컴퓨터교육학회 논문지 2025년 제28권 제2호
https://doi.org/10.32431/kace.2025.28.2.007

# 생성형 AI를 위한 Retrieval-Augmented Generation (RAG) 기술 동향 및 전망

## Trends and Prospects of Retrieval-Augmented Generation (RAG) for Generative AI

윤여찬† · 김수균††
Yeochan Yoon† · Sookyun Kim††

**요약**

본 논문은 생성형 AI 분야에서 Retrieval-Augmented Generation(RAG) 기술의 발전과 응용 사례를 탐구한다. GPT-3와 GPT-4와 같은 대규모 언어 모델(LLM)이 뛰어난 언어 생성 능력을 보여주고 있지만, 사실성 부족과 실시간 지식 통합의 한계는 혁신적인 해결책을 필요로 한다. RAG는 검색 메커니즘과 생성 모델을 결합함으로써 이러한 문제를 해결하며, 외부 지식 소스를 실시간으로 참조하여 생성 결과의 신뢰성을 높인다. 논문은 RAG의 아키텍처를 심층적으로 분석하며, 검색 모듈, 증강된 컨텍스트 구성, 그리고 생성 메커니즘을 포함한 주요 구성 요소를 다룬다. 또한 BLEU, ROUGE, 사실성 평가와 같은 주요 평가 지표를 검토하고, 헬스케어, 교육, 금융과 같은 다양한 도메인에서의 실제 연구 사례를 조명한다. 이와 함께 지식 베이스의 신뢰성, 개인정보 보호 문제, 인프라 비용과 같은 도전 과제를 제시하고, 지식 그래프 및 멀티모달 데이터 통합과 같은 미래 연구 방향에 대한 통찰을 제공한다. 본 연구는 RAG의 강점, 한계, 그리고 잠재적 응용 가능성을 체계적으로 분석함으로써, 생성형 AI의 사실성, 신뢰성, 사용성을 향상시키기 위해 RAG 시스템을 도입하려는 연구자와 실무자들에게 실질적인 가이드를 제공하는 것을 목표로 한다.

**주제어** Retrieval-Augmented Generation, 생성형 AI, 대규모 언어 모델

**ABSTRACT**

This paper explores the technological advancements and applications of Retrieval-Augmented Generation (RAG) in the field of Generative AI. As large language models (LLMs) like GPT-3 and GPT-4 demonstrate remarkable language generation capabilities, their limitations in factual accuracy and real-time knowledge integration necessitate innovative solutions. RAG addresses these challenges by combining retrieval mechanisms with generation models, enabling real-time referencing of external knowledge sources and improving the reliability of generated content. The paper provides an in-depth analysis of RAG's architecture, encompassing retrieval modules, augmented context construction, and generation mechanisms. It also examines key evaluation metrics such as BLEU, ROUGE, and factuality assessments, while highlighting practical implementations across domains like healthcare, education, and finance. Furthermore, it identifies challenges such as knowledge base reliability, privacy issues, and infrastructure costs, offering insights into future research directions, including integration with knowledge graphs and multimodal data. By systematically analyzing RAG's strengths, limitations, and potential applications, this study aims to guide researchers and practitioners in adopting RAG systems to enhance generative AI's factuality, trustworthiness, and usability.

**Keywords** Retrieval-Augmented Generation, Generative AI, Large Language Models

†정회원 제주대학교 소프트웨어학부 인공지능전공 교수(교신저자)
††정회원 제주대학교 소프트웨어학부 컴퓨터공학 전공 교수

논문투고 2025년 01월 03일
심사완료 2025년 02월 18일
게재확정 2025년 02월 19일
발행일자 2025년 02월 28일

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-00245316).

## 1. 서론

인공지능(AI) 분야는 지난 수년간 딥러닝 기술의 발전과 대규모 언어 모델 연구의 급격한 성장에 힘입어 새로운 전기를 맞이했다. 특히 자연어를 처리하고 생성하는 능력을 크게 진전시킨 생성형 AI(Generative AI)는 대화형 챗봇, 자동 요약 시스템, 창의적 텍스트 생성 도구 등 다양한 애플리케이션으로 영역을 확장하며 주목받고 있다. 이처럼 생성형 AI 모델이 사회 전반에 두드러진 영향을 미치고 있음에도, 단순히 학습된 파라미터에 근거한 통계적 예측 방식은 지식 부족, 사실적 정보 전달의 어려움, 객관성 및 신뢰도 부족과 같은 문제를 야기한다. 위와 같은 문제들을 극복하기 위한 방안으로 Retrieval-Augmented Generation(이하 RAG)이 등장했다. RAG는 전통적인 정보 검색 기법을 생성형 모델에 결합하여, 모델이 외부 지식 소스를 역동적으로 참조하도록 유도한다. 이 과정을 통해 과거의 학습 시점 이후 새롭게 축적된 지식을 반영할 수 있으며, 잘못된 정보를 생성하는 현상을 완화하여 보다 정확한 답변과 텍스트를 생성할 수 있다는 점에서 주목할 가치가 있다.

본 논문은 생성형 AI의 한계를 보완하는 대표적인 해결책으로 언급되는 RAG의 기술적 특성과 동향을 면밀히 고찰하고, 이를 실제 서비스나 산업 현장에 적용할 때 고려해야 할 제반 이슈와 미래 연구 과제를 제시하는 데 목적이 있다. 구체적으로는 RAG의 기본 구조와 구성 요소가 무엇인지, 현재 RAG와 관련하여 어떠한 연구 및 기술이 주목받고 있으며 실제로 어디에서 어떠한 방식으로 응용되고 있는지, 그리고 RAG가 마주하고 있는 한계와 윤리적·산업적 과제는 무엇인지에 대한 세부적인 논의를 전개한다.

본 논문에서 다루는 내용은 대규모 언어 모델과 검색 기술을 결합하는 전략과 관련 구현 사례를 중점적으로 다룬다. 아울러 Hugging Face[1], LangChain[2] 등 주요 프레임워크와 라이브러리를 활용한 RAG 구현 방식을 살펴보고, 다양한 산업군에서 RAG가 어떠한 방식으로 활용되고 있는지 소개한다. 이후 RAG 시스템을 평가하는 데에 적합한 정량적·정성적 지표를 논의하며, 앞으로의 연구 방향과 과제를 제시함으로써 연구자와 산업 실무자의 이해를 높이고자 한다.

본 논문은 총 8장으로 구성되어 있으며, 2장에서는 생성형 AI와 대규모 언어 모델 전반의 이론적 배경을 살펴본다. 3장에서는 RAG 시스템의 전형적인 아키텍처와 이를 구성하는 Retrieval, Context Augmentation, Generation 세 단계의 핵심 기술을 설명한다. 4장에서는 최신 RAG 기술과 다양한 응용 사례를 소개하고, 5장에서는 정량·정성적 지표를 비롯한 RAG 시스템 평가 방법을 중점적으로 다룬다. 6장에서는 RAG 도입 과정에서 발생할 수 있는 문제점과 한계를 분석하고, 7장에서는 향후 발전 가능성과 지식 그래프, 멀티모달 데이터 등과의 결합 가능성을 제시한다. 8장에서는 앞선 논의 전반을 요약하고, RAG 기술이 갖는 의의와 연구 및 산업적 한계를 제시한다.

본 논문은 생성형 AI 분야에서 요구되는 사실성과 신뢰도를 높이기 위해 RAG를 어떻게 설계하고 활용할 수 있는지에 관해 종합적이고 체계적인 시야를 제공한다. 기존 연구의 경우 단순한 사실전달이나 정보정달을 중점적으로 하지만 본 논문의 경우 다양한 기술 사례와 연구 동향을 폭넓게 분석하고, 최신 연구에서 주목받고 있는 RAG 연구방향을 동시에 제시함에 있어 기여도가 있다. 결론적으로 본 논문이 RAG라는 개념을 이해하고 실제 환경에 적용하고자 하는 연구자와 실무자에게 유용한 안내서가 되기를 기대한다.

## 2. 이론적 배경

### 2.1 생성형 AI 개념 및 동향

생성형 AI(Generative AI)는 기계가 입력된 텍스트나 이미지를 단순 분석하는 데 그치지 않고, 새로운 데이터를 직접 만들어 내는 능력을 지향하는 기술로 정의된다. 이 개념은 과거부터 존재했으나, 딥러닝과 대규모 언어 모델(LLM)의 발전에 힘입어 최근 급격히 주목받고 있다. 생성형 AI의 대표적인 예시로는 GPT[3] 계열이 있다. GPT 모델은 초기에 언어 모델링 과제를 풀기 위해 설계되었지만, 거대한 데이터셋을 활용한 사전 학습(Pre-training)과 적절한 미세 조정(Fine-tuning)을 통해 높은 수준의 자연어 생성 능력을 보여 준다. 반면, BERT[4] 계열 모델은 애초에 양방향성(Bidirectional) 인코더 구조를 채택해 텍스트 이해에 강점을 보이도록 설계되었다. 따라서 BERT는 자연어 생성보다는 분류나 추론, 문장 유사도 계산 등에 더욱 효과적이며, 최근에는 BERT 기반 모델을 생성형 태스크에 접목하기 위한 다양한 변형 기법이 연구되고 있다.

생성형 AI의 비약적인 발전을 이끈 주요 동력 중 하나는 대규모 언어 모델의 진화이다. GPT-3, GPT-4와 같은 초거대 모델들은 파라미터 수와 학습 데이터의 확장을 통해 이전 세대 모델과 구분될 정도로 뛰어난 텍스트 이해력과 생성 능력을 확보했다. 이러한 LLM들은 문제 해결 능력이 향상되었으며, 번역, 요약, 글쓰기 지원, 교육용 챗봇 등 다양한 분야에서 폭넓게 활용되고 있다. 그러나 이처럼 우수한 성능에도 불구하고, 생성형 모델들은 여전히 몇 가지 문제점을 안고 있다. 가장 대표적인 예는 정확성과 사실성(Factuality) 부문에서 나타나는 취약점이다. 모델이 자연어 표현을 매우 유창하게 생성하더라도, 실제 지식과 상충되는 내용을 혼합해 결과를 도출하는 이른바 'Hallucination' 현상을 완전히 피하기는 쉽지 않다. 또한 최신 정보를 즉각적으로 반영하기 어려운 구조적 제약 때문에 빠르게 변화하는 지식 환경에서 활용할 경우 사실 기반 정보 부족 문제가 부각될 수 있다. 이와 같은 한계에 더해, 모델이 어떤 과정을 통해 특정 답변을 생성했는지 명확히 설명하기 어렵다는 점도 실무 현장에서 종종 우려되는 부분이다.

### 2.2 RAG의 개념

RAG는 생성형 AI가 지닌 한계를 보완하기 위해 고안된 개념으로, 생성 모델 내부의 통계적 패턴만으로 답변을 생성하는 방식에서 벗어나 외부 지식의 실시간 참조를 가능하게 한다. 즉, 모델이 질문에 답변하기 전에 대규모 문서나 데이터베이스에서 해당 질의와 관련성 높은 정보를 검색하고, 이러한 외부 지식을 바탕으로 텍스트를 생성하도록 설계한다. 이 접근법을 통해 특정 시점 이후에 발생한 새로운 사건에 대한 정보를 추출하거나, 모델이 학습하지 못한 전문 지식을 신속하게 보강할 수 있다는 장점이 부각된다.

RAG의 작동 방식을 이해하기 위해서는 크게 두 가지 요소인 정보 검색과 대규모 언어 모델의 결합 구조를 살펴볼 필요가 있다. 먼저 IR 기술은 전통적으로 TF-IDF[5]나 BM25[6] 같은 키워드 중심의 점수화 방식을 활용했으며, 최근에는 BERT, Sentence Transformers[7] 같은 딥러닝 기반 모델을 사용해 문장 임베딩을 구하고, 벡터 유사도 검색으로 관련 문서를 탐색하는 방식이 각광받고 있다. 이후 이렇게 찾아낸 문서(또는 문서의 일부)는 생성 모델에 입력될 맥락(Context)으로 제공되며, 모델은 질문과 함께 주어진 맥락을 통합해 최종 답변을 만든다.

RAG를 구현하는 접근 방식은 다양하다. 예컨대 “OpenAI + Vector Store”라는 구조는 OpenAI에서 제공하는 대규모 언어 모델 API를 호출하기 전에, 사용자가 자체적으로 구축한 벡터 스토어(Vector Store)에서 관련 문서를 검색해 모델에 제공하는 과정을 거친다. Retrieval이 끝난 뒤에는 Reranking 과정을 통해 가장 적합한 문서 또는 요약본을 최종적으로 선택하고, 이를 바탕으로 생성 단계에 돌입한다. Retrieval과 Reranking, Generation 세 단계를 순차적으로 진행하는 모델도 존재하며, 최적화 과정에서 피드백 루프나 별도의 필터링 모듈을 추가해 정확도를 높이는 방식도 시도되고 있다. 이 모든 접근 방식의 공통점은 단순히 내부 파라미터에 근거하지 않고, 외부 지식을 얼마만큼 효율적이고 신뢰도 높게 참조하느냐에 따라 생성 결과가 달라진다는 점이다.

결국 RAG는 대규모 언어 모델의 장점을 극대화하면서도, 최신 정보 반영과 정확성 보장을 위한 정보 검색 과정을 결합함으로써 기존 생성형 모델이 노출한 취약점을 보완할 수 있는 잠재력을 지닌다. 이러한 특성으로 인해 다양한 산업 분야에서 빠르게 도입 사례가 늘어나고 있으며, 연구 환경에서도 체계적인 성능 평가와 최적화 기법 개발을 통해 새로운 가능성을 모색하고 있다.

## 3. RAG 아키텍처 및 주요 구성 요소

### 3.1 Retrieval 모듈

RAG 시스템의 첫 번째 단계인 Retrieval 모듈은 모델이 외부 지식소스로부터 필요한 정보를 얻는 관문 역할을 수행한다. 생성형 모델은 학습된 파라미터에만 의존할 경우 지식의 최신성이나 정확성이 떨어질 수 있는데, Retrieval 모듈은 이를 보완하기 위해 대규모 문서 집합이나 데이터베이스에서 적절한 자료를 찾아낸다. 이러한 검색 과정은 크게 전통적인 키워드 기반 검색과 임베딩 기반 검색으로 나눌 수 있다.

전통적인 키워드 기반 검색은 BM25나 TF-IDF 방식을 활용하며, 문서의 단어 빈도와 역문서 빈도를 고려해 특정 질의어와의 매칭 점수를 계산한다. 이 방식은 구현이 단순하고 해석이 용이하다는 장점이 있지만, 질의어와 문서 간의 유사성이 단순 단어 매칭에 의존하기 때문에 의미적 관계를 충분히 반영하기 어렵다는 한계를 지닌다. 반면 임베딩 기반 검색(Vector Similarity Search)은 자연어 문장이나 단어를 벡터 형태로 변환한 뒤, 이 벡터 간의 유사도(코사인 유사도 등)를 측정하여 관련 문서를 검색하는 방식을 취한다. 이러한 임베딩은 BERT, Sentence Transformers 등 다양한 모델로부터 얻을 수 있으며, FAISS[8]나 Annoy[9] 같은 벡터 인덱싱 라이브러리를 통해 대규모 벡터 데이터를 효율적으로 검색할 수 있다.

검색을 통해 추출된 후보 문서가 많을 경우에는 Ranking

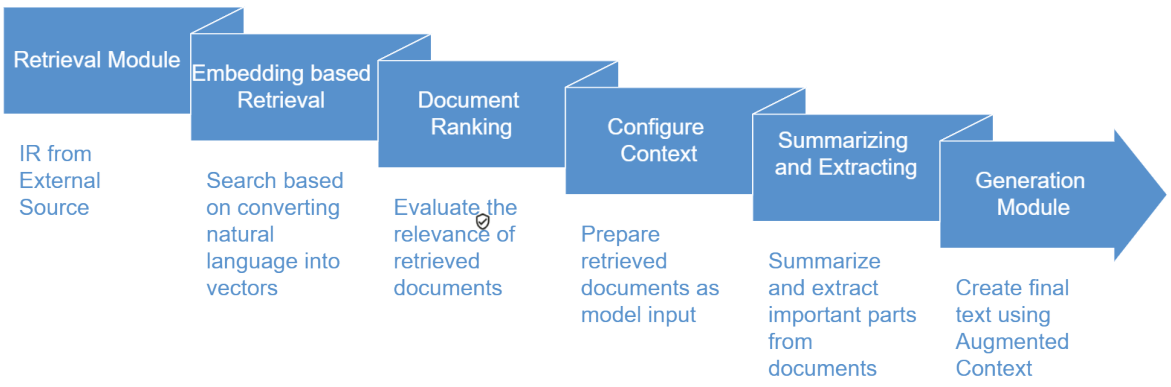


Figure 1. RAG 시스템 워크플로우

기법을 통해 가장 관련도가 높은 문서를 우선적으로 선별한다. Ranking은 검색된 결과를 질의 의도나 문맥에 비추어 재정렬하는 과정이며, 전통적인 확률 모델부터 학습 기반의 랭킹 모델까지 다양한 방법이 사용된다. Retrieval 단계에서의 정확도와 효율성은 곧 RAG의 전체 성능으로 직결되므로, 사용 환경과 데이터의 특성에 맞는 검색 알고리즘 및 인덱싱 전략을 선택하는 것이 중요하다.

### 3.2 Augmented Context 구성

Retrieval 모듈을 통해 선정된 문서는 곧바로 생성 모델에 투입되지 않고, Augmented Context라는 형태로 가공되어 주입된다. Augmented Context는 모델이 답변을 생성할 때 참고할 수 있도록, 검색된 문서나 문서 일부를 요약·추출·필터링한 결과물로 구성된다. 이 과정은 RAG 시스템의 품질을 결정짓는 핵심 요소 중 하나로, 생성 모델이 답변을 형성하는 데에 꼭 필요한 정보만 간결하고 효율적으로 제공해야 한다는 점에서 세심한 기획이 요구된다.

Augmented Context를 만드는 가장 직관적인 방식은 검색된 문서에서 연관된 단락이나 문장을 직접 추출한 뒤 모델에 전달하는 것이다. 특정 분야나 문서 구조가 정형화되어 있지 않을 경우, 이와 같은 방식으로 신속하게 RAG를 구축할 수 있다는 장점이 있다. 다만 검색 과정에서 비교적 긴 문서가 다수 추출될 경우에는 요약 기법을 통해 핵심 내용만을 간추려 모델에 넘기는 것이 효과적일 수 있다. 또한 민감 정보나 노이즈가 포함된 텍스트가 의도치 않게 모델에 전달될 수 있으므로, 이 단계에서 적절한 필터링 작업도 수행된다.

프롬프트 설계(프롬프트 엔지니어링)는 Augmented Context 구성과 밀접하게 연관되는 주제다. 생성 모델이 참고할 컨텍스트를 어떤 형식으로 주입하고, 질의나 답변 요구사항을 어떻게 설정하느냐에 따라 모델의 응답 품질이 크게 달라질 수 있다. 따라서 질의와 컨텍스트의 맥락을 최대한 자연스럽게 연결하고, 모델이 참조해야 할 정보를 명시적으로 제공하는 과정을 통해, 실제 사용자 의도를 더욱 정확하게 반영하는 답변을 도출할 수 있다.

### 3.3 Generation 모듈(생성 파트)

Generation 모듈은 앞서 구성된 Augmented Context를 활용해 최종 답변이나 텍스트를 생성한다. GPT 계열, T5[10] 계열, BERT 기반 생성 모델 등 다양한 대규모 언어 모델이 활용될 수 있으며, 최근에는 파인튜닝된 사내 전용 모델이나 최신 API를 사용하는 사례도 늘고 있다. 이 단계에서 모델은 Retrieval과 Augmented Context를 통해 획득한 정보와 자체적으로 학습된 지식을 종합해 답변을 생성하는데, 사실성(factuality)과 맥락 일관성을 최대한 유지하도록 설계하는 것이 중요하다.

생성된 문장은 별도의 후처리 과정을 거쳐 사용자에게 전달되거나, 시스템 내부에서 다른 모듈에 연계될 수 있다. 예컨대, 답변이 지나치게 장황하거나 문맥에서 벗어나는 내용이 포함되어 있을 경우에는 간단한 규칙 기반 필터나 추가 모델을 활용해 문장 길이를 조정하거나 오탈자를 교정한다. 오히려 생성 모델의 결과물을 인간이 직접 검토하는 Human-in-the-loop 프로세스를 도입해, 모델이 생성한 답변의 품질과 정확성을 높이는 전략을 적용하기도 한다.

Generation 모듈은 결국 사용자에게 전달되는 최종 산출물을 만들어내는 단계이기 때문에, 전체 RAG 시스템의 가치를 결정하는 매우 중요한 요소라고 볼 수 있다. Retrieval이나 Augmented Context 구성에서 부족함이 있어도, 이 단계에서 부분적으로 보완할 가능성이 있기 때문이다. 반면 Retrieval 과정에서 치명적인 정보 누락이 발생하거나, Augmented Context에 맞지 않는 질의가 들어오면 Generation 단계의 최적화만으로는 한계가 존재한다. 그렇기 때문에 RAG 시스템은 Retrieval, Augmented Context, Generation 세 단계가 긴밀히 연동되어야 최고의 성능을 발휘할 수 있다.

## 4. RAG 기술 및 연구동향

### 4.1 주요 RAG 프레임워크 및 라이브러리

RAG을 구현할 때에는 대규모 언어 모델과 검색 기술을 긴밀히 결합하는 과정이 필수적이다. 이를 효과적으로 지원하는 대표적인 라이브러리로 Hugging Face Transformers와 FAISS를 꼽을 수 있다. 먼저 Hugging Face Transformers는 다양한 사전 학습 언어 모델과 이를 활용하는 파이프라인을 손쉽게 구성할 수 있도록 하는 프레임워크로서, 텍스트 임베딩 생성부터 텍스트 생성에 이르는 폭넓은 기능을 제공한다. 생성형 모델과 함께 FAISS를 결합하면 대규모 벡터 데이터베이스를 효율적으로 검색할 수 있으며, 정확도와 처리 속도 측면에서 우수한 성능을 보이는 것이 특징이다. 특히 다차원 벡터 인덱싱 및 검색을 지원하므로 의미 기반 검색을 구현하는 데 적합하다는 장점이 부각된다.

LangChain과 LlamaIndex[11] 역시 RAG 구현을 손쉽게 지원하는 툴로 주목받고 있다. LangChain은 여러 프롬프트를 순차적으로 연결하여 복잡한 작업을 단계별로 처리하는 프롬프트 체이닝(Prompt Chaining) 개념을 도입하여, 하나의 멀티스텝 워크플로우 안에서 정보 검색과 생성 과정을 유연하게 연결해 준다. 이 과정에서 임베딩 모델, 스토리지 백엔드, 생성 모델 등을 모듈화하여 개발자 혹은 연구자가 원하는 파이프라인을 구성하기 쉽게 돕는 것이 LangChain의 큰 장점이다. 반면 LlamaIndex는 문서 데이터베이스를 쉽게 관리하고, 해당 데이터에서 필요한 정보를 검색하여 생성 모델로 전달하는 과정을 단순화함으로써 RAG 활용 진입 장벽을 낮춘다는 의의를 가진다.

### 4.2 산업 분야별 최신 연구동향

RAG 기술은 다양한 산업 분야에서 활용되며, 특히 지식 기반이 방대하거나 최신성을 요구하는 업무에서 가치를 발휘한다. 의료 분야에서 RAG는 대규모 언어 모델의 한계를 보완하기 위해 외부 데이터베이스에서 정보를 검색해 응답을 생성하는 기술이다. 이 기술은 특히 의료 데이터의 복잡성과 전문성을 처리하는 데 유용하며, 최근 다양한 연구를 통해 그 가능성이 입증되고 있다.

#### 4.2.1 의료분야 RAG 연구동향

2024년 1월 발표된 Ke의 연구[12]에서는 수술 전 의학 정보를 다루는 LLM-RAG 모델이 개발되었다. 이 모델은 35개의 수술 전 지침을 데이터로 활용했으며, 생성된 응답의 정확도가 인간 전문가의 답변을 초과하는 91.4%로 평가되었다. 평균 응답 시간은 15~20초로, 실시간 의사결정을 지원하는 데 유용하다는 결과가 나왔다.

Zhu의 연구[13]에서는 멀티모달 전자의무기록(EHR) 데이터를 분석하는 REALM 프레임워크가 제안되었다. 이 프레임워크는 임상 노트와 시계열 데이터를 통합하여 예측 성능을 높였으며, 외부 지식 그래프를 활용해 의료 문맥을 강화했다. 이를 통해 입원 중 사망률과 30일 재입원율 예측에서 우수한 성능을 보였다. Zhu의 다른 연구[14]에서는 RAG를 활용하여 전자의무기록의 다중 모드 데이터를 통합적으로 분석하고 예측 모델링의 성능을 향상시키는 방법을 제안하였다. 텍스트, 이미지, 구조화된 데이터를 결합하여 의료 예측의 정확도와 효율성을 극대화했다. 실험 결과, 제안된 EMERGE 모델은 기존 방법 대비 우수한 성능을 보이며 의료 데이터 분석의 새로운 가능성을 입증했다.

제약 산업에서는 규제 준수 프로세스를 지원하기 위해 QA-RAG 모델[15]이 제안되었다. 이 모델은 사용자 질문에 대해 관련 문서를 검색하고 신뢰할 수 있는 답변을 제공하도록 설계되었다. 기존 방법보다 높은 정확도를 기록했으며, 규제 환경에서의 실질적인 활용 가능성을 보여주었다.

Xiong의 연구[16]에서는 의료 분야에서 RAG 시스템의 성능을 평가하기 위해 7,663개의 질문을 포함한 MIRAGE 벤치마크를 제안하였다. 이를 통해 다양한 조합의 코퍼스, 검색기, 언어 모델을 실험하여, RAG가 GPT-3.5와 Mixtral의 성능을 GPT-4 수준으로 향상시킬 수 있음을 보였다.

Long[17]은 의료 분야에 최적화된 RAG 프레임워크인 Bailicai를 소개하였다. 이 프레임워크는 LLM의 성능을 향상시키며, '환각' 문제를 완화하는 데 효과적이었다.

MedGraphRAG[18]는 그래프 기반 RAG 프레임워크를 도입하여, 의료 LLM의 신뢰성과 안전성을 높이는 방법을 제안하였다.

Ngo[19]의 연구에서는 LLM이 의료 질문 응답 시나리오를 처리하는 능력을 평가하기 위한 MedRGB 벤치마크를 소개하였다.

Miao[20]는 KDIGO 2023 가이드라인을 기반으로 만성 신장 질환 관리에 특화된 RAG 시스템을 개발했다. 이 시스템은 질문에 따라 관련 가이드라인 문서를 검색하고, 정확한 답변을 생성하도록 설계되었다. 임상 시나리오에서 92.3%의 응답 정확도를 기록하며 기존 시스템을 능가하는 성능을 보였다. 연구는 신장학 외 다양한 의료 분야에서 RAG의 잠재력을 보여줬다.

#### 4.2.2 금융분야 RAG 연구동향

금융 분야에서도 RAG가 점차 도입되고 있다. 기업들은 고객 응대 FAQ 시스템이나 시장 리포트 자동 생성 도구에 RAG 개념을 적용함으로써, 수시로 갱신되는 증권 분석 리포트나 경제 지표 관련 정보를 지속적으로 업데이트하고 투자자 혹은 내부 직원에게 배포한다. 이런 방식은 복잡하고 방대한 데이터베이스를 기반으로 한 정확한 응답이 필수적인 금융 산업 특성과 부합한다는 장점이 있다.

Wang[21]은 금융 도메인에서 RAG 시스템을 평가하기 위한 벤치마크인 OmniEval을 제안하였다. OmniEval은 쿼리를 다섯 가지 작업 클래스와 16가지 금융 주제로 분류하여 다양한 시나리오에서의 RAG 성능을 체계적으로 평가한다. 또한 GPT-4 기반의 자동 생성과 인간 주석을 결합하여 평가 데이터의 신뢰성을 높였다. 연구는 금융 분야에서 RAG 시스템의 성능 변화와 수직 도메인에서의 능력 향상 가능성을 강조한다.

Li[22]는 금융 분석을 위한 AlphaFin 데이터셋을 공개하고, Retrieval-Augmented Stock-Chain 프레임워크를 제안하였다. AlphaFin 데이터셋은 전통적인 연구 데이터셋, 실시간 금융 데이터, 그리고 사고의 연쇄(Chain-of-Thought) 데이터를 결합하여 LLM의 금융 분석 능력을 높이는 데 기여하며 RAG 기술을 통합하여 금융 분석 작업에서 성능을 개선하는 방법을 제시한다.

Kang[23]은 금융 작업에서 LLM의 환각(hallucination) 문제를 실증적으로 조사한다. 금융 개념 설명, 역사적 주가 조회 등의 작업에서 LLM이 환각 현상을 보이는 사례를 제시하며, 이를 완화하기 위한 방법으로 소수 샷 학습(few-shot learning), DoLa(Decoding by Contrasting Layers), RAG, 프롬프트 기반 도구 학습 등을 평가하였다. 연구는 금융 작업에서 LLM의 환각 문제가 심각하며 이를 해결하기 위한 추가 연구의 필요성을 강조한다.

Settiy[24]의 논문은 금융 문서에서 검색 증강 생성(RAG) 모델의 성능을 개선하기 위한 기법을 제안한다. 주요 방법으로 텍스트 청킹 최적화, 쿼리 확장, 메타데이터 추가, 재랭킹 알고리즘, 임베딩 미세 조정을 제시한다. 이를 통해 LLM 기반 금융 질문 응답 시스템의 정확성과 신뢰성을 향상시킨다.

임재훈[25]이 제안한 논문은 AutoRAG 프레임워크를 활용해 금융 문서에 최적화된 검색 증강 생성(RAG) 시스

템을 구현하는 연구를 다룬다. AutoRAG는 RAG 시스템 구축 과정을 자동화하여 최적의 파이프라인을 탐색하며, Advanced RAG가 Naive RAG보다 우수한 성능을 보임을 확인했다. 이를 통해 금융 도메인에서 RAG 시스템의 효율성과 정확성을 향상시키는 방법을 제시한다.

Yepes[26]은 금융 보고서에서 검색 증강 생성(RAG)의 효율성을 높이기 위해 문서를 구조적 요소별로 분할하는 새로운 청킹(chunking) 방법을 제안한다. 전통적인 문단 단위 분할 대신, 문서 이해 모델을 통해 주석된 요소 유형에 따라 분할하여 정보 검색의 정확성과 문맥성을 향상시킨다. 연구 결과, 이러한 요소 기반 청킹이 금융 보고서에서 RAG의 성능을 크게 개선함을 확인하였다.

Zhao[27]는 금융 감성 분석에서 대형 언어 모델(LLM)의 성능을 향상시키기 위해, 인간의 지시를 반영한 Instruction Tuning과 주식 시장의 피드백을 활용한 강화 학습을 통해 모델을 조정하는 방법을 제안한다. 이를 통해 LLM이 텍스트 데이터의 내재된 감성을 더 정확하게 파악하여, 기존 최첨단 모델 대비 정확도와 F1 점수가 1%에서 6% 향상되었으며, 주가 움직임 예측에서도 우수한 성과를 보였다. 또한, 이러한 감성 분석 결과를 기반으로 한 포트폴리오는 상승장에서는 S&P 500 대비 샤프 지수가 3.61% 높았고, 하락장에서는 손실이 5배 감소하는 등 시장 상황에 따른 탄력성을 입증하였다.

#### 4.2.3 교육분야 RAG 연구동향

교육 현장에서도 RAG는 Q&A 시스템이나 과제 피드백 도구로 응용되고 있다. 학습자가 질문을 제시하면, RAG 모델이 교과서, 학습 자료, 논문, 온라인 백과사전 등을 검색해 관련 근거를 추출하고, 이를 토대로 답변을 구성한다. 이를 통해 교사 또는 교수자가 제공하는 일방적인 정답 외에도, 다각적인 관점이나 심층 정보를 자동으로 제공할 수 있다는 점이 주목받고 있다.

Miladi[28]의 연구는 RAG 기술을 활용하여 대규모 공개 온라인 강좌(MOOCs) 환경에서 GPT 모델의 정확성을 높이는 방법을 조사한다. 연구 결과, GPT-3.5 단독 사용 시 60%였던 정확도가 RAG를 통합한 GPT-4에서는 80%까지 향상되었으며, 이는 교육 콘텐츠 생성에서 RAG 기술이 가진 잠재력을 보여준다.

Henkel[29]은 RAG를 활용하여 수학 질문-답변 시스템의 응답 품질을 향상시키는 방안을 다룬다. 고품질 오픈소스 수학 교재를 기반으로, 실제 학생 질문에 대한 RAG 응답의 효과를 평가한 결과, RAG가 응답 품질을 높이는 데 유효했으나, 교육 자원과의 정확한 일치와 학생 선호 간의 균형을 고려해야 함을 강조했다.

Manathunga[30]은 의학교육 분야에서 RAG를 적용하는 방법을 제안하였다. 연구진은 대표 벡터를 활용한 추출적 및 요약적 방식의 결합 기법을 제안하며, 이를 통해 대규모 비정형 텍스트 데이터를 요약하면서 언어 모델의 환각 문제와 유해 응답 생성을 줄이는 데 중점을 두었다.

OwlMentor[31]는 대학생들이 과학 논문을 더 잘 이해하도록 돕기 위해 RAG 기반 학습 플랫폼으로 이를 대학 강의에 적용해 학습 효과와 사용 참여도를 평가했다. 연구 결과, RAG 시스템이 학습자의 과학 텍스트 이해와 몰입도를 높이는 데 효과적임을 보여줬으며, 학습자가 기술을 수용하는 과정에서 발생하는 변화도 관찰했다. 하지만 AI 도구가 효과적인 학습을 위해 검증된 교육적 전략에 기반해야 한다는 점도 강조했다.

HiTA[32]는 RAG 기반 교육 플랫폼으로, 교육자를 중심으로 한 AI 지원 학습 시스템이다. HiTA는 교육자의 전문성을 강화하고, 학생들에게는 교육자 감독 하에 맞춤형 학습 지원을 제공한다. 콜로라도 광업대학((Colorado School of Mines)에서 6명의 교육자와 400명의 학생을 대상으로 한 두 학기 동안의 파일럿 연구 결과, 97% 이상의 학생들이 HiTA를 유용하다고 평가했으며, 80% 이상이 ChatGPT보다 더 도움이 된다고 응답했다.

Modran[33]은 RAG 접근법을 활용한 지능형 챗봇 튜터링 시스템을 개발하여, 전통적인 튜터링의 한계와 일반적인 LLM(대규모 언어 모델)의 단점을 극복하고자 했다. 이 시스템은 학습자의 이해도와 참여도를 높이기 위해 정확하고 맥락에 맞는 맞춤형 지원을 제공하며, 학습자와의 상호작용을 통해 지속적으로 성능을 향상시킨다. 이를 통해 대학생들의 학습 경험을 풍부하게 하고, 개인 맞춤형 학습을 촉진하며, 학습 참여도와 성과를 향상시키는 것을 목표로 한다.

#### 4.2.4 기업분야 RAG 연구동향

비즈니스 분야에서는 문서 요약과 고객 응대 챗봇이 두드러진 사례로 손꼽힌다. 전사적으로 보유한 방대한 문서(매뉴얼, 기획안, 계약서 등)를 신속히 검색하고 요약해 주는 시스템을 RAG로 구현하면 지식 관리가 한결 효율적으로 이루어진다. 또한 전자상거래나 고객센터 챗봇에 RAG 접근법을 접목할 경우, 제품 설명서나 사용자 리뷰와 같은 외부 문서를 토대로 답변을 제공함으로써 고객 만족도를 높일 수 있다. 이광우[34]는 생성형 AI의 부정확한 정보 제공과 정보 유출 우려를 해결하기 위해, 한국어 문장 임베딩을 활용한 지식 데이터베이스를 구축하고 최적화된 검색을 통해 관련 정보를 제공하는 RAG 기반 질의응답 시스템을 설계하였다. 이를 통해 기업 내부 정보의 효율적 업데이트와 기밀 정보 보호를 강화하였다.

정천수[35]의 논문은 LangChain 프레임워크를 이용하여 LLM 애플리케이션 아키텍처를 설계하고, RAG 모델을 통해 기업 내부 데이터의 활용을 극대화하는 생성형 AI 서비스 구현 방법을 제시하였다. 이를 통해 정보 부족 문제를 극복하고, 유사 문맥 추천 및 질의응답 시스템의 성능을 향상시켰다.

Bruckhaus[36]는 RAG 기술을 기업 환경에 적용할 때 데이터 보안, 정확성, 확장성 등의 문제로 인해 한계가 발생함을 지적하였다. 이를 해결하기 위해 의미적 검색, 하

이브리드 쿼리, 최적화된 검색 기법 등의 발전이 필요하며, 평가 프레임워크를 통해 기업 맞춤형 RAG 솔루션의 유효성을 검증할 것을 제안하였다.

Purwar[37]는 오픈소스 LLM을 활용한 RAG 시스템이 기업 특화 데이터셋에서 얼마나 효과적인지 비교 분석하였다. 연구 결과, 오픈소스 LLM과 효율적인 임베딩 기법을 결합하면 RAG 시스템의 정확도와 효율성을 크게 향상시킬 수 있음을 발견하였다.

Raina[38]는 기업 환경에서 RAG의 검색 정확도를 높이기 위해 문서를 원자 단위로 분해하고, 생성된 질문을 활용하여 관련 정보를 검색하는 방법을 제안하였다. 이를 통해 검색 단계에서의 정확도를 높여, 최종 응답의 품질을 향상시켰다.

Ahmad[39]는 다문화 기업 환경에서 RAG 모델을 활용하여 다국어 정보 검색을 개선하는 방법을 제시한다. 다양한 언어와 문해력 수준을 고려하여 데이터 제공 전략, 최신 정보 반영, 환각 현상 완화, 오류 응답 방지, 전달 속도 최적화 등의 요소를 통합하였다. 이를 통해 다문화 조직 내에서 정보의 신속하고 정확한 전달을 실현하고자 한다.

## 5. 성능 평가 방법

### 5.1 객관적 평가 지표

RAG 시스템을 평가하는 대표적인 방법은 자연어 생성 모델 분야에서 널리 사용되는 객관적 지표를 활용하는 것이다. 우선 BLEU[40], ROUGE[41], METEOR[42]와 같은 전통적 지표를 들 수 있는데, 이들은 모델이 생성해 낸 텍스트와 기준 정답(Reference) 간의 유사도를 통계적으로 측정하여, 어휘나 구문 단위에서의 일치 정도를 수치화한다. 예컨대 BLEU는 기계 번역에서 시작된 지표로서 n-gram 단위의 정합성을 중심으로 하며, ROUGE는 요약 과제에서 주요 개념어의 회수나 매칭 정도를 보는 방식이 특징적이다. METEOR는 어휘 동의어와 어간 처리 등을 추가로 고려해 조금 더 유연하게 표현 간 유사도를 판단한다.

이 밖에도 정답이 명확하게 정의되어 있는 과제(예: 특정 질문에 대한 답안)에 대해서는 정확도(Precision), 재현율(Recall), F1-score 같은 지표가 활용된다. 예를 들어 정보 검색 과정에서 문서 후보를 선정하는 Retrieval 단계의 성능을 확인하기 위해, 검색된 문서가 실제로 유효 정답을 포함하고 있는지에 대한 정확도나 재현율을 측정할 수 있다. 나아가 질문응답(QA) 실험이나 사실성(Factuality)이 중요한 시나리오에서는 모델의 답변이 정확히 근거 문서에 부합하는지를 정밀하게 평가하기 위해, 라벨링된 데이터셋을 기준으로 한 F1-score 계산이 사용되기도 한다. 이처럼 객관적 지표들은 모델 개발 과정에서 빠르게 성능 변화를 추적하고, 서로 다른 시스템 간의 우열을 비교하기에 유용하나, 실제 사용자의 체감 만족도나 장문의 응답 품질을 모두 반영하기에는 한계가 있다.

생성된 응답과 근거 데이터 간의 정합성을 평가하기 위해, FEQA(Factual Error QA)와 FactCC(Factual Consistency Classifier)와 같은 도구를 활용할 수 있다. FEQA는 생성된 문장을 질문-응답 방식으로 변환하여 사실성을 평가하며[43], FactCC는 문장쌍 간의 정합성을 분류하는 모델로서 사용된다[44].

### 5.2 주관적/인간 평가 지표

모델이 생성한 결과물의 실제 사용성이나 맥락적 자연스러움을 평가하기 위해서는 주관적 지표나 인간 평가가 필수적이다. 생성형 AI에서 흔히 거론되는 문제 중 하나는 '환각(Hallucination)' 현상으로, 외부 근거 문서에 존재하지 않는 정보를 마치 사실인 양 생성하는 경우가 이에 해당한다. 따라서 사람이 직접 모델의 답변을 점검하여 얼마나 사실에 부합하는지를 확인하는 '사실성(Factuality)' 평가가 중요하며, 의료나 법률, 금융처럼 오류 용납도가 낮은 분야에서는 반드시 수행해야 할 프로세스다.

추가로 모델이 생성한 텍스트가 사용자 입장에서 얼마나 쓰기 편하고 자연스러운지에 대한 주관적 평가는 설문조사나 사용자 피드백 인터뷰 형태로 진행된다. 이 과정에서 답변의 맥락 일관성(Context Coherence), 동일 문장이나 의미가 불필요하게 반복되는지 여부, 문체가 사용자에게 적절하고 이해하기 쉬운 방식으로 구성되어 있는지 등을 점검하게 된다. 실제 서비스 현장에서는 사용자들에게 일정 기간 동안 시스템을 사용하도록 한 뒤, 만족도 점수나 자유 양식의 의견을 받아 모델 개선 방향을 수립하는 경우가 많다.

### 5.3 성능 향상 전략

RAG 시스템의 성능을 높이기 위해서는 Retrieval 단계와 Generation 단계를 통합적으로 개선하는 전략이 필요하다. 먼저 Negative examples(거짓 정보)를 충분히 포함한 데이터셋을 구성해, 모델이 허위 사실을 출력하는 빈도를 낮추는 방향으로 학습 혹은 파인튜닝하는 방안을 생각해 볼 수 있다[45]. Retrieval 과정에서 노이즈가 심한 문서나 신뢰도가 낮은 자료가 선정되지 않도록 필터링 기법을 적용하고, Retrieval 결과물을 다시 한 번 Reranking하거나 검증하는 모듈을 추가하는 것도 한 방법이다[46].

Generation 단계에서는 파인튜닝된 언어 모델의 하이퍼파라미터를 조정하거나, 지식 그래프(Knowledge Graph) 같은 구조화된 데이터를 결합해 답변의 정확성을 높이는 시도가 이루어진다[47]. 모델이 참조하는 외부 정보를 체계화해, 특정 개념이나 관계를 명시적으로 주입함으로써 문맥적 혼동이나 사실 오류를 줄일 수 있다. 상황에 따라서는 Retrieval과 Generation 단계가 분리되어 있기보다, 두 모듈이 상호작용하면서 학습되도록 엔드투엔드 파이프라인을 구축하는 방안도 모색된다[48]. 다만 엔드투엔드 접근법은 구현 복잡도가 높고 대규모 연산 자원을 필요로 하므로, 시스템 규모나 도입 목적에 맞추어 최적의 전략을 도입하는 것이 중요하다.

## 6. 문제점 및 한계

### 6.1 지식 베이스의 신뢰도 및 지속적 업데이트 문제

RAG 모델이 의존하는 핵심 자원 중 하나는 외부 지식 베이스다. 이 지식 베이스가 최신 정보로 지속적으로 갱신되지 않으면, 모델은 시대착오적인 내용을 바탕으로 답변을 생성하거나 최근 사건에 대한 질의를 적절히 처리하지 못하는 문제가 발생한다. 예를 들어 특정 시점 이후에 발표된 새로운 학술 논문이나 업데이트된 규정이 반영되지 않은 상태에서 모델이 질의에 응답할 경우, 의도치 않게 잘못된 정보를 제공하게 된다. 더욱이 지식 베이스 내 문서가 오래되거나 특정 관점을 편향적으로 담고 있는 경우, 모델은 해당 오류를 그대로 답변에 반영할 가능성이 높아진다. 결국 사용자는 RAG 시스템이 최신 지식에 근거해 정확하고 공정한 정보를 제공해 줄 것이라고 기대하지만, 실제로는 미갱신된 데이터 혹은 편향된 자료 때문에 유의미한 오류가 발생할 수 있다.

이 같은 문제를 완화하기 위해서는 데이터 수집 및 관리 정책이 체계적으로 마련되어야 한다. 실시간(또는 주기적)으로 신뢰도 높은 소스에서 지식을 자동 혹은 반자동으로 업데이트하고, 갱신된 내용이 검색 과정에서 우선적으로 반영될 수 있도록 인덱싱 재생성을 진행해야 한다. 또한 권위 있는 자료인지, 혹은 여러 관점을 균형 있게 포함하고 있는지 등을 검토하여 지식 베이스의 품질을 지속적으로 모니터링하는 작업이 뒤따라야 한다.

### 6.2 프라이버시 및 보안 이슈

RAG 시스템은 검색 단계에서 방대한 문서를 다루기 때문에, 검색 대상 데이터에 민감하거나 개인정보가 포함되어 있을 가능성이 있다. 예를 들어 헬스케어 분야에서 환자 관련 기록을 활용하는 시스템이라면, 환자 정보가 쉽게 노출될 위험을 내포한다. 이는 검색 허용 범위와 접근 권한을 어떻게 설정하는지, 그리고 생성된 응답 중에서 민감 정보가 무방비로 노출되지는 않는지 등을 엄격히 관리해야 함을 의미한다.

특히 클라우드 기반으로 RAG 시스템을 운영하는 경우, 대규모 인덱스와 언어 모델을 외부 서버에서 호스팅하는 일이 많아지면서 보안 문제가 더욱 중요해진다. 검색 쿼리 및 결과가 오가면서 발생할 수 있는 정보 유출, 혹은 보안 취약점을 악용한 공격 사례가 보고되는 상황에서, 개인정보 보호법 혹은 산업별 규제(GDPR, HIPAA 등)를 준수하기 위한 별도의 솔루션이 요구된다. 이를 위해 검색 접근 제어, 데이터 암호화, 민감 문서 필터링 등을 포함하는 전반적인 보안 아키텍처 설계가 필수적이다.

### 6.3 해석 가능성(Explainability) 및 투명성(Transparency)

RAG 시스템은 Retrieval 단계와 Generation 단계를 결합해 최종 답변을 생성하기 때문에, 어떤 문서가 어떻게 검색되었고 어떤 방식으로 답변에 반영되었는지를 설명하기가 쉽지 않다. 특히 대규모 언어 모델이 내부 파라미터를 바탕으로 단어 시퀀스를 생성하는 과정은 '블랙박스'에 가깝다는 평가를 받고 있으며, 검색 모듈 역시 의미 벡터를 이용한 매칭 방식을 적용할 경우 사용자 입장에서 구체적인 매칭 이유를 파악하기가 까다롭다.

이런 상황에서 해석 가능성과 투명성을 높이기 위해서는, Retrieval 단계에서 어떤 문서(또는 문서의 특정 부분)가 선택되었는지와 그 근거 점수를 간략히 제시해 주거나, 모델이 생성 과정에서 어떤 맥락 정보를 활용했는지를 추적할 수 있는 로깅(logging)·비주얼라이제이션 기법을 도입하는 편이 좋다. 예를 들어 시스템이 "왜 이 문단을 참조해 답변을 도출했는가?"에 관한 근거를 사용자에게 제공하면, 모델에 대한 신뢰도가 높아지고, 필요 시 시스템 개선 방향을 잡기에도 용이하다[49].

### 6.4 고비용/고성능 인프라 의존 문제

RAG 시스템은 대규모 언어 모델뿐만 아니라 대규모 인덱스(벡터 스토어, 텍스트 검색 인덱스 등) 역시 실시간으로 운용해야 하기 때문에, 막대한 연산 자원과 스토리지, 네트워크 트래픽을 요구하는 경우가 많다. 이는 스타트업이나 소규모 연구팀이 RAG를 대규모로 구축하기 어려운 장벽이 되기도 하며, 이미 운영 중인 기업이라도 클라우드 비용이나 GPU 클러스터 구축 비용이 급증할 위험이 있다.

또한 대규모 언어 모델이 실행되는 동안 응답 지연(latency)이 길어지거나, 벡터 검색 과정에서 인덱스가 제대로 최적화되어 있지 않으면 쿼리 처리 속도가 현저히 떨어지는 상황이 발생할 수 있다. 이는 실시간 서비스가 중요한 도메인에서 사용자 경험을 저하시킬 우려가 크다. 이러한 문제를 해결하기 위해서는 효율적인 인덱싱 알고리즘(예: FAISS, Annoy, ScaNN[50] 등)과 분산 처리 기술을 도입해야 하며, 고성능 모델의 양자화(quantization)[51]나 지식 증류(knowledge distillation)[52]를 활용해 경량화 모델을 확보하려는 노력도 병행되어야 한다.

## 7. 기술발전 전망

RAG 기술은 생성형 AI의 한계를 보완하기 위해 출현했으며, 이미 여러 산업 분야에서 실무적으로 의미 있는 성공 사례를 만들어 내고 있다. 그러나 기술적·윤리적 문제를 완전히 해결하기 위해서는 다양한 연구와 실험이 앞으로도 지속적으로 이루어져야 한다. 특히 지식 그래프와 멀티모달 데이터의 결합, 시맨틱 서치 기술의 고도화, 개인화된 사용자 경험 제공, 그리고 거짓 정보 최소화와 윤리적 이슈 대응 전략 마련 등이 차세대 RAG의 핵심 방향성으로 꼽힌다.

가장 먼저 지식 그래프(Knowledge Graph)와의 결합[47] 가능성이 주목된다. 텍스트 중심의 문서 검색을 통해 얻은 정보만으로는 사실 관계나 개념 간 연관성을 완벽히 파악하기 어려운 경우가 많다. 예컨대 의학 분야에서 질병,

증상, 약물, 상호 금기사항 등은 구조화된 그래프 형태로 관계를 맺고 있으며, 이를 단순 텍스트 검색보다 효율적으로 관리할 수 있다. RAG 모델이 이러한 지식 그래프를 직접 참조할 수 있도록 설계하면, 검색된 문서를 기반으로 한 텍스트생성 과정에서 보다 정확하고 체계적인 근거를 활용하게 된다. 이는 결과물의 사실성을 높이고, 잘못된 정보가 포함될 확률을 줄이며, 향후에는 대규모 언어 모델이 그래프 상의 노드와 에지 정보를 바탕으로 추론 과정을 좀 더 명시적으로 전개하도록 돕는 역할을 할 수 있다.

다음으로 멀티모달(Multimodal) RAG[53]에 대한 기대도 높아지고 있다. 지금까지의 RAG 연구는 주로 텍스트 문서를 검색하고, 텍스트 기반 대화나 답변을 생성하는 데 집중되어 왔다. 그러나 실제 응용 시나리오에서는 이미지, 음성, 비디오, 표(스프레드시트) 등 텍스트 외에도 다양하고 복합적인 형태의 데이터가 중요한 역할을 차지한다. 예컨대 상품 추천 서비스를 예로 들면, 사용자 후기가 텍스트로 제공되는 동시에 제품 사진이나 사용 동영상이 함께 제시될 수 있다. 멀티모달 RAG는 이러한 비정형 데이터를 검색하고, 해당 정보를 언어 모델의 맥락에 자연스럽게 통합함으로써 보다 풍부하고 설득력 있는 응답을 만들어 낼 가능성을 열어 준다.

시맨틱 서치 기술[54]의 발전도 차세대 RAG 구현에서 큰 비중을 차지할 것으로 보인다. 전통적인 키워드 기반 검색이나 단순 임베딩 매칭을 넘어, LLM과 결합된 고도화된 임베딩을 활용하거나 메타데이터와 문서 속 구조적 정보를 정교하게 반영하는 방식이 제안되고 있다. 이를 통해 검색 정확도를 높이고, Retrieval 단계에서 후보 문서가 지나치게 많아지는 문제를 완화할 수 있으며, 궁극적으로는 Generation 단계에서 더욱 정밀한 근거를 활용하게 된다. 이러한 시맨틱 서치 기술은 지식 그래프와 함께 사용될 수도 있으며, 대규모 인덱스를 다룰 때 발생하는 속도 저하 문제와 결합되어 다양한 최적화 기법 연구로 이어질 것으로 전망된다.

사용자 맞춤형 개인화[55] 역시 빠질 수 없는 미래 과제이다. 개인별 성향, 지식 수준, 과거 대화 기록 등을 RAG에 반영하는 방식을 통해, 동일한 질문이라도 사용자 맥락에 맞추어 최적화된 답변을 제시할 수 있다는 가능성이 제기된다. 예컨대 헬스케어 분야에서 사용자의 건강 상태나 이력, 식습관 정보를 참고하여 의료 정보 검색 및 안내 수준을 조절하거나, 교육 분야에서 학생별 학습 수준과 취약점을 인식하고 맞춤형 피드백을 제공하는 식의 응용이 가능하다. 이를 위해서는 개인정보 보호 문제와 안전 장치 마련이 병행되어야 한다는 점이 중요한 고려 사항이 될 것이다.

마지막으로 거짓 정보를 최소화하고 오남용을 방지하기 위한 지속적인 품질 관리와 윤리적 이슈 해결 방안이 필수적으로 논의되어야 한다. RAG가 접속하는 외부 지식 소스에 부정확하거나 편향된 내용이 포함되어 있으면, 검색 단계와 생성 단계가 아무리 발전되어 있다 해도 오류를 전파하게 되기 쉽다. 이를 해소하기 위해서는 신뢰할 수 있는 출처로 지식을 엄선하고, 일관적인 검증(verification) 과정을 도입하며, 결과물을 인간이 검수하는 Human-in-the-loop[56] 방식을 적용하는 등의 적극적인 조치가 필요하다. 또한 인종, 성별, 문화적 편향이 시스템에 내재되지 않도록 주의 깊은 점검과 개선 노력이 지속되어야 하며, 이를 지원하기 위한 정책과 법·제도적 장치 역시 발전이 요구된다.

이상과 같은 미래 지향점은 서로 긴밀히 연결되어 있으며, RAG가 학계와 산업계 전반에 확고히 자리매김하기 위해서는 여러 측면에서의 통합적인 노력이 뒤따라야 한다. 지식 그래프, 멀티모달, 시맨틱 서치, 맞춤형 개인화, 윤리적·법적 장치 등은 단독으로도 흥미로운 연구 주제지만, RAG의 실효성을 획기적으로 높이는 열쇠가 되리라는 점에서 앞으로 더욱 주목받을 것으로 보인다. 궁극적으로는 신뢰성 높은 외부 지식과 창의적 언어 생성 능력을 결합한 차세대 RAG 시스템이, 폭넓은 사용자 집단에게 진정한 가치를 제공하는 새로운 지평을 열어 갈 것이라는 기대가 크다.

## 8. 결론

본 논문은 생성형 AI(Generative AI)의 한계를 보완하기 위한 기술적 해법으로 대두된 Retrieval-Augmented Generation(이하 RAG)의 개념과 동향을 살펴보고, 이를 구현·적용하는 과정에서 발생하는 다양한 문제와 해결 전략을 조망하였다. 먼저 RAG의 등장 배경으로, 대규모 언어 모델(LLM)이 보여 주는 뛰어난 언어 생성 능력에도 불구하고 최신 정보 반영과 사실 기반 답변에 취약할 수 있다는 점이 지적되었다. 외부 검색 모듈을 생성 모델에 결합함으로써 이러한 문제를 해결하려는 RAG 개념은 이미 정보 검색(IR) 분야에서 축적된 기술적 자산을 적극적으로 활용하며, 산업 현장과 학계에서 빠르게 주목받고 있다.

논문에서는 RAG를 구성하는 Retrieval, Augmented Context, Generation 모듈의 세부 아키텍처를 정리하고, Hugging Face Transformers + FAISS, LangChain, LlamaIndex 등 주요 프레임워크를 통해 구현되는 흐름을 확인하였다. 아울러 헬스케어, 금융, 교육, 일반 비즈니스 분야에서 실제 적용 사례가 증가하고 있음을 언급하며, 이를 뒷받침하는 연구 논문 동향도 간단히 소개했다. 전통적인 자연어 생성 평가 지표인 BLEU, ROUGE, METEOR에 더해 사실성(Factuality), 사용자 만족도, 문체 적절성 등을 포함한 주관적 평가가 병행되어야 한다는 점과, Retrieval 및 Generation 모듈 간 긴밀한 최적화를 통해 거짓 정보와 노이즈를 최소화하는 것이 중요한 과제라는 점도 확인하였다.

RAG는 생성형 AI가 단순히 사전 학습 파라미터에 의존하지 않고, 외부 지식 소스를 능동적으로 활용한다는 점에서 의의와 가치를 지닌다. 해당 기술이 발전함에 따라, 지식 그래프나 멀티모달 데이터와의 결합, 시맨틱 서치의 고도화, 사용자 맞춤형 개인화, 윤리적 이슈 해소 등과 같은 다양한 주제에서 혁신적인 응용 가능성이 열리고 있다. 그러나 본

논문에서 제시한 구조나 사례가 모든 상황에 보편적으로 적용될 수 있는지는 여전히 한계가 있으며, 특정 분야나 데이터셋 특성에 따라 다른 접근법이 필요할 수 있음을 인정해야 한다. 또한 최신 정보를 실시간으로 반영하는 문제, 민감한 문서나 개인정보를 다루는 과정에서 발생하는 보안·프라이버시 우려, 모델 자체의 투명성과 해석 가능성 부족 등은 여전히 해결해야 할 과제로 남아 있다.

향후 연구에서는 다양한 사용자 집단과 데이터 환경을 고려하여 RAG를 실제 서비스로 확장하는 실증 연구가 필요하다. 산업 현장에서의 도입 사례를 분석해, 구체적인 성능 지표와 비용·효율 측면에서의 이점을 비교 검토하는 후속 연구가 이루어진다면, RAG 기술의 활용 가치를 더욱 명확히 입증할 수 있을 것이다. 또한 윤리적·법적 규제와 맞물린 문제는 단순히 기술 개발만으로는 해결하기 어려운 측면이 있어, 관련 분야 전문가와의 협업과 거버넌스 체계 수립이 필수적이다. 정리하자면, RAG는 생성형 AI가 처한 근본적인 한계를 효율적으로 보완할 수 있는 잠재력을 갖추고 있으며, 올바른 데이터 관리와 책임 있는 기술 운영 체계가 마련된다면 더욱 폭넓은 영역에서 혁신적인 가치 창출에 기여할 것으로 기대된다.

## 참고문헌


[1] Hugging Face. (n.d.). *Transformers: State-of-the-art natural language processing for PyTorch*, TensorFlow, and JAX. Retrieved from https://huggingface.co

[2] LangChain. (n.d.). LangChain: *Building applications with LLMs through composability*. Retrieved from https://github.com/hwchase17/langchain

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1(Long and Short Papers),* 4171-4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[5] Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, 18*(11), 613-620. https://doi.org/10.1145/361219.361220

[6] Robertson, S., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129-146. https://doi.org/10.1002/asi.4630270302

[7] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 3982-3992. https://doi.org/10.18653/v1/D19-1410

[8] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data,* 7(3), 535-547. https://doi.org/10.1109/TBDATA.2019.2921572

[9] Spotify. (n.d.). *Annoy: Approximate Nearest Neighbors in C++/Python*. Retrieved from https://github.com/spotify/annoy

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

[11] LlamaIndex. (n.d.). *LlamaIndex: A data framework for LLM applications*. Retrieved from https://github.com/jerryjliu/llama_index

[12] Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., ... & Ting, D. S. W. (2024). Development and Testing of Retrieval Augmented Generation in Large Language Models--A Case Study Report. *arXiv preprint arXiv:2402.01733*. https://doi.org/10.48550/arXiv.2402.01733

[13] Zhu, Y., Ren, C., Xie, S., Liu, S., Ji, H., Wang, Z., ... & Pan, C. (2024). REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. *arXiv preprint arXiv:2402.07016*. https://doi.org/10.48550/arXiv.2402.07016

[14] Zhu, Y., Ren, C., Wang, Z., Zheng, X., Xie, S., Feng, J., ... & Pan, C. (2024). EMERGE: Integrating RAG for Improved Multimodal EHR Predictive Modeling. *arXiv preprint arXiv:2406.00036*. https://doi.org/10.48550/arXiv.2406.00036

[15] Kim, J., & Min, M. (2024). From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process. *arXiv preprint arXiv:2402.01717*. https://doi.org/10.48550/arXiv.2402.01717

[16] Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*. https://doi.org/10.48550/arXiv.2402.13178

[17] Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., & Ting, D. S. W. (2024). Development and testing of retrieval augmented generation in large language models: A case study report. *arXiv preprint, arXiv:2402.01733*. https://doi.org/10.48550/arXiv.2402.01733

[18] Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., & Grau, V. (2024). Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*. https://doi.org/10.48550/arXiv.2408.04187

[19] Ngo, N. T., Van Nguyen, C., Dernoncourt, F., & Nguyen, T. H. (2024). Comprehensive and Practical Evaluation of Retrieval-Augmented Generation Systems for Medical

Question Answering. *arXiv preprint arXiv:2411.09213.* https://doi.org/10.48550/arXiv.2411.09213

[20] Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O. A., & Cheungpasitporn, W. (2024). Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. *Medicina, 60*(3), 445. https://doi.org/10.3390/medicina60030445

[21] Wang, S., Tan, J., Dou, Z., & Wen, J. R. (2024). OmniEval: An Omnidirectional and Automatic RAG Evaluation Benchmark in Financial Domain. *arXiv preprint arXiv:2412.13018.* https://doi.org/10.48550/arXiv.2412.13018

[22] Li, X., Li, Z., Shi, C., Xu, Y., Du, Q., Tan, M., Huang, J., & Lin, W. (2024). AlphaFin: Benchmarking financial analysis with retrieval-augmented Stock-Chain framework. *arXiv preprint, arXiv:2403.12582.* https://doi.org/10.48550/arXiv.2403.12582

[23] Kang, H., & Liu, X.-Y. (2023). Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint, arXiv:2311.15548*. https://doi.org/10.48550/arXiv.2311.15548

[24] Setty, S., Thakkar, H., Lee, A., Chung, E., & Vidra, N. (2024). Improving retrieval for RAG-based question answering models on financial documents. *arXiv preprint, arXiv:2404.07221.* https://doi.org/10.48550/arXiv.2404.07221

[25] Lim, J.-H., & Suh, J.-W. (2024). A Study on implementing the most optimized RAG system for financial document using AutoRAG. *Annual Conference of KIPS*, 521–522. https://doi.org/10.3745/PKIPS.Y2024M10A.521

[26] Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024). Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*. https://doi.org/10.48550/arXiv.2402.05131

[27] Zhao, Z., & Welsch, R. E. (2024). Aligning LLMs with Human Instructions and Stock Market Feedback in Financial Sentiment Analysis. *arXiv preprint arXiv:2410.14926*. https://doi.org/10.48550/arXiv.2410.14926

[28] Miladi, F., Psyché, V., & Lemire, D. (2024). Leveraging GPT-4 for accuracy in education: A comparative study on retrieval-augmented generation in MOOCs. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky (Vol. 2150). Communications in Computer and Information Science. Springer,* Cham. https://doi.org/10.1007/978-3-031-64315-6_40

[29] Henkel, O., Levonian, Z., Li, C., & Postle, M. (2024). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *Proceedings of the 17th International Conference on Educational Data Mining,* 315–320. https://doi.org/10.5281/zenodo.12729824

[30] Manathunga, S. S., & Illangasekara, Y. A. (2023). Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education. *arXiv preprint arXiv:2308.00479.* https://doi.org/10.48550/arXiv.2308.00479

[31] Thüs, D., Malone, S., & Brünken, R. (2024). Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature. *Frontiers in Psychology, 15,* 1474892. https://doi.org/10.3389/fpsyg.2024.1474892

[32] Liu, C., Hoang, L., Stolman, A., & Wu, B. (2024, July). HiTA: A RAG-Based Educational Platform that Centers Educators in the Instructional Loop. *In International Conference on Artificial Intelligence in Education* (pp. 405-412). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_37

[33] Modran, H., Bogdan, I. C., Ursutiu, D., Samoila, C., & Modran, P.-L. (2024). LLM intelligent agent tutoring in higher education courses using a RAG approach. *Preprints*. https://doi.org/10.20944/preprints202407.0519.v1

[34] Yi, G.-W., & Kim, S. K. (2024). Design of a question-answering system based on RAG model for domestic companies. *Journal of the Korea Society of Computer and Information, 29*(7), 81–88. https://doi.org/10.9708/jksci.2024.29.07.081

[35] Jeong, C. (2023). Generative AI service implementation using LLM application architecture: Based on RAG model and LangChain framework. *Journal of Intelligence and Information Systems, 29*(4), 129–164. https://doi.org/10.13088/jiis.2023.29.4.129

[36] Bruckhaus, T. (2024). RAG Does Not Work for Enterprises. *arXiv preprint arXiv:2406.04369.* https://doi.org/10.48550/arXiv.2406.04369

[37] Purwar, A. (2024). Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability. *arXiv preprint arXiv:2406.11424*. https://doi.org/10.48550/arXiv.2406.11424

[38] Raina, V., & Gales, M. (2024). Question-based retrieval using atomic units for enterprise RAG. *arXiv preprint, arXiv:2405.12363.* https://doi.org/10.48550/arXiv.2405.12363

[39] Ahmad, S. R. (2024). Enhancing Multilingual Information Retrieval in Mixed Human Resources Environments: A RAG Model Implementation for Multicultural Enterprise. *arXiv preprint arXiv:2401.01511.* https://doi.org/10.48550/arXiv.2401.01511

[40] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* 311–318. https://doi.org/10.3115/1073083.1073135

[41] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop,* 74–81. https://aclanthology.org/W04-1013/.

[42] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization,* 65–72. https://aclanthology.org/W05-0909/.

[43] Durmus, E., He, H., & Diab, M. (2020). FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL),* 5055-5070. https://doi.org/10.18653/v1/2020.acl-main.454

[44] Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the Factual Consistency of Abstractive Text Summarization. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 9332-9346. https://doi.org/10.18653/v1/2020.emnlp-main.750

[45] Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., & Silvestri, F. (2024). The power of noise: Redefining retrieval for RAG systems. *In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 719–729). Washington, DC, USA. ACM. https://doi.org/10.1145/3626772.3657834

[46] Dong, J., Fatemi, B., Perozzi, B., Yang, L. F., & Tsitsulin, A. (2024). Don't Forget to Connect! Improving RAG with Graph-based Reranking. *arXiv preprint arXiv:2405.18414.* https://doi.org/10.48550/arXiv.2405.18414

[47] Hussien, M. M., Melo, A. N., Ballardini, A. L., Maldonado, C. S., Izquierdo, R., & Sotelo, M. Á. (2025). RAG-based explainable prediction of road users' behaviors for automated driving using knowledge graphs and large language models. *Expert Systems with Applications, 265,* 125914. https://doi.org/10.1016/j.eswa.2024.125914

[48] Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics, 11,* 1-17. https://doi.org/10.1162/tacl_a_00530

[49] Singhal, R., Patwa, P., Patwa, P., Chadha, A., & Das, A. (2024). Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. *arXiv preprint arXiv:2408.12060.* https://doi.org/10.48550/arXiv.2408.12060

[50] Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., & Kumar, S. (2019). Accelerating large-scale inference with anisotropic vector quantization. *arXiv preprint arXiv:1908. 10396.* https://doi.org/10.48550/arXiv.1908.10396

[51] Lang, J., Guo, Z., & Huang, S. (2024). A Comprehensive Study on Quantization Techniques for Large Language Models. *arXiv preprint arXiv:2411.02530.* https://doi.org/10.48550/arXiv.2411.02530

[52] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531.* https://doi.org/10.48550/arXiv.1503.02531

[53] Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., ... & Yao, H. (2024). Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085.* https://doi.org/10.48550/arXiv.2410.13085

[54] Sawarkar, K., Mangal, A., & Solanki, S. (2024). Blended RAG: Improving RAG (Retriever-Augmented Generation) accuracy with semantic search and hybrid query-based retrievers. *In Proceedings of the 7th IEEE International Conference on Multimedia Information Processing and Retrieval.* IEEE. https://doi.org/10.48550/arXiv.2404.07220

[55] Wang, H., Huang, W., Deng, Y., Wang, R., Wang, Z., Wang, Y., ... & Wong, K. F. (2024). Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv preprint arXiv:2401.13256.* https://doi.org/10.48550/arXiv.2401.13256

[56] Afzal, A., Kowsik, A., Fani, R., & Matthes, F. (2024). Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human in the Loop. *arXiv preprint arXiv:2407.05925.* https://doi.org/10.48550/arXiv.2407.05925

윤여찬

· 2004년 고려대학교 컴퓨터학과 졸업(학사)
· 2007년 고려대학교 컴퓨터학과 졸업(석사)
· 2020년 고려대학교 컴퓨터학과 졸업(박사)
· 2007~2022년 한국전자통신연구원 책임연구원
· 2022년 ~ 현재 제주대학교 인공지능전공 조교수

관심분야 : 자연어처리, 멀티모달 딥러닝
ycyoon@jejunu.ac.kr

김수균

· 2006년 고려대학교 컴퓨터학과 졸업(박사)
· 2006~2008년 삼성전자 책임연구원
· 2008~2020년 배재대학교 게임공학과
· 2020년 ~ 현재 제주대학교 컴퓨터공학전공 교수

관심분야 : 그래픽스, 컴퓨터비전
kimsk@jejunu.ac.kr