JDFR Journal of Data Forensics Research
데이터포렌식연구, 1권 1호, 2024년 12월, 47-62면
Journal of Data Forensics Research, Volume 1, Issue 1, December 2024, Pages 47-62.
https://doi.org/10.12972/JDFR.2024.1.1.4
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

수사 지원을 위한 거대언어모델 및 검색증강생성 기반 대화형 인공지능 에이전트 시스템

김지형[1], 김민수[1], 계효선[2], 최원기[1], 이승우[3], 고재진[4], 이상신[5]

[1]한국전자기술연구원 자율지능시스템연구센터 선임연구원
[2]한국전자기술연구원 자율지능시스템연구센터 전임연구원
[3]한국전자기술연구원 자율지능시스템연구센터 책임연구원
[4]한국전자기술연구원 융합시스템연구본부 본부장
[5]한국전자기술연구원 자율지능시스템연구센터 센터장

교신저자: 이상신, sslee@keti.re.kr

요약

범죄수사는 다양한 법령에 대한 깊은 이해와 높은 수준의 전문성을 요구하는 분야이다. 이에 따라 수사 현장에서 관련 지식을 바탕으로 효과적인 질의응답 서비스를 제공할 수 있는 대화형 인공지능 에이전트에 대한 필요성이 증가하고 있다. 본 연구에서는 Large Language Model (LLM)과 Retrieval-Augmented Generation (RAG)을 활용한 대화형 에이전트 시스템의 참조 구조와 성능 분석을 위한 벤치마크를 제안하였다. 이를 통해 수사 분야에 특화된 LLM 기반 인공지능 시스템 개발에 실질적인 지침을 제공하고자 한다. 또한, GPT-4o, EXAONE 3.0 등 최신 모델들을 적용하여 성능을 분석하였고, 실제 현장에서 활용 가능한 인공지능 시스템의 개발 가능성을 확인하였다.

주제어

인공지능, 거대언어모델, 검색증강생성, 수사지원, 대화형 에이전트

Received: November 27, 2024
Revised: December 16, 2024
Accepted: December 23, 2024
Published: December 31, 2024

Original Article

A Conversational AI Agent System Based on Large Language Models and Retrieval-Augmented Generation to Support Investigation

Jeehyeong Kim[1], Minsu Kim[1], Hyoseon Kye[2], Won-Gi Choi[1], Seungwoo Lee[3], Jaejin Ko[4], Sang-Shin Lee[5]

[1]Senior Research Fellow, Autonomous Intelligence System Research Center, Korea Electronics Technology Institute, Republic of Korea
[2]Researcher, Autonomous Intelligence System Research Center, Korea Electronics Technology Institute, Republic of Korea
[3]Principal Research Fellow, Autonomous Intelligence System Research Center, Korea Electronics Technology Institute, Republic of Korea
[4]Vice President of Convergence Systems Research Center, Korea Electronics Technology Institute, Republic of Korea
[5]Center Director, Autonomous Intelligence System Research Center, Korea Electronics Technology Institute, Republic of Korea

Corresponding Author: Sang-Shin Lee, sslee@keti.re.kr

ABSTRACT

Criminal investigation is a field that requires a high level of expertise and a thorough understanding of various legal frameworks. Thus, there is a growing need for an artificial intelligence (AI) agent service that can effectively provide knowledge-based Q&A support within investigative contexts. This study proposes a reference architecture and benchmark for evaluating the performance of an AI chatbot system utilizing Large Language Models (LLM) and Retrieval-Augmented Generation (RAG). Through this, we offer practical guidelines for developing LLM-based AI systems specifically tailored to investigative work. Additionally, by applying and testing recent models such as GPT-4o and EXAONE 3.0, we demonstrate the feasibility of developing practical, field-ready AI systems for this domain.

KEYWORDS

Artificial Intelligence, Large Language Model, Retrieval-Augmented Generation, Supporting Investigation, Conversational Agent

1. 서론

최근 범죄 수법이 고도화되고, 수사 과정에서 디지털 증거의 양이 방대해지고 유형이 다양해짐에 따라 수사 지원을 위한 인공지능 기술이 주목받고 있다 [1]. 거대언어모델(LLM, Large Language Model)의 성능이 비약적으로 향상됨에 따라 대량의 자연어 및 멀티모달 정보를 빠르게 분석할 수 있게 되었으며, 이러한 기술을 적용한 수사 지원에 대한 수요가 증가하고 있다. 특히 다양한 디지털 증거와 멀티모달 데이터를 신속하게 분석하여 수사에 직접 활용하거나 증거를 찾아내는 분야가 주목받고 있다. 또한, 여러 매뉴얼과 관련 법령을 분석하여 현장 관계자를 보조하며 질의에 대한 응답을 통해 의사결정을 지원하는 대화형 에이전트 서비스에 대한 필요성도 커지고 있다 [2-4]. 새로운 범죄 유형이 등장하고, 디지털 기술의 발전에 따라 범죄 수법이 복잡해지면서, 경찰은 과거보다 다양한 분야에 걸쳐 전문적 지식이 요구되는 수사를 수행해야 하는 상황에 놓여 있다. 이에 따라 특정 전문 분야에서 수사를 담당하는 특별사법경찰제도의 중요성도 커지고 있다. 특별사법경찰제도는 형사소송법 제197조에 따라 전문 지식이 요구되는 분야에서 행정 공무원이 일반 사법경찰을 대신해 사법경찰권을 행사할 수 있도록 한 제도이다 [5]. 특별사법경찰은 식품, 환경, 위생, 지식재산권, 건설 등 다양한 행정 분야에서 주요 수사 권한을 행사하며, 범죄 예방 및 대응에서 핵심적인 역할을 담당하고 있다. 그러나 특별사법경찰은 수사에 대한 방대한 법령과 복잡한 규정, 다양한 매뉴얼을 신속히 숙지해야 하는 부담이 크다. 각 분야별로 적용되는 법령과 규정이 복잡하게 얽혀 있어 다양한 상황에 맞춰 정확히 해석하고 적용해야 하며, 관련 매뉴얼의 내용이 광범위해 추가적인 교육과 지식이 요구된다.

새로운 유형의 범죄가 발생할 때 이를 신속히 이해하고 대응하는 데도 어려움이 있어 현장 대응력을 강화하기 위한 지원이 필요하다. 이러한 상황에서 LLM 기반 인공지능 기술은 법령 해석, 수사 절차 관련 질의응답 등을 통해 특별사법경찰의 업무를 효과적으로 지원할 수 있는 중요한 도구가 될 수 있다. LLM은 높은 언어 이해력과 질의응답 기능을 바탕으로 사용자의 질문을 이해하고, 관련 정보를 신속히 찾거나 요약하여 제공할 수 있다 [6]. 이를 통해 특별사법경찰이 방대한 규정과 절차를 효과적으로 참고하고, 수사 과정에서 필요한 지식을 즉각적으로 얻을 수 있도록 지원할 수 있다. 특별사법경찰을 위한 수사 지원 도구로 LLM이 활용되기 위해서는 일반적인 자연어 처리 능력 외에도 법률 지식과 수사 용어 이해 등 수사 분야에 특화된 전문성이 요구된다. 이를 위해 검색증강생성(RAG, Retrieval Augmented Generation) 기술이 적극적으로 적용될 수 있다. RAG는 LLM이 사용자 질의에 대한 응답을 생성할 때 외부 지식베이스를 참조함으로써 최신성과 정확성을 보장할 수 있다 [7]. 특별사법경찰을 위한 LLM 기반 수사 지원 서비스에서 RAG가 필수적인 이유는 크게 세 가지이다. 첫째, 범죄 수사에는 피해자 신상 등 공개가 어려운 민감 정보가 포함될 수 있는데, RAG를 활용하면 LLM이 민감 정보를 학습하지 않고도 지식베이스 참조를 통해 신뢰성 있는 응답을 생성할 수 있다. 둘째, LLM이 학습을 통해 최신 법령과 판례를 모두 반영하는 것은 매우 어렵다. 필요한 정보를 데이터베이스화하여 LLM이 질의응답 시 이를 참조함으로써 최신 정보가 반영된 응답을 제공할 수 있다. 마지막 이유는 환각(Hallucination)의 예방이다. LLM은 언어 모델 특성상 사실이 아닌 텍스트를 생성할 가능성이 있으며, 이는 수사지원이라는 분야에서 치명적으로 작용할 수 있다. RAG는 이러한 환각을 예방할 수 있도록 도와줄 수 있다.

LLM과 RAG를 결합하여 수사지원 분야를 위한 인공지능 시스템 구현에는 몇 가지 기술적인 어려움이 존재한다. 수사 및 법률 지식은 일반적인 자연어와는 달리 전문 용어 위주로 구성되어 있으며, 상대적으로 복잡한 맥락을 가지고 있다. 따라서 RAG 성능에 따라 환각 현상으로 인해 LLM의 응답에서 중요한 법적 맥락의 누락, 응답 근거에 대한 신뢰성 저하 등 수사 지식 보완 측

면에서 문제가 발생할 수 있다. RAG에 수사 실무 매뉴얼, 실제 수사 사례 등 다양한 데이터를 저장해서 활용한다고 해도 LLM의 잘못된 해석과 확장 추론으로 인해 응답의 품질이 낮아질 수 있으며, RAG를 통해 검색된 참조 문서의 내용 간에 중복되는 부분이나 충돌하는 부분이 있는 경우, LLM이 이를 적절히 조정하고 통합된 관점에서 일관성 있는 응답을 생성해야 하는 요구사항이 있다.

본 논문에서는 수사 지원을 위한 거대언어모델 및 검색증강생성 기반 대화형 인공지능 에이전트 시스템에 대한 참조구조와 성능평가를 위한 벤치마크를 제안한다. 해당 대화형 에이전트 시스템은 다양한 기술들이 복합적으로 고려되어야 하며, 주어진 배경지식 및 질의 유형에 따라 어떤 설계가 최적의 성능을 도출할지 예측하기가 어렵다. 따라서 기본적인 시스템 구조인 참조구조와, 이러한 시스템의 성능을 평가할 질의응답 벤치마크 데이터셋 구축이 필요하다. 본 연구에서는 이를 통해 LLM 및 RAG를 활용한 수사지원 대화형 에이전트 시스템을 구현하고 성능을 분석하여, 실제 현장에서 활용가능한 수준의 인공지능 시스템에 대한 개발 가능성을 확인하고자 한다.

2. 관련 연구

2.1. Large Language Model (LLM)

LLM은 방대한 텍스트 데이터셋을 학습하여 언어를 이해하고 생성하는 인공지능 모델로, 자연어 처리(NLP, Natural Language Processing) 및 다양한 언어 기반 작업에서 탁월한 성능을 발휘한다. 초기 NLP 모델은 한정된 매개변수로 특정 작업에 특화된 성능을 보여주었으나, LLM은 수십억에서 수조 개의 매개변수를 활용하여 방대한 데이터에서 다양한 언어 패턴을 학습할 수 있게 되었다. 이러한 발전의 기반에는 트랜스포머(Transformer) 아키텍처가 중요한 역할을 하였다. 트랜스포머는 Self-attention 메커니즘을 통해 문장 내 모든 단어 간 관계를 효율적으로 계산하여 대규모 병렬 학습이 가능하도록 설계되었다 [8]. 이로써 기존 RNN 기반 모델보다 더 빠르고 효율적인 학습이 가능해졌다. 또한, Multi-Head Attention과 Positional Encoding과 같은 기법은 LLM이 문맥을 효과적으로 이해하고 문장 내 의미 관계를 학습할 수 있도록 돕는다. 이를 통해 LLM은 번역, 요약, 질문 답변과 같은 다양한 언어 작업에서 높은 성능을 발휘하며, 사용자 의도를 정확히 파악하고 자연스러운 응답을 생성할 수 있는 능력을 갖추게 되었다.

LLM 연구는 다양한 작업에서의 적응성 강화와 모델의 성능 및 효율성 향상에 초점을 맞추어 빠르게 발전해왔다. LLM을 기반으로 한 대화형 에이전트 연구 또한 다양한 분야에서 활발히 진행되며, 성능 향상과 실용적 응용을 목표로 빠르게 발전하고 있다. OpenAI의 ChatGPT는 GPT 모델의 버전을 빠르게 업그레이드 해나가면서 성능이 급속도로 향상되고 있으며, GPT 모델은 매개변수와 알고리즘의 발전에 따라 더 정교한 언어 이해와 응답 생성을 가능하게 하고 있다. GPT-1은 트랜스포머 아키텍처를 기반으로 1억 1천 7백만 개의 매개변수를 사용해 대규모 언어 모델의 가능성을 입증했다. GPT-2는 15억 개의 매개변수를 활용하여 자연스러운 텍스트 생성과 언어 모델의 응답 일관성을 크게 개선했으며, GPT-3는 1750억 개의 매개변수를 사용하여 복잡한 언어 처리 능력을 갖추어 자연스러운 텍스트 생성과 다양한 작업에서 높은 성능을 보였다 [9]. GPT-4는 멀티모달 학습을 지원하며 텍스트와 이미지를 함께 처리할 수 있는 능력을 갖추어 정밀한 언어를 기반으로 한 맞춤형 응답을 제공할 수 있도록 설계되었다. GPT-4o(omni)는 이름에서 알 수 있듯이 자연어, 청각, 시각 등 여러 종류의 데이터를 이해하고 처리하여 답변을 제공한다. 이후 공개된 o1 preview와 경량화 버전인 o1-mini는 코딩, 수

학, 과학 등 다양한 분야에서 복잡하고 논리적인 추론 능력을 바탕으로 특정 벤치마크 데이터셋에서 인간 전문가와 비교 할 만한 강력한 성능을 가진다 [10]. 이 외에도, Anthropic의 Claude는 안전성과 윤리적 대응을 중시하는 독자적 접근을 통해 독보적인 연구 방향을 제시하고 있다 [11]. Claude는 예측 불가능한 응답을 최소화하고 신뢰성 있는 대화를 제공하도록 설계되었으며, 이를 통해 사용자와의 상호작용에서 윤리적 지침을 준수하는 AI로 자리 잡고 있다. Microsoft의 Bing Chat은 GPT-4를 기반으로 실시간 웹 검색을 활용하여 최신 정보를 반영한 응답을 생성함으로써 정보의 정확성과 최신성을 보장한다. 이러한 실시간 검색 기능은 Bing Chat을 지식 제공 및 정보 질의응답 분야에서 중요한 AI 도구로 발전시키고 있다 [12]. Google의 Bard는 LaMDA (Language Models for Dialog Applications) [13] 모델을 기반으로 자연스러운 질문 응답을 지원하며, Google 검색과의 연계를 통해 폭넓고 심층적인 정보를 제공하는 특징이 있다. Meta의 Llama (Large Language Model Meta AI) 시리즈는 효율성과 비용 최적화를 목표로 개발된 모델로, GPT와 유사한 성능을 유지하면서도 경량화된 구조를 채택한 것이 특징이다 [14]. Llama 1은 학습 데이터의 효율성을 극대화하고 비용 절감을 목표로 하였으며, 연구자와 개발자에게 오픈 액세스를 제공함으로써 학술 연구와 응용 개발을 촉진하였다. Llama 2는 고품질 데이터로 학습하여 성능을 강화하였으며, 비용과 자원 효율성을 개선하면서도 GPT-4에 필적하는 성능을 보였다. 최신 버전인 Llama 3은 멀티모달 기능을 추가하여 텍스트뿐만 아니라 이미지 등 다양한 데이터 유형을 처리할 수 있으며, 이를 통해 더욱 정밀하고 풍부한 언어 이해와 응답 생성을 가능하게 했다. 이처럼 다양한 LLM 기반 대화형 에이전트들은 각기 다른 접근법과 기술을 통해 질의응답 기능을 지속적으로 발전시키고 있으며, 대화형 인공지능의 실용성과 응용 가능성을 한층 넓히고 있다.

국내 연구진들의 한국어 언어 모델 연구 역시 LLM 분야에서 중요한 발전을 이루고 있다. 대표적인 예로 LG의 EXAONE [15], NAVER의 CLOVA X [16], Bllossom [17] 등이 있으며, 이들 모델은 한국어 환경에 맞춘 자연어 처리 성능을 크게 향상시켰다. 특히 LG의 EXAONE은 한국어와 영어를 모두 이해하고 응답할 수 있는 이중언어 모델로, 추론 효율성을 크게 개선하여 처리 속도와 자원 사용의 최적화를 달성하였다. EXAONE 3.0은 다양한 도메인의 국내외 전문적인 데이터를 학습하여 다양한 산업 분야에서 실질적으로 응용될 수 있는 능력을 갖추었다고 평가되며 한국어 NLP 성능의 새로운 기준을 제시하고 있다.

2024년 이후 LLM 연구는 다양한 산업 및 응용 분야에서의 활용성을 지속적으로 확장하고 있다. 거대 언어 모델은 사용자 맞춤형 서비스와 상호작용형 애플리케이션에서 점점 더 중요해지고 있으며, 특히 법률, 의료, 교육 등 전문 분야에서의 채택이 늘어나고 있다. LLM이 특정 도메인 지식에 특화되도록 설계되고, 사용자 경험을 개선하기 위해 인간의 의도와 맥락을 보다 잘 이해할 수 있는 방향으로 발전하고 있다. 최근의 LLM 연구는 특정 산업과 전문 분야의 필요를 충족시키기 위해 도메인 맞춤형 언어 모델을 구축하는 데 초점을 두고 있으며, 이는 일반적인 LLM의 넓은 언어 이해를 넘어서 해당 분야의 복잡한 용어와 전문 지식을 세밀하게 다룰 수 있는 능력을 요구한다. 이 때 LLM 자체의 성능뿐만 아니라 RAG [18] 및 Chain of Thought (CoT) [19], Tree of Thought (ToT) [20]과 같은 기술이 함께 요구되기도 한다. 도메인 특화 데이터로 학습하는 LLM 모델은 일반 모델보다 더 높은 정확도와 일관성을 제공하며, 산업 내에서의 신뢰성과 활용도를 크게 높일 수 있다. 하지만 이러한 도메인 맞춤형 모델을 구축하는 과정에서 모델 드리프트(Model Drift)와 카타스트로픽 망각(Catastrophic Forgetting) 문제가 발생하는 등 개발에 대한 어려움이 있다 [21]. 모델 드리프트는 LLM이 학습된 데이터와 다른 분포의 새로운 데이터에 노출될 때 모델의 성능이 저하되는 현상을 의미하며, 이는 특정 도메인에 최적화된 모델일수록 일반화 능력에서 한계를 드러낼 수 있다. 또한, 카타스트로픽 망각은

LLM이 새로운 정보를 학습할 때 기존에 학습한 내용을 잊어버리는 현상으로, 특히 지속적 학습 환경에서 더욱 심각한 문제가 될 수 있다. 따라서, LLM 구축에 있어 이러한 모델 드리프트와 카타스트로픽 망각 문제를 잘 고려하는 것은 필수적이며, 이를 해결하기 위한 지속적인 연구와 혁신적인 접근이 필요하다.

2.2. Retrieval-Augmented Generation (RAG)

RAG [18]는 대규모 언어 모델의 한계를 보완하는 접근 방식으로, 실시간 검색 기능을 통해 외부 데이터베이스에 접근하고 그 정보를 텍스트 생성에 반영함으로써 훈련 후 최신 정보가 업데이트되지 않는 모델의 한계를 극복한다. 이는 고정된 훈련 데이터에만 의존하는 기존 LLM의 단점을 해결하기 위해 도입되었으며, 특히 최신 정보가 필요한 질의응답 시스템과 같이 빠르게 변화하는 응용 분야에서 필수적인 기술로 자리 잡고 있다. RAG는 '검색 단계'와 '생성 단계'로 구성되며, 검색 단계에서 벡터 데이터베이스(Vector DB)나 그래프 데이터베이스(Graph DB) 같은 외부 데이터베이스를 통해 관련 정보를 추출하고, 생성 단계에서 언어 모델이 이를 바탕으로 답변을 생성한다. Vector DB는 유사성 검색에, Graph DB는 복잡한 관계를 다루는 데 적합하다.

Vector DB는 비정형 데이터를 벡터 임베딩으로 변환해 고차원 벡터 공간에서 효율적인 유사성 검색을 가능하게 하며, NLP 및 컴퓨터 비전 응용에서 임베딩 생성 및 검색의 속도와 정확성을 높이는 데 기여하고 있다 [22]. Vector DB는 딥러닝 기반 임베딩을 활용해 문맥적 유사성을 반영한 검색 결과를 제공하며, 예를 들어 BERT나 GPT 같은 LLM에서 생성된 임베딩을 데이터베이스에 저장하여 문서 검색이나 추천 시스템에 사용된다. 이러한 연구들은 임베딩의 최적화를 통해 문서 검색의 효율성을 향상시키고, Vector DB와 언어 모델 간 상호작용의 개선에 초점을 맞추고 있다. 한편, Graph DB는 데이터 간의 복잡한 관계를 효율적으로 다루고 관리할 수 있는 구조로, 노드와 엣지의 그래프 구조를 통해 소셜 네트워크 분석, 추천 시스템, 지식 그래프와 같은 관계형 데이터 응용에 강점을 보인다 [23]. Graph DB는 고유의 쿼리 언어와 네이티브 그래프 엔진을 활용해 다양한 관계형 데이터를 유연하게 처리할 수 있는 기능을 제공하며, 특히 정보 간의 관계를 시각화하고 분석하는 데 중요한 역할을 한다. 최신 연구는 이러한 데이터베이스의 성능과 확장성을 높이기 위해 복합적인 관계 및 계층적 데이터를 효과적으로 다룰 수 있는 모델링 기법에 중점을 두고 있다. Comprehensive RAG Benchmark (CRAG) [24]는 RAG 연구의 발전을 위해 설계된 벤치마크로, 4,409개의 질문-답변 쌍과 지식 그래프(KG, Knowledge Graph)를 모사한 API를 포함하여 사실적이고 다양한 질의응답 작업을 다룬다. CRAG는 금융, 스포츠, 음악, 영화, 오픈 도메인을 포함한 다섯 개의 주요 도메인과 여덟 가지 질문 유형을 정의함으로써 RAG 시스템의 강점과 한계를 심층적으로 이해할 수 있도록 한다. 이를 통해 모델이 단순한 질문뿐만 아니라 복잡하고 동적인 질문에도 신뢰성 있는 답변을 제공할 수 있는지를 평가할 수 있다. CRAG에서는 질문 유형을 단순 사실을 묻는 단순 질문과 특정 조건이 포함된 조건부 단순 질문, 복수의 객체나 엔티티를 요구하는 집합형 질문, 두 엔티티의 비교를 요구하는 비교 질문, 검색 결과를 종합하여 답변하는 집계 질문, 다수의 정보를 연결해 답변을 생성하는 multi-hop 질문, 검색된 정보를 가공하거나 추론해야 하는 후처리 필요 질문, 그리고 잘못된 전제를 포함한 잘못된 전제 질문으로 정의했다. 이러한 질문 유형은 정보의 시간적 역동성 등을 반영하여 RAG 모델의 정보 검색 및 응답 생성 능력을 정밀하게 평가하도록 설계되었다.

RAG 시스템은 실시간 정보 검색 기능을 LLM의 응답 생성에 통합함으로써 다양한 상용 서비스에서 활용되고 있다. RAG의 적용은 특히 정보의 최신성과 정확성이 중요한 서비스에서 두드

러진다. Perplexity AI는 실시간 답변 제공, 답변의 투명성 확보 및 환각 현상 (Hallucination) 완화를 위해 RAG 기술을 적용한 검색 엔진을 개발하였다. 이 엔진은 사용자의 질문을 분석하여 핵심 내용을 파악한 후, 관련 정보를 웹에서 검색하고 이를 기반으로 정확하고 신뢰성 있는 응답을 생성하도록 설계되었다. Microsoft의 Bing Chat 또한 실시간 검색 기능을 통해 RAG 기술의 개념을 적용하였다. 이처럼 RAG 연구는 대규모 언어 모델의 한계를 보완하고, 다양한 산업 분야에서 신뢰성과 최신성을 강화한 정보 검색 및 생성 시스템의 발전을 이끌고 있다.

2.3. Chain of Thought (CoT)

Chain of Thought(CoT) [19]는 언어 모델이 복잡한 질문이나 논리적 사고가 필요한 상황에서 일련의 사고 과정을 단계별로 진행하도록 유도하는 방법으로, 고난도 추론 문제 해결에 효과적이다. CoT는 단순한 답변 생성이 아닌 문제를 여러 단계로 분해하여 중간 결과를 도출함으로써 보다 정확하고 논리적인 결론에 도달하게 한다. 이는 LLM의 직관적 답변 의존 한계를 극복하고 논리적 사고를 통해 복잡한 질문에 대한 정확한 답변을 가능하게 한다. CoT는 기존 언어 모델이 복잡한 논리적 추론에서 성능이 제한적이라는 문제에서 출발했다. CoT의 아키텍처는 문제를 단계별로 해석하고 결과를 축적하여 최종 결론에 이르는 구조로, 각 중간 단계의 정보를 참조해 정확성을 높인다. '단계별 추론 프로세스'를 사용하여 각 단계의 결과를 순차적으로 다음 단계로 넘기며, 프로세스 기반 토큰화와 중간 단계 결과 저장 및 참조, 결과 검증 등을 통해 복잡한 문제 해결을 지원한다.

[19]의 연구는 Chain of Thought Prompting 기법을 통해 ChatGPT, LaMDA, PaLM와 같은 기존 언어 모델의 추론 성능을 향상시킬 수 있음을 증명하였다. CoT는 특히 고난도의 수학 문제와 논리적 추론을 요구하는 작업에서 모델의 성능을 크게 개선하는 데 효과적임이 입증되었다. 이러한 연구는 LLM의 추론 능력을 강화하고 다양한 도메인에서의 적용 가능성을 확대할 수 있는 중요한 방법론으로써 CoT의 가능성을 제시하였다. 이를 확장한 Tree of Thought (ToT) [20]는 언어 모델이 복잡한 문제를 해결할 때 더 정교한 추론 과정을 수행할 수 있도록 고안된 방법론이다. ToT는 CoT가 선형적인 단계별 추론 과정을 통해 문제를 해결하는 데 비해, 다양한 경로를 탐색하여 최적의 결론에 도달할 수 있는 비선형적인 사고 과정을 포함한다. 이 접근법은 언어 모델이 여러 경로로 탐색하고 각 경로에서의 중간 결과를 평가하면서 다양한 옵션 중 최적의 답을 선택할 수 있게 한다. 각 단계는 다양한 선택지를 제공하며, 모델은 이를 바탕으로 새로운 결정을 내리며 점진적으로 문제를 해결해 나간다. ToT는 특히 논리적 사고, 의사결정, 문제 해결에서 더 높은 정확성과 효율성을 필요로 하는 응용 분야에 적합하다. CoT가 문제를 단계별로 해결하는 사고 과정이라면, ToT는 이를 확장하여 다양한 가능성을 검토하고 최적의 해결책을 탐색하는 복합적인 사고 과정을 포함한다는 점에서 차별화된다. CoT는 LLM이 특정 도메인에서 외부 데이터를 활용할 때 논리적이고 간결한 답변을 생성하는 데 적합하다. 비선형적인 추론 과정은 검색한 정보를 결합하고 평가하는 데 복잡성을 증가시킬 수 있기 때문에, CoT는 보다 효율적인 방법이 될 수 있다.

3. 본론

수사지원과 같이 활용 분야와 활용처가 어느 정도 특정되어 있는 대화형 인공지능 에이전트 시스템에서는 LLM, RAG, CoT와 같은 기술을 목적과 해당 분야의 특징에 맞게 최적화하여 활용하는 것이 중요하다. 본 연구에서는 해당 기술들의 적용이 가능한 대화형 에이전트 시스템의

기본적인 참조 구조를 제안하고자 한다.

LLM을 효과적으로 활용하는 시스템을 구축하기 위해서는 RAG 기반의 데이터 흐름의 설계가 필수적이다. 방대한 데이터를 바탕으로 사건의 정황을 분석하거나 관련 배경 정보를 참고해야 하는 경우, 효율적인 RAG 설계를 통해 필요한 데이터를 신속하고 정확하게 추출할 수 있어야 한다. 특히, 실시간으로 변화하는 정보를 반영해야 하는 상황에서는 실시간 검색 기능이 연동된 RAG 시스템이 필수적이다. 수사 지원 시스템의 경우 다양한 종류의 데이터 시스템이 존재할 수 있으며, 연동 방식과 정보 검색 방법이 각각 다를 수 있다. RAG 수행 과정에서 과도한 정보를 추론의 배경 지식으로 활용할 경우, LLM이 맥락의 핵심을 놓쳐 응답의 품질이 저하될 가능성이 있으며, 입력 토큰 수 증가로 인해 처리 시간이 길어져 대화형 에이전트 시스템의 응답 지연이 발생할 수 있다.

복잡한 논리적 추론이 요구되는 질의에 대응하기 위해 CoT 기법을 활용하는 것이 효과적일 수 있다. CoT 기법은 모델이 단계별 사고 과정을 통해 복잡한 문제를 논리적으로 해결하도록 지원하여, 대화형 에이전트의 응답 정확성과 신뢰성을 향상시킨다. 여러 요소를 동시에 고려하는 것보다 CoT를 활용해 순차적으로 분석하도록 설계하면 응답의 품질을 높일 수 있다. 그러나 CoT가 지나치게 복잡하게 설계되면 유지보수가 어려워지고, 응답 지연 시간이 길어지는 문제가 발생할 수 있다. 마지막으로, 적절한 LLM 모델의 준비가 중요하다. 먼저, OpenAI의 GPT-4o와 같은 외부 API를 연동하여 사용할 수 있는 모델을 선택할 것인지, 자체 구축한 sLLM 모델을 활용할 것인지를 결정해야 한다. 외부 API를 이용한 LLM 모델은 대체로 성능이 우수하고 편리하다는 장점이 있지만, 비용 문제와 보안 이슈가 발생할 수 있다. 특히, 데이터 유출이 허용되지 않는 상황에서는 외부 API 기반 LLM의 사용이 제한될 수 있다. 이러한 이유로 대부분의 수사 지원 시스템은 sLLM을 기반으로 구축하는 것이 필요하다. 공개된 모델을 우선적으로 고려할 수 있으며, 충분한 GPU 자원과 잘 정제된 대량의 데이터가 확보된 경우에는 파인튜닝을 통해 모델을 자체적으로 구축하는 방안을 검토할 수 있다. 본 연구에서는 수사 지원 대화형 인공지능 에이전트 시스템의 기본 참조 구조를 정의하고 이를 구현하였다. 또한, 특별사법 경찰 매뉴얼을 기반으로 한 기초적인 RAG 기반 QA 벤치마크를 설계하고, 이를 통해 참조 구조의 구현 결과에 대한 성능 시험을 수행하였다. LLM 모델로는 SOTA 수준인 GPT-4o부터 대표적인 한글 지원 sLLM 모델인 EXAONE까지 다양한 모델들의 성능을 비교하였다.

3.1. 수사지원 QA 벤치마크 데이터셋 구성

Table 1은 본 연구에서 제안하는 수사지원 QA 벤치마크의 요약이다. LLM 및 RAG를 활용하여 수사 지원 대화형 인공지능 에이전트를 구축할 때는 고려해야 할 요소가 많고, 각 요소가 결과에 어떤 영향을 미칠지 직관적으로 예측하기 어렵다. Vector DB, 임베딩 모델, 프롬프팅, LLM 모델 등 각 구성 요소가 성능에 영향을 주며, 이러한 요소들 간의 상관관계가 독립적이지 않기 때문에 현장 요구사항을 충족하는 시스템을 구현하는 과정에서 설계 및 구현의 복잡도가 기하급수적으로 증가한다. 따라서, 대화형 인공지능 에이전트 시스템의 구현에 앞서, 성능을 정의하고 측정할 수 있는 수단이 필요하다. 기존에 LLM 및 RAG를 평가하는 다양한 벤치마크가 공개되어 있으나, 한글 기반의 수사 지원 분야에 특화된 벤치마크는 전무하다. 본 연구에서는 이러한 수사 지원 분야에서 LLM 및 RAG를 활용한 대화형 인공지능 에이전트 시스템의 성능을 평가하기 위한 벤치마크를 제안한다. 벤치마크의 배경 지식은 대표적인 특별사법경찰 관련 법률인 “사법경찰관리의 직무를 수행할 자와 그 직무범위에 관한 법률”을 포함한 6개 법령으로 선정하였으며, 이를 기반으로 632개의 벤치마크 질의문을 정의하였다. 질의문은 복합 추론을

최대한 배제하고, 문서 확인만으로 답변이 가능한 기초적인 내용으로 구성하였다. 질의문은 주로 직무 범위, 수사 절차, 행정 절차, 법적 의무 등에 대해 질의하는 형식으로 구성하였으며, 각 항목은 관련 법령의 근거를 포함하여 질의(Q) - 응답(A) - 근거(R) 형태로 구성된다. 해당 벤치마크는 LLM 및 RAG의 기본적인 질의응답 기능 평가를 위해 설계되었으며, 이를 통해 수사 지원 분야에서 대화형 에이전트 시스템 개발의 성능 기준을 정의할 수 있다.

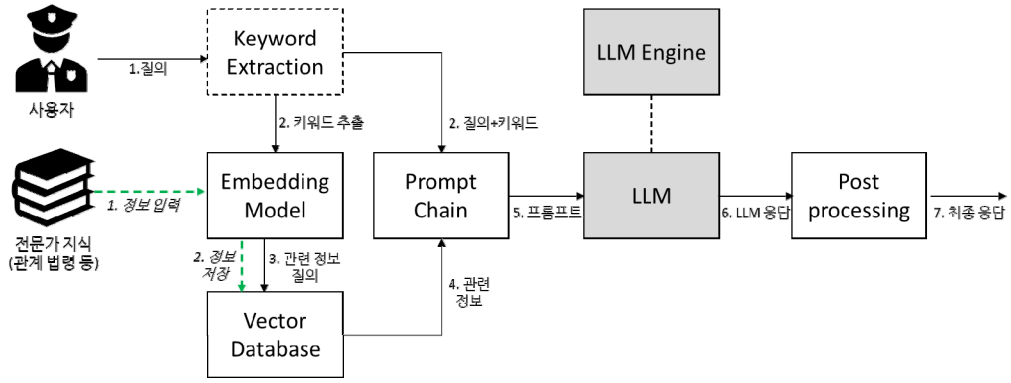
<Table 1> QA benchmark dataset for supporting investigative

번호	관계 법령	질의문수	예시질의	모범답안	근거
1	사법경찰관리의 직무를 수행할 자와 그 직무범위에 관한 법률	120	사법경찰업무 관련 산림보호공무원의 직무범위는?	소속 관서의 임야에서 산림 및 임산물에 관한 범죄를 수사 가능	제6조제5항 가목
2	특별사법경찰관리 지명절차 등에 관한 지침	97	벌금형을 받은 경우 지명에 부적격할 수 있나요?	벌금 200만 원 이상의 형을 받고 5년 경과 전에는 지명 부적격	제5조의2제1항 제1호
3	검사의 사법경찰관리에 대한 수사지휘 및 사법경찰관리의 수사준칙에 관한 규정	100	특별사법경찰관의 주요 임무는 무엇인가요?	범인과 범죄 사실을 수사하고 증거를 수집하는 업무를 수행	1장제2조제1항
4	(경찰청)범죄수사규칙	115	변호인 접견을 처리할 때 요구해야 하는 서류는 무엇인가요?	변호사 신분증과 접견신청서를 확인해야 합니다.	제80조 2항
5	체포구속 업무처리 지침	100	긴급체포 시 어떤 조건이 충족되어야 합니까?	긴급체포는 범죄의 중대성, 증거 인멸 및 도주의 염려, 그리고 체포 긴급성이 있을 때 가능합니다.	제3장 제1조 가, 나, 다항
6	인권보호수사규칙	100	장시간 조사 제한이 있나요?	조사 시간은 총 12시간을 초과하지 않아야 하며, 특별한 사유가 없으면 실제 조사 시간은 8시간을 넘지 않아야 합니다.	제44조제1항

3.2. 수사지원 대화형 인공지능 에이전트 시스템 참조구조

Figure 1은 본 연구에서 제안하는 수사 지원 대화형 인공지능 에이전트 시스템의 참조 구조를 나타낸다. 사용자가 자연어로 질의하면, Vector DB에 저장되어 있는 배경 지식에서 관련 정보를 검색하고 이를 바탕으로 LLM이 응답을 생성한다. 먼저, 사용자의 질의에서 키워드를 추출하여 임베딩한 후, 이를 기반으로 Vector DB에서 관련 정보를 찾는다. 일반적인 RAG에서는 사용자의 질의를 그대로 활용하지만, 수사 지원 시스템의 경우 도메인에 특화된 단어가 많이 사용되고, 비슷한 형태를 가진 단어라도 전혀 다른 의미를 지닐 수 있기 때문에 질의를 그대로 사용하는 것보다 특화된 단어로 변환하는 것이 상대적으로 유리하다. 전문가 지식으로 활용되는 관계 법령과 예상되는 사용자의 질의가 비교적 단순한 구조라면 일부 단계를 생략할 수도 있다.

Vector DB를 활용하기 위해서는 반드시 임베딩 모델을 사용해야 한다. 임베딩 모델은 문장이나 단어를 고차원의 벡터로 변환하는 역할을 하며, 한글 임베딩 모델의 성능은 전체 RAG의 성능에 영향을 미칠 수 있다. 다른 벤치마크에서 성능이 우수한 임베딩 모델일지라도, 수사 지원과 같은 특수한 도메인의 언어 체계에서는 다른 성능을 보일 가능성이 있다. 따라서 Vector DB에 관계 법령 등의 전문가 지식을 사전에 입력할 때에도 임베딩 모델을 거쳐야 하므로, RAG 성능이 저하될 경우 임베딩 모델의 교체를 고려하는 것이 필요하다.



<Figure 1>The proposed reference architecture for an investigative support conversational AI agent

관계 법령 등의 전문가 지식을 Vector DB에 사전 입력 할때에도 다양한 고려사항이 있다. 문서의 형태로 데이터를 입력 할 때, 특수문자가 함께 입력되는데 통상적으로 Vector DB는 Vector를 기반으로 관련 정보를 검색하기 때문에 특수 문자를 제거하는 것이 필수적이지는 않으나, 상황에 따라서는 Vector DB에 입력되는 정보들이 잘 전처리 되는것이 RAG 성능 향상에 도움이 될 수 있다. 일반적으로 Vector DB에 질의를 할 때는 연관성이 높은 순서대로 n개의 단락을 뽑는 방식으로 진행한다. 따라서 전문가 지식을 입력할 때 중복된 정보가 저장되지 않도록 하는 것이 중요하다. Vector DB를 통해서 관련 정보를 검색한 후, 사용자의 질의 정보와 질의 정보의 키워드 등을 활용하여 Prompt Chain을 통해 실제적으로 LLM에 입력될 프롬프트를 생성해야 한다. 프롬프트 엔지니어링에는 굉장히다양한 기법과 방법론이 있으며, 사용자 질의 패턴, RAG로 부터 획득한 관련정보, LLM 모델 종류 등으로 최적의 프롬프트가 달라질 수 있다. Step by Step과 같은 단어들을 넣어서 성능을 향상시키는 방법론부터, 복잡도가 높은 CoT 응용 기법들까지 다양한 기술들이 성능향상을 위해 고려될 수 있다. 이후 완성된 프롬프트는 LLM에 쿼리문으로 전달되어 응답을 요청하게 된다. 수사지원 대화형 인공지능 에이전트의 경우, 보안상의 이유로 대부분 sLLM을 고려할 수 밖에 없다. 서비스하고자 하는 사용자의 예상 규모에 따라 sLLM을 어떻게 잘 서빙 할 것인지에 대해서도 많은 고려 요소가 있다. 자체 GPU서버를 활용할때의 방법론과 클라우드를 활용할때의 방법론이 다를 수 있다.

LLM이 응답을 생성하면, 해당 응답내용에 대한 후처리 절차가 반드시 필요하다. 확률적으로 LLM이 정해진 포맷대로 출력을 하지 않을 가능성이 있어, 상황에 따라서는 복잡한 파싱기능이 구현되어야 할 수 있다.

3.3. 성능평가 방법

성능 평가는 사람이 직접 평가하는 Human Evaluation으로 진행하였다. 평가 기준은

CRAG [24]를 참고하여 Table 2와 같이 정의하였다. Perfect, Acceptable, Missing, Incorrect의 네 단계로 응답의 품질을 평가한다. Perfect는 충분히 완벽한 대답을 한 상황이고, Acceptable은 유용한 대답을 했으나 필요한 정보가 충분히 포함되지 못한 상황을 의미한다. Missing은 정보가 없거나 모른다 등의 대답이 나왔을 경우를 의미한다. 기본적으로 해당 벤치마크는 관계 법령을 기준으로 만들어졌으므로, 정보가 없다고 나온 것은 RAG로 관련 정보를 찾지 못했다는 의미이다. 제대로된 정보를 찾았어도 LLM이 Missing으로 판단할 가능성이 있으나, 본 연구의 실험환경 상 RAG에서 한번에 가져오는 데이터의 양이 크지 않았기 때문에 그런 경우는 발견되지 않았다. Incorrect의 경우, 틀린 대답을 하는 것은 서비스 관점에서 모르는 것보다 훨씬 치명적이므로 -1점으로 설정하였다.

<Table 2> The scoring table based on investigative support QA benchmark

평가	내용	Human Evaluation
Perfect	사용자의 질문에 올바르게 대답	1
Acceptable	질문에 유용한 답변을 제공했으나 사소한 오류가 있음	0.5
Missing	모르겠습니다 등의 시스템 오류	0
Incorrect	적절하지 않은 답변 또는 관련 없는 정보 제공	-1

3.4. 실험 결과

본 연구에서는 제안하는 대화형 에이전트 시스템 참조구조를 구현하고, 정의한 수사지원 QA 벤치마크 데이터셋을 기반으로 GPT-4o와 GPT-mini를 테스트하였고, 추가로 공개 한글 sLLM인 EXAONE 3.0 7.2B [15] 모델을 적용하여 평가하였다. GPT-4o는 현재 사용가능한 LLM 모델중 SOTA 수준의 모델로 평가받고 있으나, API 비용이 비싼 단점이 있고, GPT-mini는 성능은 다소 낮으나, 가격이 GPT-4o 대비 매우 저렴하다는 장점이 있다. EXAONE 3.0 의 경우 공개된 한글 기반 sLLM 모델 중 세계 최고 수준 성능을 보이는 대표 모델로써, 본 연구의 적용 모델로 선정하였다. Table 3에서는 해당 벤치마크를 기반으로, 본 연구에서 제안하는 참조구조의 구현 결과물에 대한 성능평가 결과를 나타내고 있다. Accuracy는 틀린 대답을 하는 Incorrect를 제외한 Perfect, Acceptable, Missing 의 결과에 대한 전체 데이터 대비 비율로 정의하였다. GPT-4o가 0.95로 가장 높았고, EXAONE 이 0.92로 그다음, GPT-mini 가 0.82로 가장 낮았다. GPT-4o가 Accuracy가 높았던 이유는 Missing의 비율이 높아서 Incorrect의 비율이 낮게 나왔기 때문이라고 볼 수 있다. GPT-mini의 경우, Incorrect 가 18%를 기록하게 되면서 Accuracy는 0.82이며 Score는 0.56으로 나타났다. 특이한 점은 Missing이 1% 수준이라는 것인데, RAG이 정상동작 하지 않아서 배경지식이 적절하지 않을 경우, Incorrect로 분류되는 수준의 답변을 할 확률이 높았다. 해당 현상이 전형적인 환각(Hallucination)으로 볼 수 있다. EXAONE의 경우 Accuracy가 0.92 였고 Score도 0.77 수준으로 나타났다. Incorrect가 GPT-4o 보다는 높았지만, GPT-mini 보다는 절반수준으로 낮았다. Acceptable의 경우 다른 모델과 큰 차이가 없었는데, Perfect가 다른 모델들에 비해 오히려 더 높게 나왔다. 동일한 RAG이 적용되었는데, 다른 모델들과 비교하였을 때 Missing에 비해 Perfect가 높은 것은 RAG을 실패한 상황에서도 모델의 자체 추론으로 Perfect 에 해당하는 정답을 추론하는 비율이 높았다는 것을 암시한다.

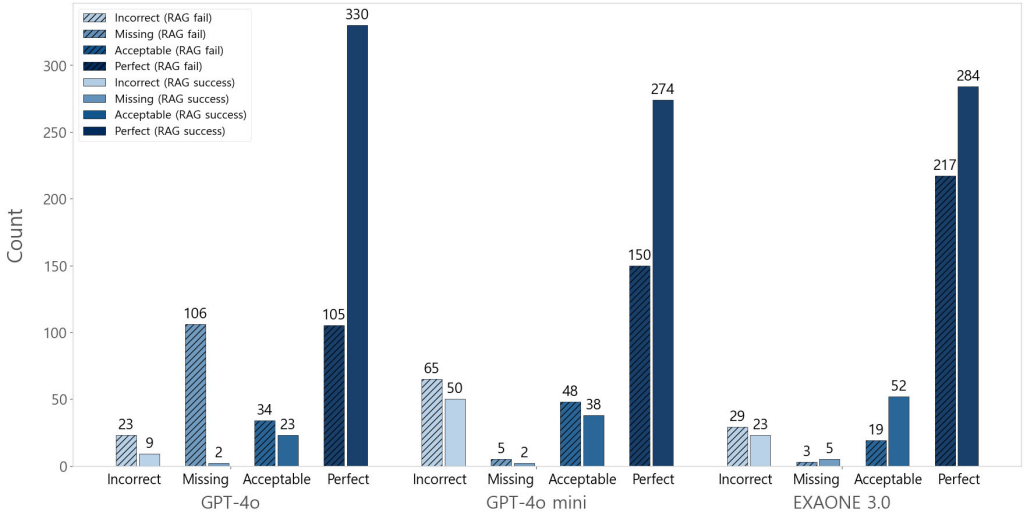
<Table 3> Performance comparison table by LLM model. Accuracy represents the ratio of Perfect, Acceptable, and Missing relative to the entire dataset. Incorrect, Missing, Acceptable, and Perfect are each displayed as count / ratio relative to the entire dataset.

Model	Accuracy	Score (H.E)	Incorrect	Missing	Acceptable	Perfect
GPT-4o	0.95	0.68	32 / 5%	108 / 17%	57 / 9%	435 / 69%
GPT-mini	0.82	0.56	115 / 18%	7 / 1%	86 / 14%	435 / 67%
EXAONE	0.92	0.77	52 / 8%	8 / 1%	71 / 11%	501 / 79%

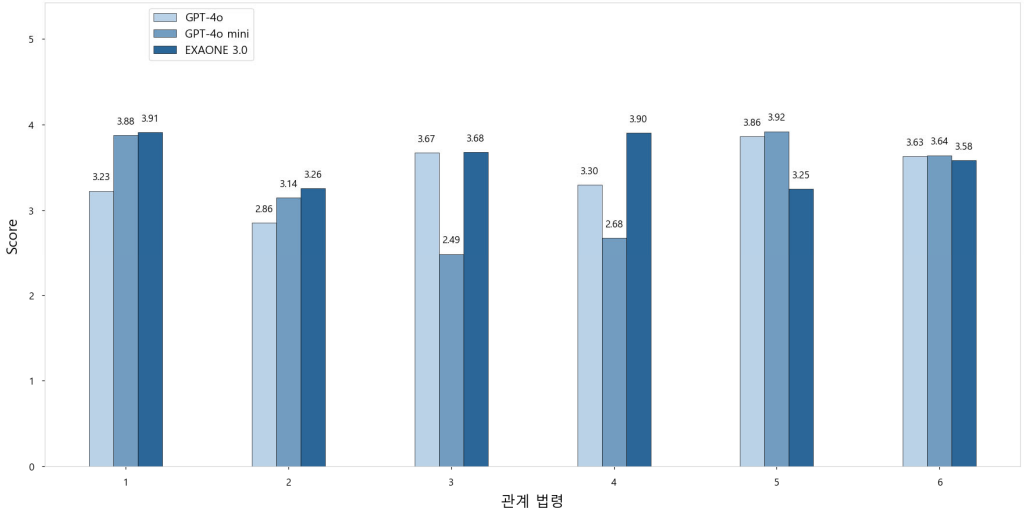
관련 정보 없이 해당 벤치마크 QA의 응답 품질이 다른 모델 대비 높다는 것은 EXAONE 모델이 해당 데이터를 사전학습에 활용했을 가능성이 높다는 것을 의미한다. 현재 EXAONE 의 구체적인 학습데이터는 공개되어 있지 않으나, 경찰 관련 법령은 온라인에 공개되어 있는 데이터가 많으므로 사전학습에 활용했을 가능성이 있다. 그렇지 않다면, Missing으로 집계된 응답의 개수가 작은 것으로 미루어 볼 때, GPT-mini 와 큰 차이가 나지 않았을 확률이 크다.

Figure 2와 3은 각각 RAG의 성공 여부에 따른 모델 별 성능과 관계 법령에 따른 모델별 성능 차이를 나타내고 있다. Vector DB로부터 조회한 내용이 질의와 직접적인 연관이 없는 경우를 RAG fail 로 나타내어 그림에서 빗금 영역으로 표기하였다. GPT-4o 의 경우, missing으로 판단한 사항들의 대부분은 RAG에 실패한 사항이었다. RAG에 실패한 질의 중 약 105건은 Perfect 품질의 응답을 도출했으며 34건은 Acceptable 품질의 응답을 도출했다. 이는 벤치마크의 구성이 비교적 간단하게 되어 있기 때문에, 배경지식이 없어도 GPT-4o 수준의 SOTA 모델이라면 Perfect 수준의 답변이 가능한 경우가 어느 정도 있을 수 있음을 의미한다. 배경지식이 주어지는 경우 Perfect 품질의 응답을 생성하는 경우는 다른 모델 대비 훨씬 좋은 성능을 보였다. GPT-4o mini의 경우, RAG를 실패한 질의에 대해서 Missing으로 판단하는 비율이 현저히 낮았다. 하지만 RAG를 실패한다고 해서 모든 결과가 Incorrect의 응답을 생성하지는 않았다. Incorrect는 RAG를 실패한 경우 65건, RAG를 성공한 경우가 50건으로, RAG의 실패가 Incorrect 품질 응답 생성에 많은 영향을 끼치는 것은 맞지만, 절대적이라고 보기는 어려웠다. 반면 RAG를 실패한 경우에 Perfect 품질의 비율을 나타낸 경우가 150건으로, GPT-4o 대비 훨씬 높게 집계되었다. 이는 모델 성능의 한계로 인해, RAG를 실패한 상황에 대해 Missing으로 판단하지 못하고 응답생성을 시도했는데, 꽤 높은 비율로 Perfect 품질의 응답생성에 성공했다는 의미이다. RAG를 성공한 경우에 GPT-4o 대비 Perfect 건수가 약 60건 낮은데, 모델성능의 차이로 인해 RAG에 성공했음에도 불구하고 Incorrect로 분류하는 비율이 높았다는 것을 의미한다. EXAONE 3.0 모델의 경우 Missing이 낮은 것은 GPT-4o mini와 유사한데, RAG를 실패했을 때 Perfect 품질의 응답을 생성하는 비율이 확실히 다른 모델 대비 높은 것을 확인할 수 있다. 이는 해당 모델이 배경지식과 유사한 데이터로 사전학습되어, RAG에 실패한 상황에서도 모델이 잠재적으로 학습된 정보를 이용하여 적절한 대답을 도출했다고 해석할 수 있다. 모델의 성능을 의미하는 RAG를 성공한 상황의 Perfect 비율이나, RAG를 실패한 상황에서 Missing으로 판단하는 비율은 GPT-4o mini와 유사한 성능을 보이고 있는 것으로 보아, 기본적인 모델의 성능은 적어도 본 벤치마크 및 RAG 시스템을 기준으로 봤을 때 EXAONE 3.0 이 GPT-4o mini와 유사하다고 판단할 수 있다. Figure 3에서는 6개의 관계법령에 대해, 각 관계법령 별로 모델에 대한 Score를 비교하고 있다. 관계 법령 3과 4에서 GPT-4o mini의 성능이 대폭 떨어진 반면, EXAONE 3.0은 GPT-4o와 비슷하거나 오히려 더 나은 성능을 보였다는 점이다. 관계 법령 3이 “검사의 사법경찰관리에 대한 수사지휘 및 사법경찰관리의 수사준칙에 관한 규정” 이고, 관계 법령 4가 “(경찰청)범죄수사규칙”임을 고려하면, 우연히 수사에 관련된 법령들에 한해

EXAONE 3.0 의 성능이 좋게 나온 것은 관련된 문서가 EXAONE 3.0 에 사용되었을 확률이 높다는 것을 의미한다. 그 밖에 관계 법령 1에서 GPT-4o에 비해 GPT-4o mini와 EXAONE 3.0의 성능이 높게 나온 것을 확인할 수 있다. 실험 결과, GPT-4o mini나 EXAONE 3.0이 RAG를 실패한 상황에서 추론을 시도하여 답변의 품질이 좋은 경우가 관계 법령 1에 많이 나타나고 있었다. 관계 법령 1은 "사법경찰관리의 직무를 수행할 자와 그 직무범위에 관한 법률"로써, 공무원이 사법경찰관리 지명을 받는 경우, 본래의 해당 공무원의 업무 범위에서 사법경찰 관리 업무를 수행할 수 있다는 내용을 담고 있다. 다양한 직무의 범위를 나열하고 있으나, 해당 공무원은 본 업무의 분야에서 사법경찰관리 직무를 수행하게 되어있는 비교적 간단한 논리구조를 가지고 있으므로, GPT-4o mini나 EXAONE 3.0이 missing 판단을 하지 못하고 추론을 시도했을 때 괜찮은 품질의 응답을 생성할 확률이 높았다고 보여진다. 수사에 관한 내용을 다루는 관계법령 3이나 4의 경우, 관계 법령 1에 비해 비교적 복잡한데, 해당 상황에서 GPT-4o mini의 성능이 떨어지는 것을 확인할 수 있다.



<Figure 2> Benchmark performance comparison by model based on RAG success rate (total RAG successes: 364, approximately 58%)



<Figure 3> Benchmark performance comparison by model according to relevant legislation

4. 결론

본 연구는 수사 지원을 위한 LLM 및 RAG 기반 대화형 에이전트 시스템에 대한 기초적인 참조구조와 벤치마크 데이터를 제안하고, 해당 시스템에 GPT-4o, GPT-4o mini, EXAONE 3.0 모델들 적용 시 나타나는 성능 변화에 대해 분석하였다. 실험을 통해 GPT-4o와 EXAONE이 높은 성능을 보였으나, 비용과 보안성 문제로 인해 GPT-mini와 같은 경량 모델이 특정 상황에서는 더 적합할 수 있음을 확인하였다.

향후 연구로는 벤치마크 데이터셋에 대한 고도화를 진행할 예정이다. 실제 수사지원 분야 관련하여 수요가 있는 데이터들을 선정하고, 전문가 자문을 통해 고도화된 벤치마크 QA 데이터셋을 만들어 이를 공개할 예정이다. 이를 통해 해당 분야에서 다양한 연구가 수행되길 기대한다. 또한, 해당 벤치마크를 바탕으로 LLM 및 RAG 시스템의 성능평가를 자동적으로 수행할 수 있는 프레임워크를 개발하고자 한다. 이를 통해 다양한 임베딩 모델들과 RAG 기법 및 CoT 등 프롬프팅 기법들을 테스트하여, 수사 지원 분야에서 이러한 기술들이 어떠한 영향을 끼치는지 분석할 예정이다. 이러한 연구들을 종합하여, 최종적으로는 현장에서 활용가능한 수준으로 수사지원분야에 최적화된 대화형 에이전트 시스템을 연구개발 하고자 한다.

참고문헌(References)

[1] An DU, Leem CS. 2019. Artificial Intelligence Algorithms, Model-Based Social Data Collection and Content Exploration. The Korea Journal of BigData, 4(2), 23-34.

[2] Zhao C. 2022. Perspective on Nonstationary Process Monitoring in the Era of Industrial Artificial Intelligence. Journal of Process Control, 116, 255-272.

[3] Al-Surmi A, Bashiri M, Koliousis I. 2021. AI-Based Decision Making: Combining Strategies to Improve Operational Performance. International Journal of Production Research, 60(14), 4464-4486.

[4] Ahmad SF, Han H, Alam MM, et al. 2023. Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and Safety in Education. Humanities and Social Sciences Communications, 10(1), 1-14.

[5] Ko HH. 2015. A Comparative Review on the Special Judicial Police System. International Law Review, 7(1), 35-63.

[6] Yao Y, Duan J, Xu K, et al. 2024. A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. High-Confidence Computing, 4(2), 1-21.

[7] Fan W, Ding Y, Ning L, et al. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, (Barcelona,) 6491-6501.

[8] Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, 5998-6008.

[9] Brown T, Mann B, Ryder N, et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020

[10] Islam R, Moushi OM. 2024. Gpt-4o: The cutting-edge advancement in multimodal LLM. Authorea Preprints.

[11] Priyanshu A, Maurya Y, Hong Z. 2024. AI Governance and Accountability: An Analysis of Anthropic's Claude. arXiv preprint arXiv:2407.01557, 2024

[12] Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. 2024. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. European Journal of Orthodontics, cjae017.

[13] Thoppilan R, Freitas D, Hall J, et al. 2022. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239, 2022

[14] Touvron H, Martin L, Stone K, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023

[15] An SY, Bae KH, Choi EB, et al. 2024. EXAONE 3.0 7.8 B Instruction Tuned Language Model. arXiv e-prints arXiv:2408.03541, 2024

[16] Yoo KM, Han JG, In SK, et al. 2024. HyperCLOVA X Technical Report. arXiv preprint arXiv:2404.01954, 2024.

[17] Choi CS, Jeong YB, Park SY, et al. 2024. Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean. arXiv preprint arXiv:2403.10882, 2024.

[18] Lewis P, Perez E, Piktus A, et al.2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. in Proceedings of the Advances in Neural Information Processing Systems 33, (Virtual Conference), 9459-9474.

[19] Wei J, Wang X, Schuurmans D, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. in Proceedings of the Advances in Neural Information Processing Systems 35, (Virtual Conference), 24824-24837.

[20] Yao S, Zhu X, Narang A, et al. Tree of thoughts: Deliberate problem solving with large language models. in Proceedings of the Advances in Neural Information Processing Systems 36(Virtual Conference).

[21] Luo Y, Yang Z, Meng F, et al. 2023. An empirical study of catastrophic forgetting in large

language models during continual fine-tuning. arXiv preprint arXiv:2308.08747, 2023.

[22] Han YK, Liu CJ, Wang P. 2023 A comprehensive survey on vector database: Storage and retrieval technique, challenge. arXiv preprint arXiv:2310.11703, 2023.

[23] Vicknair C, Macias M, Zhao Z, et al. 2010. A comparison of a graph database and a relational database: a data provenance perspective. in Proceedings of the 48th Annual Southeast Regional Conference.

[24] Yang X, Sun K, Xin H, et al. 2024. CRAG-Comprehensive RAG Benchmark. arXiv preprint arXiv:2406.04744, 2024.