG는 외부 데이터를 검색해 활용하는 과정에서 개인정보 문제가 발생할 수 있다. 검색 과정에서 개인정보가 유출되거나, 민감한 데이터를 잘못 사용할 위험이 있다. 이를 해결하기 위해서는 데이터 익명화 기술이나 개인정보 필터링 시스템을 적용하고, 필요시 사용자 동의를 받는 체계를 갖추는 것이 필요하다.

iv. 신뢰성과 편향성 문제

RAG는 강력한 도구이며 신뢰할 수 있는 외부 소스를 활용해 응답을 생성하기 때문에 LLM에 비해 더 높은 신뢰성과 정확성을 제공하지만, 여전히 사용하는 데이터 소스의 신뢰성과 편향성 문제 및 생성된 답변의 신뢰성을 완전히 보장하기에는 어려운 한계가 있다. 모델이 편향된 데이터에 기반해 응답을 생성할 경우, 결과물 역시 편향될 가능성이 높다. 따라서 다양하고 신뢰할 수 있는 데이터 소스를 선택하고, 정보의 교차 검증을 통해 신뢰성을 높이는 것이 중요하다. 또한, 데이터 편향성을 줄이기 위해 편향성 감지 및 완화 알고리즘을 적용하는 노력이 필요하다.

더불어, RAG가 생성하는 답변이 특히 법률이나 의료 분야 등의 중요한 결정에 사용될 경우, 그 신뢰성을 검증하는 과정이 필수적이다. 이를 위해 사람이 개입하여 검토하거나 자동화된 검증 시스템을 도입하는 방법을 고려할 수 있다. 이러한 기술적 접근과 함께 검색 품질, 계산 복잡도, 개인정보 보호, 신뢰성과 편향성 문제를 해결하기 위한 정책적 노력이 뒷받침되어야 한다. 이를 통해 RAG를 더욱 안전하고 효율적으로 활용할 수 있을 것이다.

v. 지식의 최신성과 정확성 유지

RAG는 대규모 텍스트 데이터베이스를 기반으로 하기 때문에, 이 데이터베이스의 최신성과 정확성이 중요하다. 만약 최신 정보로 갱신되지 않거나 오류가 포함되어 있다면, 잘못된 답변이 생성될 위험이 있다. 이를 해결하려면 데이터베이스를 정기적으로 업데이트하고, 최신 정보를 지속적으로 반영하도록 체계적으로 관리하는 것이 중요하다.

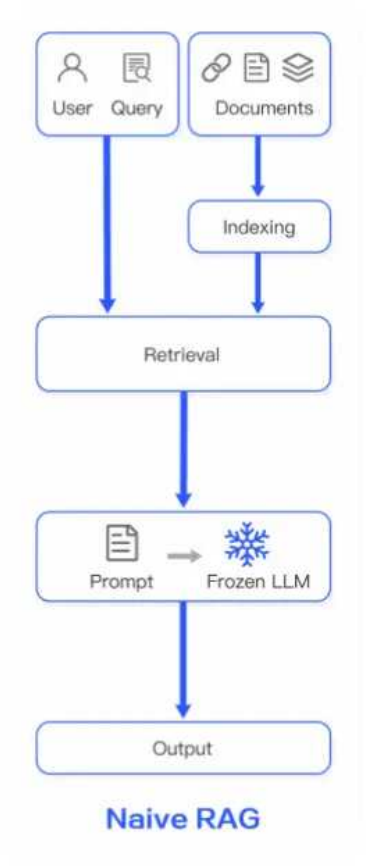
F RAG의 진화: 패러다임의 변화

앞서 보았듯이 RAG는 LLM의 여러 단점들을 보완할 수 있지만, 여전히 많은 한계가 존재하는 것이 사실이다. 그리고 이러한 단점들을 극복하기 위해 많은 시도가 있었다. 그 중 RAG 시스템의 성능, 비용 효율성을 위해 발전된 패러다임을 소개한다.

RAG 시스템은 성능, 비용 효율성을 위해 기존 Naive RAG에서 Advanced RAG, Modular RAG로 발전해 왔다. [25)]

ⅰ. Naive RAG

Naive RAG는 전통적인 인덱싱, 검색 및 생성 단계를 따른다. 즉, 사용자 입력으로 관련 문서를 찾아서 프롬프트와 결합하고 모델에 전달하여 최종 응답을 생성한다. 애플리케이션에서 여러 단계의 대화 상호 작용이 포함된 경우 대화 내역을 프롬프트에 통합할 수 있다.



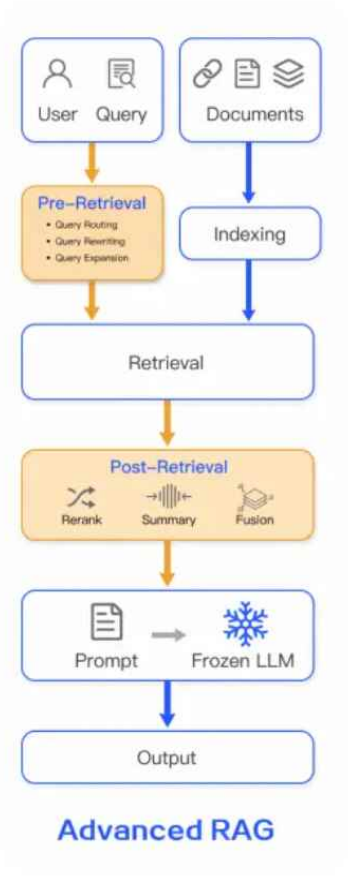
Naive RAG에는 낮은 정밀도(잘못 정렬된 문서 검색) 및 낮은 재현율(모든 관련 문서 검색 실패)과 같은 한계가 있다. RAG 시스템이 해결하려고 했던 중요한 문제인, LLM으로 오래된 정보가 전달되는 것이 발생할 수도 있다. 이로 인해 할루시네이션 문제가 발생하고 부정확한 응답이 발생할 수 있다.

그리고 증강을 적용할 때 중복 및 반복 문제가 발생할 수도 있다. 검색된 여러 구절을 사용할 때 순위 매기기와 스타일/톤을 조정하는 것도 중요하다. 또 다른 문제는 생성 작업이 증강 정보에 지나친 의존으로 인해 검색된 콘텐츠를 반복하게 하는 모델로 이어지지 말아야 한다는 것이다.

25) https://arxiv.org/pdf/2312.10997

ii. Advanced RAG

Advanced RAG는 Naive RAG의 한계를 극복하기 위해 다양한 최적화 전략을 도입했다. 사전 검색(Pre-Retrieval) 단계에서 질의 재구성(query rewriting)과 질의 확장(query expansion)을 통해 사용자 질문의 명확성을 높이고, 사후 검색(Post-Retrieval) 단계에서는 검색된 문서를 재정렬(re-ranking)하거나 핵심 정보를 강조하여 검색 품질을 개선하였다. 또한, 슬라이딩 윈도우 방식과 세부적인 텍스트 분할, 메타데이터 활용을 통해 인덱싱 효율성을 강화하였으며, 과도한 문맥 정보가 노이즈로 작용하는 문제를 방지하기 위해 적절한 정보를 선택적으로 활용했다. 이러한 개선을 통해 Advanced RAG는 검색 정확성과 생성 품질을 높이고, 정보를 효율적으로 통합하며, 더 신뢰할 수 있는 응답을 생성할 수 있게 되었다.



1) 사전 검색(Pre-Retrieval)

사전 검색 단계에서는 검색 효율성과 결과 품질을 높이기 위해 데이터 인덱싱 최적화와 임베딩 전략이 사용된다. 데이터 인덱싱 최적화는 텍스트 세분화, 인덱스 구조 개선, 메타데이터 활용, 문서 정렬 최적화, 혼합 검색 등을 통해 정확성과 맥락 적합성을 강화한다. 특히, 청크 크기 조정과 메타데이터 필터링은 검색 성능을 크게 향상시킨다.

임베딩은 검색된 문서와 질문 간의 의미론적 유사성을 극대화하여 결과의 신뢰성과 관련성을 높이는 핵심 기술이다. 이를 위해 정밀 조정 임베딩과 동적 임베딩 전략이 활용되며, 도메인 특화 데이터나 맥락 민감도를 반영한 모델 조정을 통해 검색의 효율성을 극대화한다. 이러한 최적화는 RAG 시스템의 전반적인 성능을 향상시키는 데 기여한다.

2) 사후 검색(Post-Retrieval)

검색 후 단계는 Advanced RAG에서 매우 중요한 과정으로, 검색된 문서의 중요한 정보를 질의와 결합하여 LLM에 효율적으로 입력하는 것을 목표로 한다. 모든 검색 결과를 한꺼번에 제공하면 비효율적이며 컨텍스트 윈도우를 초과할 수 있기 때문에, 관련 문서에 대한 후처리가 필요하다.

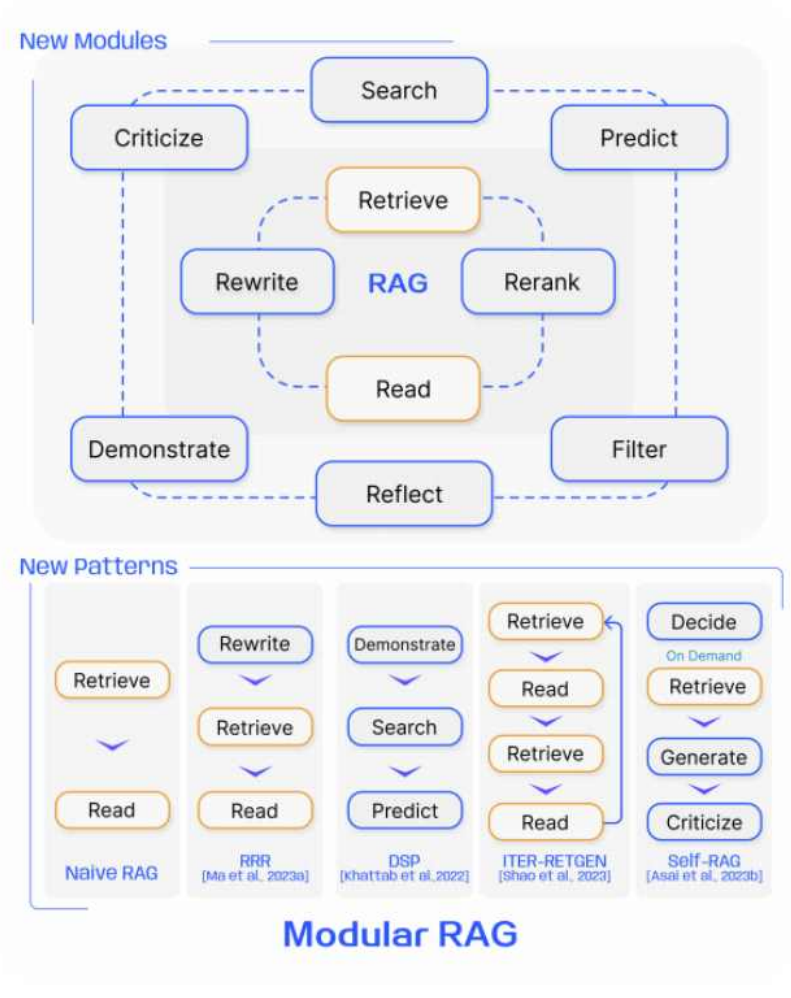
이 과정은 크게 순위 다시 매기기(ReRank)와 프롬프트 압축(Prompt Compression)으로 나뉜다. 순위 다시 매기기는 검색된 문서 중 관련성이 높은 정보를 앞부분에 배치해 LLM의 처리 효율성을 높이는 방법으로, LangChain, LlamaIndex 등에서 구현되어 있다. 프롬프트 압축은 잡음을 줄이고 중요한 정보를 강조하여 컨텍스트 길이를 줄이는 방식으로, LLMLingua와 같은 모델을 활용해 주요 정보를 선별한다. 이러한 최적화는 RAG 시스템의 응답 정확성과 처리 효율성을 크게 향상시킨다.

3) RAG 파이프라인 최적화 (RAG Pipeline Optimization)

RAG 파이프라인 최적화는 RAG 시스템의 효율성과 정보 품질을 높이기 위해 검색 전략과 프로세스를 개선하는 것을 목표로 한다. 이를 위해 하이브리드 검색을 활용해 키워드, 의미론적, 벡터 검색을 혼합하여 가장 관련성 높은 정보를 검색하며, 재귀적 검색으로 초기 단계에서 작은 문서 블록을 검색하고, 후속 단계에서 더 많은 맥락 정보를 제공해 효율성과 맥락적 풍부함을 조화롭게 결합한다. 또한, 역추적 프롬프트(StepBack-prompt)를 도입해 일반적인 개념에 기반한 추론을 장려하고, 서브쿼리를 활용해 다양한 질의 전략을 적용하여 검색 효율성을 높인다. 마지막으로, HyDE(Hypothetical Document Embeddings)로 LLM이 생성한 가상의 문서 임베딩을 기반으로 실제 문서를 검색해, 임베딩 유사성을 통해 더 정교한 검색을 가능하게 한다. 이러한 최적화 전략은 RAG의 성능을 전반적으로 강화한다.

iii. Modular RAG

Modular RAG는 Advanced RAG의 발전된 형태로, 기존 RAG 시스템의 구조적 한계를 극복하고 더 높은 유연성과 적응성을 제공하기 위해 설계되었다. Advanced RAG가 검색과 생성의 품질을 높인 데 반해, Modular RAG는 다양한 새로운 모듈과 패턴을 도입해 특정 작업과 시나리오에 맞는 최적화를 가능하게 했다. 이를 통해 RAG 시스템은 검색, 생성, 적응성을 동시에 강화하며, 효율성과 정보 품질에서 큰 도약을 이루었다.



1) 새로운 모듈들(New Modules)

Modular RAG는 다섯 가지 주요 모듈을 추가하여 기존 시스템의 기능성을 확장했다. 첫째, 검색 모듈은 기존의 유사성 검색을 넘어 LLM이 생성한 코드, SQL 등을 활용해 외부 검색 엔진이나 지식 그래프 등 다양한 데이터를 직접 검색할 수 있도록 한다. 둘째, 메모리 모듈은 LLM의 메모리 기능을 활용해 관련 정보를 저장하고 반복적으로 활용함으로써 추론 성능을 높인다. 셋째, 추가 생성 모듈은 검색된 데이터의 중복과 잡음을 줄이는 동시에 LLM이 생성한 콘텐츠를 활용해 더 풍부한 정보를 제공한다. 넷째, 태스크 적응 모듈은 특정 작업에 맞는 쿼리와 프롬프트를 자동 생성해 다양한 다운스트림 작업에 적응할 수 있다. 마지막으로, 정렬 및 검증 모듈은 질의와 검색 결과 간의 불일치를 해결하고 정보의 신뢰성을 평가해 RAG의 강건성을 보장한다.

2) 새로운 패턴들 (New Patterns)

Modular RAG는 모듈 간 상호작용을 최적화하는 새로운 패턴을 도입했다. 모듈 추가 또는 교체를 통해 기존 검색-읽기(RR; Retrieval-Read) 구조에 새로운 기능을 통합하거나 개선할 수 있으며, 재작성-검색-읽기(RRR; Rewrite-Retrieve-Read) 방식은 검색 쿼리를 조작해 더 적합한 결과를 생성한다. 모듈 간 조직적 흐름 조정은 검색과 생성 모듈 간의 동적인 상호작용을 가능하게 하여 특정 시나리오에 맞춘 유연한 최적화를 지원한다. 이러한 구조적 유연성은 복잡한 작업에서의 적응성을 강화하고, RAG 시스템의 전반적인 성능을 크게 향상시킨다.

결론적으로, Modular RAG는 기존 Advanced RAG에서 한 단계 더 나아가, 구조적 유연성과 기능 확장을 통해 다양한 시나리오와 데이터 소스에 적합한 최적화된 검색과 생성을 가능하게 한다. 이는 RAG를 더욱 강력하고 실용적인 도구로 만들어준다.

G 국내외 주요 기업의 RAG 적용 사례

RAG 기술은 다양한 산업 분야에서 활발히 적용되고 있으며, 많은 기업들이 이를 통해 서비스 품질을 향상시키고 있다. 다음은 주요 기업들의 RAG 기술 적용 사례와 그 효과를 살펴보자.

ⅰ. 구글의 메드팜 [26)27)28)]

구글의 Med-PaLM은 대규모 언어 모델을 의료 분야에 적용한 주목할 만한 사례이다. Med-PaLM은 PaLM(Pathways Language Model)을 기반으로 개발되었으며, 의료 질문 답변 능력을 크게 향상시켰다. Med-PaLM의 핵심은 instruction prompt tuning이라는 기법을 사용해 모델을 의료 도메인에 맞게 조정한 것이다. 이는 RAG(Retrieval-Augmented Generation)의 한 형태이며, 소수의 의료 관련 예시들을 사용해 모델을 미세 조정함으로써, 의료 지식과 맥락에 대한 이해도를 높였다. 이 접근법의 효과는 여러 벤치마크에서 입증되었다. Med-PaLM은 MedQA, MedMCQA, PubMedQA 등 다양한 의료 질문 답변 데이터셋에서 최고 성능을 달성했다. 특히 미국 의사 면허 시험 스타일의 MedQA 데이터셋에서 86.5%의 정확도를 기록해, 이전 최고 성능을 19% 이상 상회했다. Med-PaLM의 성능은 단순한 정확도 향상을 넘어선다. 의사들의 평가에 따르면, Med-PaLM의 답변은 과학적 합의와의 일치도, 의학적 추론 능력, 해를 끼칠 가능성 등 여러 측면

26) Karan Singhal, el al. (2023). Towards Expert-Level Medical Question Answering with Large Language Models
27) Karan Singhal, el al. (2023). Large language models encode clinical knowledge
28) 6https://www.mobihealthnews.com/news/googles-medical-llm-proves-increase-accuracy

에서 크게 개선되었다. 예를 들어, 과학적 합의와의 일치도는 Flan-PaLM의 61.9%에서 Med-PaLM의 92.6%로 상승했다. 또한 Med-PaLM은 소비자 의료 질문에 대한 답변에서도 뛰어난 성능을 보였다. 의사들은 9개 평가 축 중 8개에서 Med-PaLM의 답변을 의사의 답변보다 선호했다. 이는 모델이 의료 정보를 정확히 이해하고 전달할 수 있음을 시사한다.

ii. IBM의 Watson Health [29)30)31)]

IBM Watson Health는 RAG 기술을 활용하여 암 진단과 치료 권장사항을 제공하는 시스템이다. 이 시스템은 환자의 건강 기록과 대규모 의학 데이터를 분석하여 개인 맞춤형 치료 계획을 제안한다. Watson for Oncology는 수많은 의학 문헌과 임상 데이터를 학습하여, 의사들이 환자에게 가장 적합한 치료법을 찾는 데 도움을 준다.

이 시스템은 환자의 진료 기록과 검사 결과를 바탕으로 가능한 치료 옵션을 제시하며, 각 옵션의 신뢰도를 색상으로 구분해 직관적으로 이해할 수 있도록 한다. 연구에 따르면, IBM Watson for Oncology는 전문 의사들의 치료 권장사항과 96%의 일치율을 보였다.

RAG 시스템의 도입으로 인해 Watson Health는 최신 의학 정보를 신속하게 분석하고 반영할 수 있어, 환자 치료에 필요한 최신 지식을 빠르게 적용할 수 있는 장점을 갖게 되었다. 이러한 기술은 의료 분야에서 보다 정확하고 효율적인 진료를 가능하게 하여, 환자들에게 더 나은 결과를 제공하는 데 기여하고 있다.

iii. Siemens의 내부 지식 관리 시스템 [32)33)34)]

Siemens는 RAG 기술을 활용하여 내부 지식 관리 시스템을 혁신하고 있다. Siemens는 방대한 양의 설계 데이터와 문서를 보유하고 있는데, 이 데이터들은 과거의 설계, 매뉴얼, 테스트 자료 등으로 구성되어 있다. 이러한 데이터들은 기업의 중요한 자산이지만, 필요한 정보를 특정 상황에 맞게 찾거나 데이터에서 트렌드를 분석하는 것은 매우 어려운 작업이었다.

29) https://www.projectpro.io/article/rag-use-cases-and-applications/1059#mcetoc_1ib25mn04c
30) http://hbeetec.com/bbs/board.php?bo_table=b0301&wr_id=9
31) https://research.ibm.com/publications/watson-for-oncology-and-breast-cancer-treatment-recommendations-agreement-with-an-expert-multidisciplinary-tumor-board
32) https://blogs.sw.siemens.com/thought-leadership/2024/09/05/riches-to-rags-understanding-retrieval-augmented-generation/
33) https://www.projectpro.io/article/rag-use-cases-and-applications/1059#mcetoc_1ib25mn04c
34) https://vorecol.com/blogs/blog-the-role-of-artificial-intelligence-in-enhancing-knowledge-management-systems-168236

Siemens는 RAG 기술을 도입함으로써 방대한 데이터를 보다 쉽게 접근할 수 있도록 했다. RAG 시스템은 문서를 작은 의미 있는 조각으로 나누어 검색 쿼리와 비교하여 가장 관련성이 높은 정보를 선별해낸다. 이렇게 선별된 정보는 생성 AI 모델에 입력되어 사람이 이해하기 쉬운 형태로 요약된다. 이를 통해 직원들은 더 이상 정확한 키워드를 사용하지 않아도, 내용 기반의 자연어 검색을 통해 필요한 정보를 쉽게 찾을 수 있게 되었다.

이 시스템은 직원들이 중요한 정보를 찾는 데 소요되는 시간을 줄여 생산성을 향상시키고, 부서 간 협업을 촉진시켰다. 또한, 새로운 세대의 엔지니어와 기술자들에게 귀중한 지식을 전수하는 데 도움을 주며, 정보 접근성을 높이고, 이를 통해 기존 데이터의 가치를 극대화시켰다.

ⅳ. Bloomberg의 Law, Gevernment [35)36)37)]

Bloomberg은 RAG 기술을 활용하여 금융 보고서와 뉴스 요약에 AI 기반 도구를 사용하고 있다. 이 시스템은 방대한 양의 금융 데이터와 시장 뉴스를 신속하게 분석하여 간결한 요약본을 생성한다. Bloomberg Law와 Bloomberg Government는 RAG 기술을 통해 법률 및 정부 정책 정보를 더욱 효과적으로 처리하고 있다.

Bloomberg Law는 약 2억 건의 법원 기록, 820만 건의 실무 지침 문서, 1,550만 건의 법원 의견, 500만 건 이상의 성문법과 규정, 7,500만 건의 EDGAR(Electronic Data Gathering, Analysis, and Retrieval: 미국 증권거래위원회(SEC)가 운영하는 전자 데이터 수집, 분석 및 검색 시스템) 제출 문서 등 방대한 양의 법률 정보를 보유하고 있다. RAG 기술은 이러한 대량의 정보를 효율적으로 처리하여 사용자에게 가장 관련성 높은 정보를 제공한다.

Bloomberg Government는 RAG를 활용하여 연방 및 주 정책 동향을 추적하고, 정책에 영향을 미치는 데 도움을 주고 있다. 이 시스템은 텍스트 비교, 정보 추출, 요약, 검색 등의 기능을 통해 정부 관련 전문가들이 복잡한 정책 정보를 신속하게 이해하고 분석할 수 있도록 지원한다.

Bloomberg은 산업 표준 RAG 프레임워크와 자체 개발한 안전장치 서비스를 통해 AI 생성 콘텐츠가 법률 전문가가 작성한 원본 콘텐츠나 일차 자료에 근거하도록 보장한다. 이를 통해 최신 정보를 신속하게 제공하면서도 높은 정확도를 유지할 수 있게 되었다.

35) https://www.harrisonclarke.com/blog/harnessing-the-power-of-rag-for-content-creation-and-summarization
36) https://pro.bloomberglaw.com/about/our-approach-to-ai/#overview
37) https://about.bgov.com/about/our-approach-to-ai/#overview

v. 포스코 홀딩스의 리포팅 및 Q&A [38)39)40)]

포스코홀딩스는 구글 클라우드 기반으로 Gemini 1.5 pro 모델을 활용하여 생성형 AI 기반의 소재 기술/산업 동향 리포팅 시스템 및 지식 검색 Q&A 포털을 개발했다.

포스코그룹의 어플라이드 AI 리서치(Applied AI Research)팀은 다국어 뉴스 수집 및 번역, 랭킹 알고리즘 기반 뉴스 추천, 본문 요약, 국가별 소재 기술/산업 일간·주간 이슈 리포트 생성 및 이메일 발송까지 자동화된 프로세스로 운영되는 리포팅 시스템을 개발했다.

연구팀은 또한 지속적으로 축적되는 최신 뉴스와 관련 문서를 종합한 지식 검색 및 Q&A 시스템을 확장 구축했다. 이 시스템은 자연어 질문에 대해 90% 이상의 검색 및 답변 정확도로 정보와 인사이트를 제공한다. 현재 포스코홀딩스를 포함한 포스코그룹 내 10여 개 계열사에서 매일 500명 이상의 직원이 이용하고 있다.

포스코홀딩스의 이 시스템은 고성능 RAG 아키텍처와 문서 처리 AI를 기반으로 한 포스코그룹의 독자적인 지식 AI 에이전트이다. 특히 이 모델은 Gemini 모델과 고성능 RAG 기술을 결합하여 타사 LLM 대비 더 나은 성능과 효율성을 보이는 것으로 확인되었다.

이 시스템의 RAG 도입으로 검색 및 답변 정확도 향상, 개인화된 정보 제공, 업무 효율성 증대, 그리고 기업 내부 지식의 효과적인 활용이 가능해졌다. 또한, 최신 정보와 기존 지식을 결합함으로써 더욱 정확하고 맥락에 맞는 응답 생성이 가능해졌다. 이를 통해 포스코홀딩스는 AI 기술을 활용한 지식 관리와 의사결정 지원 시스템의 새로운 표준을 제시하고 있다.

vi. KB국민카드의 이벤트 Q&AI [41)42)]

KB국민카드는 스켈터랩스와 협력하여 LLM 기반의 '이벤트 Q&AI' 베타서비스를 출시했다. 이 서비스는 KB국민카드와 KB페이에서 운영하는 이벤트 정보를 자연스러운 대화 형식으로 찾아주는 질의응답 시스템이다. KB페이 모바일앱을 통해 로그인한 고객은 누구나 이 서비스를 이용할 수 있다.

38) https://cloud.google.com/customers/intl/ko-kr/posco-holdings?hl=ko
39) https://www.gttkorea.com/news/articleView.html?idxno=12569
40) https://www.digitaltoday.co.kr/news/articleView.html?idxno=528095
41) https://www.aitimes.kr/news/articleView.html?idxno=29202
42) https://zdnet.co.kr/view/?no=20231025102233

이벤트 Q&AI는 스켈터랩스의 LLM 기반 질의응답 챗봇 솔루션인 '벨라 큐나(BELLA QNA)'와 KB국민카드의 기업 이벤트 정보를 API로 연동하는 방식으로 구현되었다. 이 서비스는 RAG(Retrieval Augmented Generation) 방식을 채택하여 LLM의 할루시네이션을 최소화하고 답변의 정확도를 높였다.

이벤트 Q&AI는 고객이 상담원과 대화하듯 자연스러운 대화를 통해 진행 중인 이벤트의 상세 정보를 확인할 수 있도록 설계되었다. 서비스는 어렵고 전문적인 기술 용어 사용을 최소화하고, 추가 질문에 답변을 제공하며, 고객의 의도를 파악해 고객 중심의 대화를 이어갈 수 있도록 구현되었다.

KB국민카드는 이 서비스를 통해 매일 업데이트되는 수백 가지의 이벤트에 대한 정확한 정보를 전달하여 고객 경험을 혁신하고 강화할 예정이다. 이 서비스는 고객이 직접 검색을 통해 정보를 탐색해야 하는 불편함을 해소하고, 고객의 불편사항을 신속하게 해결할 수 있게 해준다.

RAG 방식의 도입으로 이벤트 Q&AI 서비스는 이벤트 정보를 융합하여 답변 생성 기능을 개선했다. 이를 통해 고객이 원하는 정보와 기업이 제공하고 싶은 정보의 간극을 줄일 수 있게 되었다. 또한, 범용 LLM 기반 챗봇의 한계를 극복하고 기업에 최적화된 답변을 제공할 수 있게 되었다.

H 산업별 RAG 적용 가능성

RAG 기술은 다양한 산업 분야에서 혁신적인 적용 가능성을 보여주고 있다. 이 기술은 방대한 데이터베이스에서 관련 정보를 검색하고 이를 바탕으로 맥락에 맞는 응답을 생성하는 능력을 가지고 있어, 여러 산업에서 획기적인 변화를 일으킬 잠재력을 지니고 있다.

금융 분야에서 RAG는 실시간 시장 분석과 투자 전략 수립에 중요한 역할을 할 수 있다. RAG 시스템은 최신 시장 데이터, 기업 재무 정보, 뉴스 기사를 실시간으로 검색하고 분석하여 투자자에게 맞춤형 투자 전략을 제시할 수 있다. 예를 들어, 특정 기업의 주식에 대한 투자 결정을 내릴 때, RAG 시스템은 해당 기업의 최근 재무 보고서, 관련 산업 동향, 경쟁사 정보, 그리고 글로벌 경제 지표 등을 종합적으로 분석하여 투자 위험과 기회를 평가할 수 있다. 또한, 다양한 금융 규제 문서와 시장 동향을 검색하여 잠재적 리스크를 식별하고 대응 전략을 수립하는 데 도움을 줄 수 있다. 이는 금융 기관들이 복잡한 규제 환경에서 컴플라이언스를 유지하면서도 효과적인 리스크 관리를 할 수 있게 해준다.

의료 분야에서 RAG는 진단 지원과 개인화된 치료 계획 수립에 활용될 수 있다. RAG 시스템은 환자의 증상을 입력받아 관련된 의학 문헌, 임상 사례, 최신 연구 결과를 검색하고 분석하여 의사의 진단을 보조할 수 있다. 예를 들어, 희귀 질환의 경우 RAG 시스템은 전 세계의 유사 사례와 최신 치료법을 신속하게 검색하여 의사에게 제공함으로써 정확한 진단과 효과적인 치료 방법 선택을 지원할 수 있다. 또한, 환자의 유전자 정보와 관련된 최신 연구 결과를 검색하여 개인화된 치료법을 제시할 수 있다. 이는 정밀 의료의 발전에 크게 기여할 수 있는 잠재력을 가지고 있다. 더불어, RAG 기술은 의료진의 지속적인 교육과 훈련에도 활용될 수 있다. 최신 의학 연구 결과와 임상 가이드라인을 실시간으로 제공함으로써 의료진이 최신 지식을 유지하고 최선의 치료를 제공할 수 있도록 지원할 수 있다.

법률 분야에서 RAG는 법률 문서 분석과 판례 검색에 효과적으로 활용될 수 있다. RAG 시스템은 계약서를 분석하고 관련 법규와 판례를 검색하여 잠재적 위험 요소를 식별할 수 있다. 예를 들어, 복잡한 국제 계약의 경우 RAG 시스템은 관련된 여러 국가의 법규와 판례를 검토하여 계약의 적법성과 잠재적 위험을 평가할 수 있다. 또한, 특정 법률 문제와 관련된 판례를 광범위하게 검색하고 요약하여 변호사의 사건 준비를 지원할 수 있다. 이는 법률 전문가들의 업무 효율성을 크게 향상시킬 수 있다. 더불어, RAG 기술은 법률 자문 서비스의 접근성을 높이는 데도 기여할 수 있다. 기본적인 법률 질문에 대해 RAG 기반의 챗봇이 관련 법규와 판례를 검색하여 신속하게 답변을 제공함으로써, 일반 시민들의 법률 정보 접근성을 높일 수 있다.

교육 분야에서 RAG는 개인화된 학습 경험과 실시간 정보 업데이트에 활용될 수 있다. RAG 시스템은 학생의 학습 진도와 선호도에 따라 관련 학습 자료를 검색하고 추천하여 맞춤형 학습 경로를 제공할 수 있다. 예를 들어, 특정 주제에 대해 어려움을 겪는 학생에게 RAG 시스템은 해당 주제와 관련된 다양한 설명 자료, 예제, 연습 문제 등을 제공하여 학습을 지원할 수 있다. 또한, 교과 내용과 관련된 최신 연구 결과나 시사 정보를 실시간으로 검색하여 학습 자료에 통합할 수 있다. 이는 학생들에게 더욱 풍부하고 최신의 학습 경험을 제공할 수 있다. 더불어, RAG 기술은 교사들의 수업 준비와 평가 과정에도 도움을 줄 수 있다. 최신 교육 방법론과 다양한 교육 자료를 제공하여 교사들이 더욱 효과적인 수업을 설계할 수 있도록 지원할 수 있다.

미디어 및 엔터테인먼트 분야에서 RAG는 콘텐츠 제작 지원과 개인화된 추천에 활용될 수 있다. RAG 시스템은 특정 주제나 장르에 관한 방대한 정보를 검색하고 분석하여 창작자에게 아이디어와 참고 자료를 제공할 수 있다. 예를 들어, 역사 드라마를 제작할 때 RAG 시스템은 관련 시대의 역사적 사실, 문화, 의복, 언어 등에 대한 상세한 정보를 제공하여 작품의 사실성과 깊이를 높일

수 있다. 또한, 사용자의 시청 이력과 최신 트렌드 정보를 결합하여 더욱 정교한 콘텐츠 추천을 제공할 수 있다. 이는 콘텐츠 제작의 질을 높이고 사용자 경험을 개선하는 데 기여할 수 있다. 더불어, RAG 기술은 실시간 엔터테인먼트 서비스에도 적용될 수 있다. 예를 들어, 라이브 스트리밍 중 시청자의 질문에 대해 RAG 시스템이 관련 정보를 실시간으로 검색하여 제공함으로써 더욱 풍부하고 상호작용적인 콘텐츠 경험을 만들어낼 수 있다.

RAG 기술은 각 산업 분야에서 방대한 정보를 효과적으로 검색하고 분석하여 의사결정을 지원하고 개인화된 서비스를 제공하는 데 큰 잠재력을 가지고 있다. 그러나 이러한 기술의 적용에는 몇 가지 중요한 고려사항이 있다. 데이터의 정확성과 신뢰성을 보장하는 것이 중요하며, 개인정보 보호에 대한 엄격한 준수가 필요하다. 예를 들어, 의료 분야에서 환자의 민감한 개인 정보를 다룰 때는 철저한 보안 조치가 필요하다. 또한, AI 기술의 윤리적 사용에 대한 지속적인 논의와 가이드라인 수립이 필요하다. 특히 법률이나 의료 분야에서 RAG 시스템의 판단이 인간 전문가의 결정을 완전히 대체하는 것이 아니라 보조하는 역할을 해야 한다는 점을 명확히 해야 한다.

RAG 기술의 성공적인 도입을 위해서는 각 산업 분야의 전문가들과 AI 기술자들 간의 긴밀한 협력이 필요하다. 이를 통해 각 분야의 특수성을 고려한 맞춤형 RAG 시스템을 개발하고 최적화할 수 있을 것이다. 또한, 사용자들의 피드백을 지속적으로 수집하고 반영하여 시스템의 성능을 개선하는 것도 중요하다. 이는 RAG 시스템이 실제 사용 환경에서 더욱 정확하고 유용한 결과를 제공할 수 있도록 하는 데 필수적이다.

4 결론

RAG 기술의 등장은 LLM의 한계를 극복하고 AI의 활용 범위를 확장하는 중요한 전환점이 되었다. RAG는 실시간 데이터를 활용하여 정확성과 최신성을 높이는 동시에, LLM의 강력한 언어 추론 및 생성 능력을 결합함으로써 AI 시스템의 성능을 한 단계 끌어올렸다.

RAG의 주요 장점은 다음과 같다. 첫째, 실시간 데이터 활용을 통해 최신 정보 기반의 응답을 생성할 수 있다. 둘째, 응답의 근거를 제시함으로써 신뢰성과 투명성이 향상된다. 셋째, 외부 지식을 활용하여 맥락 추론 능력이 강화된다. 넷째, 다양한 산업에 쉽게 적용 가능한 높은 확장성을 지닌다. 다섯째, 선택적 데이터 활용을 통해 리소스 효율성이 높다. 여섯째, 응답 속도 향상을 위한 효율적인 다양한 기법 사용이 가능하다. 일곱째, 다양한 데이터 소스 활용을 통해 편향성을 감소시킬 수 있다. 이러한 특성들은 LLM의 할루시네이션 문제를 해결하고, AI 시스템의 전반적인 성능 향상에 기여한다.

그러나 RAG 기술의 발전과 함께 몇 가지 중요한 한계와 과제도 존재한다. 첫째, RAG의 성능은 검색 품질에 크게 의존하므로, 부적절한 검색 결과는 전체 시스템의 성능을 저하시킬 수 있다. 둘째, 실시간 검색과 정보 처리로 인한 높은 계산 복잡도는 시스템의 효율성과 확장성에 제약을 줄 수 있다. 셋째, 외부 데이터 소스 활용 시 개인정보 보호 문제가 발생할 수 있어 이에 대한 신중한 접근이 필요하다. 넷째, 검색된 정보의 신뢰성과 편향성 문제는 LLM에 비해 향상되었지만 여전히 중요한 과제로 남아있다. 다섯째, 지식 베이스의 최신성과 정확성을 지속적으로 유지하는 것이 RAG 시스템의 성능을 위해 필수적이나, 이는 상당한 노력과 리소스를 요구한다.

RAG 기술은 AI의 새로운 패러다임을 제시하며, LLM의 한계를 극복하고 더욱 정확하고 신뢰할 수 있는 AI 시스템을 구축하는 데 핵심적인 역할을 할 것이다. RAG의 다양한 장점들을 통해 여러 산업 분야에서 혁신을 가속화할 수 있지만, 앞서 언급한 한계들을 극복하기 위해서는 지속적인 연구와 기술 개발이 필요하다. 앞으로 RAG의 기술적 진보와 함께 윤리적, 사회적 측면에서의 고려도 병행되어야 할 것이며, 궁극적으로 RAG 기술의 도입은 더 나은 사회, 더 나은 미래를 만드는 데 중요한 역할을 할 것으로 기대된다.

NIA

Digital Insight 2024

RAG 기술의 등장과 발전 동향

▌발　행 : 2024.12.6.

▌발행인 : 황종성

▌발행처 : 한국지능정보사회진흥원(NIA) 인공지능정책본부 미래전략팀

▌기획 및 문의 : 정현영 주임(hyeon0@nia.or.kr)

▌작　성 : 서강대학교 장나은 교수(zangne@sogang.ac.kr)