「인터넷전자상거래연구」 제24권 제4호
2024년 8월, pp. 49~62
한국인터넷전자상거래학회
https://doi.org/10.37272/JIECR.2024.08.24.4.49

AI 챗봇 시스템 주요 기술 동향 분석을 통한 민원 처리시스템 개선

AI chatbot system improvement of complaint handling system through analysis of major technology trends

권 오 균* · 신 용 태**
Ohkyoon Kwon · Yongtae Shin

··· Abstract ···

This paper analyzes the major technology trends of AI chatbot systems and proposes improvement measures for civil complaint handling systems. In modern society, civil complaint services are an important means for governments and companies to communicate with customers, and the introduction of chatbots can play an important role in increasing the efficiency of these services and customer satisfaction. However, there is a growing need to improve work inefficiency as existing chatbot-based civil complaint handling systems have limitations in handing complex complaints. This study examines the latest AI technology, LLM, RAG, and GraphRAG, analyzing the practical usage and limitations of each technology to propose an innovative improvement plan for civil complaint handling systems. In particular, to create an efficient and accurate civil complaint handling system, analysis on the characteristics of civil complaint data and consideration of network separation environment is vital. In addition, to increase the accuracy of civil complaint handling and real-time responses, HybridRAG technology and Agent/Tool technology is proposed to handle LLM hallucination problems. As a result of the study, the proposed system is expected to significantly improve the quality of civil complaint services and customer satisfaction, and will provide important implications and directions for the future development of AI chatbot systems.

Key Words : Complaint data, LLM, AI chatbot, Intent analysis, RAG, GraphRAG, Tool systems

논문 접수일: 2024년 07월 25일, 게재 확정일: 2024년 08월 23일
* 숭실대학교 IT정책경영학과 박사과정, 주저자
** 숭실대학교 컴퓨터학부 교수, 교신저자

Ⅰ. 서 론

현대 사회에서 민원 서비스는 정부와 기업이 고객과 상호작용하는 핵심 요소 중 하나이다.

특히 기존의 통신 민원 서비스는 서비스 가입, 변경, 해지, 멤버십, 혜택, 청구, 수납 등 이와 관련된 다양한 민원을 해결하는 데 있어, 인력과 시간이 소모되는 단점이 컸다. 이를 보완하기 위한 대안으로 떠오른 것이 24시간 챗봇을 통한 민원 서비스라고 할 수 있다. 현재 많은 공공 및 금융기관, 통신사 등에서 챗봇 서비스를 도입하여 활용하고 있으며 이는 민원처리 과정에 있어 고객 만족도를 나타내는 서비스 품질과 효율적인 업무처리에 있어 긍정적인 고객 경험과 기업의 경쟁력 우위를 점유하는 데 직접적인 영향을 미친다.

그러나 이러한 챗봇을 활용한 민원 처리시스템은 역시 고객들에 까다로운 업무에 있어 복잡하고 비효율적인 프로세스, 서류 제출의 번거로움, 시간 소모가 많은 절차 등의 문제로 인해 불편함을 초래하고 있다. 또한, 민원처리를 위해 고객은 주로 영업점, 콜센터, 온라인/모바일 업무처리 시스템을 이용하며, 기업은 시스템 개발, 운영, 유지보수, 인력 등에 막대한 비용을 지불하고 있다. 이는 챗봇 민원 처리시스템이 가진 해결 과정을 보면 알 수 있는데, 챗봇이 처리할 수 있는 문제 해결의 데이터가 많지 않아 정해진 키워드에 맞는 단순한 처리 업무를 해결하는 데 그치고 있으며, 여러 다양한 민원을 처리하는 데 있어 한계를 드러내고 있다.

챗봇 민원 서비스를 구현하는 기술은 크게 두 가지로 구분할 수 있는데, 규칙 기반과 AI 기반이다. 첫 번째 규칙 기반 챗봇은 질문의 의도를 식별하고 유사성 분석과 검색 기술을 통해 답변을 찾는다. 반면에 AI 기술을 기반으로 한 챗봇은 사용자와 자연어 대화를 가능하게 하는 시스템으로 음성과 텍스트 모두 활용한다. 이러한 챗봇은 대규모 실제 대화 데이터를 학습하여 질문에 답변하며, 이는 금융 및 민원 서비스 등에 적용되고 있다(지동준, 2023).

LLM(Large Language Model)은 대규모 언어 데이터를 학습하여, 자연어 이해와 생성 능력이 크게 향상되었다. 또한, 사용자의 복잡하고 다양한 요구에 효과적으로 대응할 수 있게 되었다.

최근 복잡한 질문에도 응답할 수 있는 다양한 딥러닝 기반 챗봇이 구축되고 있으며, 그중 LLM 활용에 대한 검토가 활발히 이루어지고 있다. 하지만 LLM Hallucination의 문제를 제어하기 위해 망분리 환경에서의 학습 한계. 고객정보와 개인식별정보 학습 불가. 새로운 정보나 학습 이후 발생한 데이터 처리 미흡. 고비용의 GPU 비용 등 LLM fine-tunning에 있어 어려움이 많다(안성훈, 2022).

본 논문은 기존 민원 처리시스템인 챗봇 기술에서 나타난 한계를 해결하기 위해, 최신 LLM 기술을 활용한 민원 처리시스템 개선방안을 제시한다. 특히, 본 연구는 RAG와 GraphRAG 같은 최신 기술을 통합함으로써 LLM의 한계를 보완하고, 민원처리 정확성과 효율성 향상에 목표로 한다. 또한, 망분리 환경에서 안전한 데이터 처리를 위한 솔루션을 제공하고자 한다.

본 연구의 필요성으로 첫째, 복잡한 민원처리에 있어 기존의 규칙 기반 시스템의 한계를

극복할 수 있는 지능적이고 유연한 시스템 개발이 필수적이다. 둘째, 개인정보 보호와 규제 준수가 요구되는 망분리 환경에서 LLM을 안전하게 활용할 수 있는 기술적 해결책이 필요하다. 셋째, 고객의 요구가 점점 더 다양해지고 복잡해지는 상황에서 LLM 기반 AI 챗봇 시스템은 기업과 기관의 경쟁력을 높이는 중요한 도구로 자리 잡을 것이다.

따라서 본 연구는 기존 LLM 관련 연구의 한계를 명확히 인식하고, 이를 개선하기 위한 구체적인 기술적 방안을 제시함으로써, 민원처리시스템의 혁신 방향을 제시하고자 한다.

II. 이론적 배경

2.1 AI 챗봇 개념 및 동향

2.1.1 생성형 AI 챗봇

생성형 AI 챗봇은 대화형 AI로도 분류되며, 대화형 AI는 인간과 자연스럽게 상호작용할 수 있는 인공지능을 의미한다. (Ram, 2018)

<표 1>에 나타난 대화 스타일이나 사용 목적에 따라 분류된다. (Zaib, 2022; Clark et al., 2019).

<표 1> 대화 방식 및 사용 목적에 따른 대화형 AI (양나은, 2024)

사용 목적	대화 방식
사회적 목적 (Social Purposes)	채팅 지향적 (Chat-oriented)
업무적 목적 (Transactional Purposes)	과제 지향적 (Task-oriented)
	질의응답 (Q&A, Question Answer)

모바일 채팅 지향적(chat-oriented) 대화형 AI는 주로 자연스러운 양방향 대화를 제공하기 때문에 사회적 목적을 가진 것으로 분류된다. 반면, 과제 지향적(task-oriented) 대화형 AI 및 질의응답(QA) 대화형 AI는 비즈니스 목적을 가진 것으로 분류된다. Siri나 Bixby와 같은 과제 지향형 AI는 레스토랑 예약이나 일정 관리 등의 일상적인 작업을 처리하며, ChatGPT와 같은 질의응답 대화형 AI는 다양한 정보원을 기반으로 사용자 질문에 정확한 답변을 생성한다(양나은, 2024).

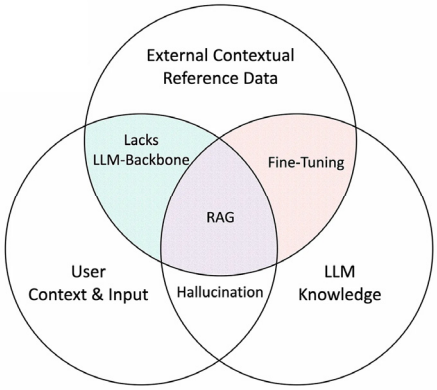
AI 챗봇은 사회적 대화를 주된 목적으로 설계되지만, 과제 수행이나 질의응답 하는 용도의 AI 챗봇은 단순히 대화형 인터페이스를 제공하는 데 그친다. 이러한 챗봇은 사용자 질문에 답변은 가능하지만, 스스로 질문을 생성하거나 제안하여 사용자와 상호작용을 하진 않는다. 반면, ChatGPT와 같은 질의응답 중심의 대화형 AI 모델은 이메일 작성, 에세이 작성, 코드 작성 등 다양한 업무를 수행할 수 있다. 또한, 자연어 처리 기능을 활용해 사회적 대화도 가능하다. 그러나 과제 지향적인 접근에서는 사회적 대화를 지원하는 양방향 상호작용이 줄어들고, 대화는 사용자의 입력(Prompt)에 따라 수행된다.

2.1.2 생성형 AI 챗봇 기술

2.1.2.1 LLM(Large Language Model)

인공지능 분야에서 ChatGPT와 같은 챗봇 서비스는 LLM의 발전으로 큰 관심을 받고 있다. 이러한 AI 모델은 대규모 언어 데이터를 바탕으로 훈련되어 자연어를 생성할 수 있다. 예를 들어, OpenAI의 GPT(Generative Pretrained Transformer) 시리즈나 Google의

BERT(Bidirectional Encoder Representations from Transformers) 같은 모델들이 있다. LLM은 기계학습과 딥러닝 기술을 기반으로 한다. 학습에는 대규모 데이터셋과 강력한 연산 능력이 필수적이다. 이 대형 언어 모델은 자연어 처리(NLP)와 자연어 생성(NLG) 작업에서 딥러닝을 활용하는 주요 모델로 사용된다. 또한, 언어의 복잡성과 관련성을 학습하는 능력을 갖추고 있다. 이 모델은 대규모 데이터로 사전 학습된 후, 세부 조정, 문맥 학습, Zero-shot, One-shot, Few-shot 학습 등의 기법을 사용한다.



<그림 1> RAG & Fine-Tuning

(출처 : https://cobusgreyling.medium.com/rag-fine-tuning-e541512e9601)

2.1.2.2 RAG(Retrieval Augmented Generation)

RAG는 외부 데이터베이스에 저장된 다양한 문서를 통합하여 복잡하고 구체적인 질문에 대해 관련된 정보를 찾아낸다. 이를 통해 신속하게 답변을 제공할 뿐만 아니라 특정 정보를 깊이 있게 이해하는 데 도움이 된다. RAG의 두 가지 주요 기술은 Vector Database와 Embedding 기술이다. 벡터 데이터베이스는 쿼리에 유사한 문서를 검색하기 위해 Vector Store를 구축하며, Embedding은 문서 데이터를 벡터 형식으로 변환하는 역할을 한다. 이 과정은 사용자의 질의를 바탕으로 유사한 문장을 찾아내는 벡터 유사도 검색 방식을 이용한다. RAG 활용은 검색과 생성을 결합하여 더욱 정확하고 의미 있는 텍스트를 생성할 수 있다. RAG 구성은 정보 검색의 장점과 텍스트 생성의 창의성을 결합하여 다양한 분야에 적용할 수 있다(정천수, 2023).

RAG의 최적 활용 방법은 Fine-Tuning과 결합하여 사용하는 것이다. <그림 1>에서 볼 수 있듯이, Fine-Tuning은 모델의 동작을 조정하는 역할을 하며, 추론 과정에서 외부 맥락 데이터를 참조하도록 지원한다. RAG와 Fine-Tuning의 조합은 최적의 해결책이 될 수 있다. 하지만 RAG는 Fine-Tuning과 비교해 데이터 추론 과정이 투명하여, LLM의 입력과 출력을 시각적으로 검토하고 추적하기 쉬워서 유용하다. Fine-Tuning은 LLM 학습 데이터를 세밀하게 조정하여 의료, 법률, 공학 등 여러 산업 분야에서 효과적으로 활용될 수 있다. 하지만 이렇게 조정된 모델은 일반적으로 높은 정확도를 유지하더라도 특정 사용자 입력에 최적화되지 않을 가능성이 있으며, 시간이 지나면서 고정된 특정 입력에 대해 맥락적 참조가 부족해질 수 있다. Fine-Tuning의 장점은 프롬프트*(AI 모델이 답변을 생성하는 데 사용하는 텍스트)에만 의존하지 않고 응답의 품질을 향상할 수 있다는 점이다. 단일 프롬프트에 모든 것을 맞추려 하기보다는 모델이 다양한 예제를 학습할 수 있게 한다.

2.1.2.3 GraphRAG

GraphRAG는 지식 그래프를 활용하여 세부

적인 지식 검색을 수행한다. 이는 분리된 정보 조각들을 연결하고 대규모 문서에 대한 종합적인 이해 개선을 목표로 한다. 지식 그래프를 사용하여, Cypher*(그래프 데이터베이스에 접근하기 위해 사용되는 그래프 쿼리 언어) 쿼리로 관련된 Node(개체)와 Edge(관계)를 검색하고, 언어 모델의 지식과 통합하는 과정을 수행한다. (Sarmah, 2024)

이 정보를 바탕으로 LLM이 질문에 대한 답변을 제공한다. GraphRAG와 RAG의 차이점은 지식 그래프에 접근하는 것으로, 이 그래프에는 텍스트뿐만 아니라 다양한 메타데이터와 관계가 저장되어 있다. GDB(Graph Database)는 복잡한 데이터 간의 관계를 더 효율적으로 처리하며, 이 점이 <표 2>에 나타나 있다. GraphRAG는 복잡한 문서 또는 Database에서 구조화된 정보를 효과적으로 처리하고 통합하는 데 장점이 있다.

<표 2> GDB 특징

구분	GDB
Schema	스키마는 애플리케이션과 함께 진화
Relationship	노드 사이의 관계를 엣지로 표현
Complex Query	속도가 빠르며 조인(Join) 불필요
Language	Cypher

(출처 : https://medium.com/@minji.sql/neo4j-소개-5a2c1b789190)

2.2 챗봇 민원처리 시스템 개요

2.2.1 주요 민원 데이터

디지털 자동화 기술 중 AI 기반 챗봇 기술이 통신 서비스 분야에 도입되면 서비스 사용자들의 접근성을 향상하고 24시간 서비스 지원 혜택을 제공함으로써 가치를 창출할 수 있다. AI 기술의 적극적 활용은 통신 서비스 분야에 상당한 기여할 것으로 예상한다. 사용자들은 특히 <표 3>에 보듯 통신상품, 정보조회, FAQ, 민원 서비스와 관련된 정보에 특히 관심이 있으며, 사용자 문의에 기반을 두어 정확한 응답과 실시간 응답을 기대한다. 통신 서비스 분야에서 온라인 서비스 활용 방법에 대한 안내 서비스는 사용자들에게 필수적인 요구 사항이다. 따라서 AI 기반 챗봇 서비스는 통신 서비스 분야에서 중요한 서비스 솔루션으로 간주할 수 있다.

<표 3> 국내 S통신사 민원 서비스

구분	요금상품	정보조회	FAQ
내용	현재 요금제 비교	기본 가입정보	로밍 신청 방법
	데이터/음성 잔여량	잔여 약정/ 할부계약	요금제 가입조건
	가입 가능한 요금제	지원금/선택 약정할인	이용가능 혜택
	요금제 변경	약정 위약금	요금제 문의
	요금제 옵션 변경	실시간/ 청구요금	멤버십 등급, 혜택

2.2.2 망분리 정책

“망분리”는 내부 정보 유출과 외부 인터넷망을 통한 불법적인 접근을 막기 위해 ‘내부망’과 ‘기관 인터넷망’을 분리하는 네트워크 보안 조치이다.(출저: 국가정보보안기본지침)

내부망과 기관 인터넷망을 분리하여 운영하는 기관은 국가정보원의 “국가 정보보안 기본지침”을 준수하는 공공기관, “전자금융감독규정”을 따르는 금융기관, “국방안보작전지령” 및 “국방사이버안보지령”을 따르는 국방부 또

는 계열기관, 그리고 “국방산업기술보호법”에 속하는 관련 기관 및 방위 산업 등이 포함된다. 특히 개인정보 처리에 있어서, 공공 및 민간 부문의 모든 개인정보 처리자는 “개인정보 보호법”에 따라 개인정보를 안전하게 처리하기 위한 보호조치를 취해야 한다.

또한, 개인정보 보호법 29조, 개인정보 기술관리적 보호조치 기준 제4조6항에 의거 전년도 말 기준 직전 3개월간 그 개인정보가 저장·관리되고 있는 이용자 수가 일일 평균 100만명 이상이거나 정보통신서비스 부문 전년도(법인인 경우에는 전 사업연도를 말한다) 매출액이 100억원 이상인 정보통신서비스 제공자 등은 개인정보 처리시스템에서 개인정보를 다운로드 또는 파기할 수 있거나 개인정보 처리시스템에 대한 접근 권한을 설정할 수 있는 개인정보 취급자의 컴퓨터 등을 물리적 또는 논리적으로 망분리 하여야 한다(이용재, 2016).

2.2.3 AI 모델 학습 과정 실태

최근 개인정보보호위원회(위원장 고학수, 이하 ‘개인정보위’)의 주요 인공지능(AI) 서비스에 대한 예비 조사 결과에 따르면, AI 모델이 정확한 응답을 제공하기 위해 많은 인력이 사용자 질문과 AI 모델의 응답을 직접 검토하고 수정하여 데이터셋을 생성하는 방식으로 훈련되고 있음이 확인되었다. 이러한 데이터셋은 AI 모델 학습 및 프롬프트 등 신속한 서비스 개선에 활용되고 있다. 그러나 입력된 데이터가 인적자원에 의해 검토되는 과정을 파악하기 어렵다. 중요한 개인정보나 이메일 주소와 같은 민감한 정보가 입력될 경우, 개인 식별자와 개인정보를 제거하는 조치 없이 AI 서비스 제공업체가 이를 데이터베이스로 변환할 때 개인정보 침해와 위험이 발생할 수 있음을 확인하였다. 참고 <표 4>

<표 4> 특정 AI 서비스 이용자 입력 데이터 분석 결과(‘23.12.30~’24.1.5.)

구분	건수	구분	건수
전화번호	672	주민등록번호	2
이메일주소	142	여권번호	34
		합계	850

AI 서비스는 이전의 서비스와는 다르게 개인정보 처리 항목, 방법, 목적, 보유, 이용 기간 등에서 큰 차이가 있다. 또한, LLM 복제 모델이나 오픈소스 형태로 배포될 때, LLM에 취약점이 발견되어도 즉시 개선에 어려울 수가 있다. 같은 LLM 기반의 AI 서비스라도 사업자에 따라 개인정보 및 아동, 민감정보에 대한 보호조치 수준이 다르다는 점이 확인되었다.

2.3 챗봇 민원 처리시스템 기술

2023년에 발표된 OpenAI ChatGPT는 사람과 같이 대화하는 경험을 제공한다. 이 때문에 각 사회영역에서 ChatGPT 모델인 LLM 사용 방식에 관한 연구가 활발히 이뤄지고 있다. 특히 자연어 처리(Natural Language Processing, NLP) 관련 다양한 작업을 수행할 수 있으며, 자연어 생성 측면에서 뛰어나고 풍부한 문맥을 이해하여 자연스러운 텍스트를 생성할 수 있다. 또한, 대규모 빅데이터와 높은 성능의 계산으로 딥러닝 학습 기술이 활용되었다. 챗봇은 비지니스의 중요한 창구이다. AI의 발전으로 NLP(Natural Language

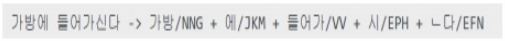
Processing): 자연어 처리기법에서 NLU (Natural Language Understanding): 자연어 이해기법, NLG(Natural Language Generation): 자연어 생성기법으로 진화하고 있다.

최근 챗봇은 머신러닝 또는 딥러닝 기술이 적용되어 사용자의 질문을 이해하고 의도 파악으로 적절한 답변을 할 수 있게 되므로 복잡하고 다양한 업무 수행을 할 수 있다. 챗봇은 규칙 기반 챗봇과 AI 기반 챗봇으로 동작 방식을 분류할 수 있다.

2.3.1 규칙 기반 챗봇 시스템 기술

2.3.1.1 자연어 처리 NLP

자연어 처리 단계에서 주요 작업은 형태소 분석(POS-Tagging)이다. <그림 2>와 같이 형태소 분석은 원시 말뭉치를 형태소로 분해하고 각 형태소에 품사 정보를 할당하여 형태소, 어근, 접두사/접미사 및 품사(POS, Part-Of Speech)와 같은 다양한 언어적 속성을 식별하는 것을 말한다.



<그림 2> 형태소 분석의 예시(김성근, 2018)

여기서 말뭉치는 특정 목적을 위해 추출된 언어 샘플 세트를 가리킨다. 형태소 분석 사전은 많은 단어를 검색 키로 사용하여 주어진 단어의 형태소 분석 결과를 검색할 수 있도록 만들어진 사전이다.

2.3.1.2 자연어 이해 NLU

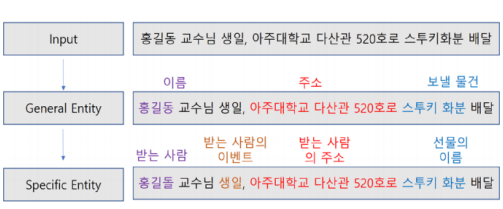
NLU는 자연어 이해를 의미한다. 이는 사용자의 입력을 NLP 모듈을 통해 분해하고 분해된 엔티티에 대한 적절한 구문 및 의미 구조를 결정하는 과정이다. NLU에는 일반개체명 인식, 세부 개체명 인식, 어휘 의미분석, 구문분석, 의도 분석과 같은 작업이 포함된다.

<그림 3>은 NLU의 기본 동작 과정을 보여주며, 구문을 분석하여 명명된 개체를 추출하고 사용자의 의도를 이해하는 것을 나타낸다.

<그림 3> 자연어 이해의 동작과정

(KMA: Korean Morphological Analyzer, NER : Named-entity Recognition) (김성근, 2018)

<그림 4>는 명명된 엔티티를 추출할 때, 챗봇은 도메인이나 작업에 효과적으로 작동하기 위해 세부 개체명 인식(Specific-NER) 시스템이 필요할 수 있다.

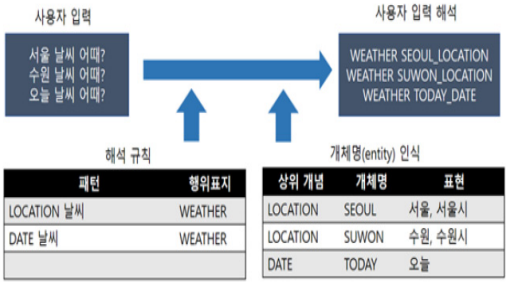


<그림 4> 세부 개체명 인식 예시(김성근, 2018)

2.3.1.3 규칙 기반 챗봇

규칙 기반 챗봇 <그림 5>는 미리 정의된 규칙을 기반으로 챗봇이 동작한다. 사용자의 입력을 해석하는 규칙과 사용자의 입력에 대해 반응하는 규칙, 응답을 생성하는 규칙을 갖고 있다.

규칙 기반 챗봇은 개발 시 많은 양의 데이터가 필요하지 않고 상대적으로 구현이 쉽다. 하지만 대화가 한정적이고 구조화되어 있다. 또한 규칙의 정도에 따라 높은 품질의 대화 서비스가 가능하다. 이는 적은 변수와 간단한 질문에 적합하므로 높은 품질의 대화를 위해서는 규칙을 잘 정의하는데 많은 시간과 인력이 필요하다.

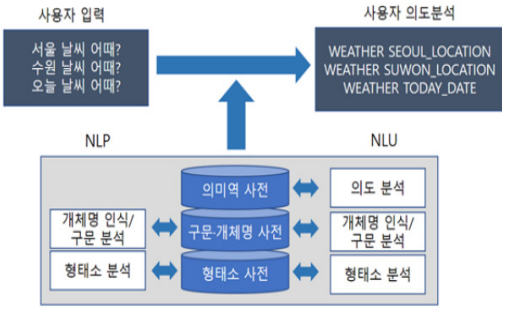


<그림 5> 규칙 기반 챗봇(김성근, 2018)

2.3.2 AI 기반 챗봇 시스템 기술

2.3.2.1 AI 기계학습 기반 챗봇

AI 기계학습 기반 챗봇 <그림 6>은 자연어 이해를 바탕으로 적절한 답변 생성 또는 답변 선택을 위한 기계학습 알고리즘을 사용한다. NLP 모듈은 사용자가 입력한 자연어를 형태소 분석 등을 분해한다. NLU 모듈은 분해된 자연어의 의미를 파악, 분석하여 정제된 정보로 최종 분류 모델에 전달하여 정의된 답변을 수행한다. 기계학습 기반 챗봇은 높은 수준의 대화가 가능한 만큼 높은 자연어 처리 기술이 필요하다. NLP 모듈과 NLU 모듈은 많은 언어학적 분석과 학습 데이터가 필요하다. 또한, 언어 데이터 특성에 맞게 구현하며, 학습된 NLP와 NLU는 다른 분야와 업무 재활용에 어려움이 있다.



<그림 6> AI 기계학습 기반 챗봇(김성근, 2018)

2.4 기존 챗봇 민원 처리시스템의 한계

기존 챗봇 민원 처리시스템은 시나리오 기반으로 제한된 범위의 입력과 답변만을 제공하고 있다. 이에 따라 사용자가 원하는 다양한 질문과 민원처리에 대응하는 데 높은 개발 비용과 한계에 직면하고 있다.

통신 서비스와 유사한 국내외 주요 은행들은 <표 5>와 같이 다양한 고객 민원 서비스를 확대하고 있다. 고객 요구에 따라 ChatGPT와 같은 첨단 기술의 영향으로 비대면 챗봇 서비스에 관한 관심이 계속해서 증가하고 있다. 이에 기업의 챗봇 서비스는 앞으로 업무가 확대될 것으로 예상한다. 챗봇을 활용한 통신 서비스 부문에서 우수한 서비스를 제공하기 위해서는 서비스 품질을 종합적으로 관리하는 것이 필요하다. 통신 서비스 품질이 비즈니스 성과와 지속가능성을 보장하는 것과 마찬가지로 통신 서비스 부문에서의 서비스 품질 향상과 고객 신뢰는 서로 연관되어 있다. 고객에게 정확한 정보를 제공함으로써 회사에 대한 신뢰가 증가하게 된다. 이 긍정적인 사이클은 고객의 만족과 삶의 질을 향상할 것이다.

LLM 모델은 학습 중에 지식을 뉴럴 신경

망에 저장한다. 학습 시간까지의 지식을 기반으로 질문에 답할 수 있지만, 학습되지 않은 데이터나 학습 이후에 나오는 데이터에 관한 질문 답변에는 한계가 있다. 또한, 잘못된 정보를 학습하여 잘못된 답변을 제공할 위험이 있으며, 이를 Hallucination(환각) 효과라고 한다. 이 두 가지 문제를 해결하기 위해 최신 데이터와 정확한 데이터가 필요하다.

<표 5> 국내외 주요은행 챗봇 서비스 현황

회사명 (국가)	서비스
NH농협 은행(한국)	상품안내, 자주 묻는 질문, 이벤트 안내, 이용시간 안내, 올원뱅크 바로가기 등을 카카오톡 기반으로 채팅을 통해 자동상담
뱅크오브 아메리카 (미국)	알림 서비스 등
토시카은행 (러시아)	잔액조회, 요금지불, 인근 현금자동입출금기(ATM) 위치 안내, 고객상담
압사은행 (남아프리카 공화국)	잔액조회, 최근 지출조회, 통신사 데이터 추가 구입, 자금이체
스코틀랜드 국립은행 (스코틀랜드)	카드분실 관리, 잠긴계정 관리, 고객상담
루나웨이 (덴마크)	잔액조회, 아마존 지출조회

출처 : 각 은행, 금융보안원

Ⅲ. AI 챗봇 민원처리 시스템 개선방안

3.1 VectorRAG와 GraphRAG

일반적으로 사용되는 RAG 방식인 Vector RAG는 텍스트, 이미지, 오디오 등 구조화되지 않은 데이터를 효과적으로 처리할 수 있으며. 데이터의 의미를 파악하는 데 특히 도움이 된다. 벡터 데이터베이스는 높은 확장 가능성과 빠른 검색 속도를 제공하며, 벡터 유사성을 기반으로 정보를 신속하게 검색할 수 있는 특징이 있다. 또한 데이터의 벡터 표현이나 임베딩을 활용하는 기계학습 모델과 결합했을 때 높은 효율성을 나타낸다. 그 결과 VectorRAG는 크고 복잡한 데이터 세트를 처리할 때 강한 이점을 가진다.

GraphRAG의 지식 그래프는 구조화된 데이터 엔티티 간의 복잡한 관계를 효과적으로 처리하고 활용하는 데 적합하다. 데이터 포인트 사이의 상호작용과 관련 정보의 중요성이 큰 상황에서 유용하다. 또한, 특정 도메인의 전문적 지식을 계층적으로 효과적으로 나타낼 수 있다. 지식 그래프는 데이터의 무결성을 유지하고 일관된 데이터 표현을 제공하는 데 적합한 도구이다(최재철, 2024).

3.2 HybridRAG

하이브리드 방식은 벡터 데이터베이스의 유연성과 확장성 및 지식 그래프의 계층적 특성을 모두 활용할 수 있다. HybridRAG는 VectorRAG와 GraphRAG의 장점을 결합하여 검색의 정확성을 높이고 관련성 있는 응답을 생성함으로써 데이터의 전반적인 품질을 향상하는 것을 목표로 한다. HybridRAG는 두 단계로 구성되어 있다. 첫 번째 단계인 VectorRAG에서는 텍스트의 유사성을 기반으로 컨텍스트를 검색하며, 이 과정에서 데이터를 작은 조각으로 나누어 각 조각을 벡터 데이터베이스에 저장된 벡터 임베딩으로 변환한다. 이후 데이터베이스에서 유사성 검색을

수행하여 가장 관련성 높은 조각들을 찾아 순위를 매긴다. 두 번째 단계인 GraphRAG에서는 지식 그래프를 활용해 데이터에서 구조화된 정보를 추출하고, 데이터 내의 엔티티와 관계를 나타낸다.

HybridRAG는 VectorRAG와 GraphRAG를 결합하여 언어 모델의 응답이 문맥적으로 정확하고 세부 정보가 충실한 응답을 생성하도록 한다.

HybridRAG의 효과는 Nifty 50 지수에 등재된 회사의 수익 전화 회의록 데이터 세트를 기반으로 광범위한 실험을 통해 입증되었다. <표 6>에서는 VectorRAG, GraphRAG, HybridRAG의 성능을 비교하였으며, 주요 지표로는 충실도(F:Faithfulness), 답변 관련성(AR:Answer Relevance), 맥락 정확도(CP:Context Precision), 맥락 회상(CR:Context Recall)이 사용되었다. 분석 결과 HybridRAG는 여러 지표에서 VectorRAG와 GraphRAG를 능가하는 성과를 보였다. 충실도 측면에서 HybridRAG는 0.96의 점수를 기록하여 생성된 응답이 제공된 문맥과 잘 일치함을 입증했다. 또한, 답변 관련성 면에서 HybridRAG는 0.96점을 기록하며 VectorRAG(0.91)와 GraphRAG(0.89)를 앞섰다. GraphRAG는 맥락 정확도에서 0.96점을 기록하며 우수한 성과를 보였으나, HybridRAG는 맥락 회상에서 VectorRAG와 함께 1.0이라는 완벽한 점수를 달성하여 뛰어난 성능을 입증하였다.

<표 6> RAG 성능지표 (출처 : Sarmah, 2024)

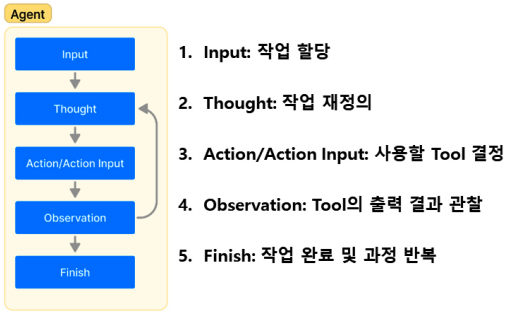
구분	VectoRAG	GraphRAG	HybridRAG
F	0.94	0.96	0.96
AR	0.91	0.89	0.96
CP	0.84	0.96	0.79
CR	1	0.85	1

실험 결과 벡터 기반과 그래프 기반 검색 기술의 강점을 결합하여, 정확하고 맥락적으로 적합한 답변을 제공할 수 있음을 보여주었다(Sarmah, 2024).

3.1 LangChain Agent/Tool

LLM은 학습 중에 저장된 데이터를 기반으로 응답을 생성한다. LLM 애플리케이션은 데이터 확장을 위해 LLM에 저장되지 않은 정보를 Google 검색엔진을 활용하거나 YouTube 비디오 스크립트를 참조할 수 있다. 또한, 회사의 내부 데이터베이스에서 정보를 조회하여 답변을 제공하기도 한다. 이러한 필요를 충족하기 위해 LangChain은 LLM이 외부 정보를 활용할 수 있도록 Agent와 Tool 같은 구성 요소를 제공한다. 이 기능은 기본 LLM에 외부 Tool을 통합하여 LLM의 기능을 확장할 수 있게 하는 LangChain의 중요한 기능 중 하나이다.

Agent는 사용자의 입력(발화)을 바탕으로 스스로 판단하여 적절한 작업을 선택하고 수행하는 역할을 한다. 제공된 Tool을 활용해 목표를 달성할 때까지 작업을 수행한다.



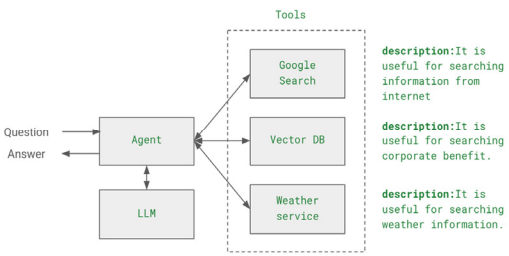
<그림 7> Agent 작업 수행 과정

(출처 : https://pkgpl.org/2023/10/26/langchain-tool들을-사용하는-agent/)

<그림 7>은 Agent가 작업을 수행하는 과정을 설명한다. 이 과정의 핵심은 LLM을 활용하여 일련의 작업을 선택하는 것에 있다. 필요한 정보를 파악하고 질문을 재구성한 후, 적절한 Tool을 사용하여 정보를 추출한다. 이후 이 정보를 분석하여 응답을 생성할 수 있는지를 판단하며, 추가 정보가 필요할 경우 질문을 다시 설정하고 적합한 Tool 선택 과정을 반복하여 최종 응답에 도달한다.

Tool을 사용하는 것은 LLM의 현재 한계를 극복할 수 있는 실질적인 방법을 제공한다. 첫째, LLM은 검색엔진 쿼리를 활용하여 최신 정보를 생성할 수 있다. 둘째, 데이터의 출처를 인용함으로 환각 문제를 줄이고 LLM의 신뢰성을 높일 수 있다. 셋째, LLM의 응답을 도출하기 위해 사용하는 API 호출을 추적하여 어느 정도 해석 가능성을 제공할 수 있다.

구글 검색엔진, 벡터 데이터베이스 및 내부 회사 데이터베이스와 같은 Tool들은 각각의 인터페이스나 API를 통해 정보를 제공한다. <그림 8>에서 볼 수 있듯이, Agent가 질문을 받으면 답변 방법을 고민한다. 이 과정에서 LLM을 사용한다. 답변이 LLM을 통해 제공될 수 있다면 직접 응답한다. LLM이 답변을 제공할 수 없다면, 등록된 외부 Tool들을 참조한다. 각 Tool은 description이라는 필드에서 수행할 수 있는 기능을 설명한다. 예를 들어, 누군가가 1988년 서울 올림픽의 최고의 선수들에 대해 알고 싶어하고 정보가 없다면, LLM은 Google 검색 도구를 사용하여 원하는 정보를 검색 결과에서 추출한다.

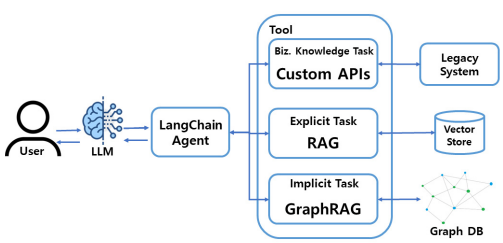


<그림 8> Agent 개념

(출처 : https://bcho.tistory.com/1426)

3.2 AI 챗봇 아키텍처 제안

<그림 9>와 같이 다양한 Tool의 기술은 민원 데이터의 업무처리, 정보제공, 질의응답 등 특성에 기반을 둔다. Agent가 답변을 생성하는 과정에 있어서 데이터 출처를 Biz. Knowledge Task, Explicit Task, Implicit Task의 3가지 Tool 방식으로 나타낼 수 있다. Biz. Knowledge Task는 Custom APIs에 사용되는 방식으로 Legacy System에서 API를 이용하여, 정확한 정보를 전달한다. 각 업무 비즈니스 특성에 맞게 제공됨으로 각 업무처리 수행에 적합하다. Explicit Task는 RAG에 사용되는 방식으로 Vector Store의 저장된 문서정보를 제공함으로 FAQ와 같은 문장정보 제공에 적합하다. Implicit Task는 GraphRAG에 사용되는 방식으로 Graph DB는 노드(node)와 엣지(edge)를 사용한 데이터와 데이터 관계를 직관적으로 표현한다. 소셜 네트워크, 추천 시스템 등 복잡한 관계의 정보를 제공하는 데 적합하다.



<그림 9> 아키텍처 제안

Agent가 각 Tool의 특성에 맞게 응답을 생성할 수 있도록, LangChain의 Tool Decorator를 사용해 Tool을 정의할 수 있다. Tool Decorator는 docstring으로 Tool 정의를 작성하면, Agent가 이를 활용해 응답을 생성하게 된다. LangChain의 다양한 기능을 얼마나 효과적으로 활용하느냐에 따라 응답 성능이 결정될 것이다.

제안된 아키텍처는 민원 데이터 작업의 특성을 분석하고 망분리 환경을 고려하였다. 이를 바탕으로 다음과 같은 구현 전략을 제안한다. 민원 데이터는 고객정보와 개인 식별 정보를 포함한 망분리 환경에서 암호화 및 복호화 모듈과 함께 안전한 API를 제공하는 것을 전제로 한다. Agent와 Tool을 활용하여, 기본 LLM에 없는 외부 데이터베이스를 연결함으로써 LLM의 기능을 확장한다. 또한, RAG와 GraphRAG 기술을 적용하여 환각 현상을 줄이고 풍성한 문장 생성을 하게 함으로써, 보다 안정적이고 맥락에 적합한 응답을 제공할 것으로 기대된다.

Ⅳ. 결론 및 시사점

본 논문은 민원 데이터를 바탕으로 LLM 기반 AI 챗봇 시스템의 주요 기술 동향을 분석하였다. 전통적인 시나리오 기반 챗봇의 한계를 극복하고 복잡한 질의를 처리할 수 있는 LLM 기반 AI 챗봇 시스템의 핵심 기술인 RAG, GraphRAG, HybridRAG, Agent/Tool 전략을 살펴보았다.

이를 위해 민원 데이터의 특성을 분석하고, 망분리 환경을 고려한 구현 전략을 제안하였다. 고객정보와 개인 식별 정보를 포함한 민원 데이터는 망분리 환경에서 암호화 및 복호화 모듈과 안전한 API 제공을 전제로 한다. 또한, Agent와 Tool을 활용하여 LLM에 외부 데이터베이스를 연결함으로써 LLM 기능을 확장하였다.

RAG와 GraphRAG 기술은 대규모 데이터에서 정확한 정보를 찾아 의미 있는 답변을 생성하는 데 효과적인 도구로 활용된다. 이 두 기술의 장점을 결합한 HybridRAG는 벡터 데이터베이스의 유연성과 확장성, 지식 그래프의 계층적 구조를 활용하여, LLM의 환각 문제를 줄일 수 있다. 또한, Agent/Tool의 사용은 정확한 답변을 제공하게 함으로써 보다 안정적이고 맥락에 맞는 응답 결과를 제공할 것으로 기대된다.

하지만 이러한 기술의 구현 복잡성과 데이터 품질이 성능에 큰 영향을 미친다. 따라서 기업과 기관들은 기술의 성공적 도입을 위해 충분한 준비와 투자가 필요하다. 또한, 망분리 환경에서의 데이터 처리와 개인정보 보호 문제도 중요한 고려사항으로, 이를 해결하기 위한 체계적인 접근이 필요하다.

결론적으로, 본 논문에서 제안한 LLM 기반 AI 챗봇 시스템은 민원 처리시스템의 효율성과 정확성을 크게 높일 가능성을 지니고 있다. 향후 연구에서는 기존 시스템 업무 수행에 필요한 LLM과 시스템 간의 인터페이스에 관한 심층적인 연구가 필요하다. LLM이 최종 답변을 생성하는 과정에서 제공되는 데이터를 활용하여, 추론 단계를 얼마나 어떻게 활용하는지 예측하는데 명확하지 않다. 따라서 앞서 제안한 Tool을 분석하고 평가하는 데 도움이 될 데이터와 개발 방법론, 기술 구성이 중요하다. Tool의 활용은 LLM의 현재 한계를 보완하고, AI 챗봇 시스템이 더욱 광범위하게 활용될 수 있는 기반을 마련할 것이다.

참고문헌

안성훈(2022). 대학 학사운영의 효과적인 안내를 위한 챗봇 시스템 구축 방안 탐색, *한국창의정보문학학회,* 8(3), 145-152.

이주승(2022). 망분리 환경에서 민감정보를 안전하게 처리하기 위한 기술적 방안. *융합보안논문지,* 23(1), 3-7.

김성근(2018). 챗봇 기술 소개 및 사례 분석 = 챗봇 개념, 기술 및 사례를 알아보자, *한국통신학회,* 35(2), 1-4.

신상수(2019). 자연어 처리 기반의 음악 추천 챗봇, *한국정보처리학회,* 26(1), 573-575.

정천수(2023). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG모델과 LangChain 프레임워크 기반, *한국지능정보시스템학회,* 29(4), 129-164.

양나은(2024). 생성형 AI 챗봇의 혼합주도적 상호작용에 따른 학습 효과와 사용자 경험에 대한 연구, *한국문화융합학회,* 46(1) 85-98.

이철승(2023). 인공지능 챗봇 발전에 따른 AI 리터러시 필요성 연구, *한국전자통신학회,* 18(3), 421-426.

김계수(2023). 공공서비스 품질 개선과 행정만족도 증진을 위한 AI채봇 개발 -관광 AI챗봇을 중심으로, *한국고객만족경영학회,* 25(1), 79-96.

지동준(2023). 메타버스 환경에서 지식 그래프 기반 AI 챗봇을 이용한 민원 서비스 구현 및 최적화, *한국산학기술학회,* 24(8), 299-305.

김홍비(2023). 거대언어모델과 문서검색 알고리즘을 활용한 한국원자력연구원 규정 질의응답 시스템 개발, *한국산업정보학회,* 28(5), 31-39.

이용재(2016). 개인정보 보호조치에 관하여 - 옥션 판결의 기술적·관리적 보호조치와 인과관계 판단을 중심으로, 사법발전재단, 1(38), 481-538.

김영승(2024). *특허 데이터를 이용한 인공지능 분야 주요 기술 동향 분석과 시사점 : 생성형 AI 기반 챗봇 기술을 중심으로* (석사학위논문, 고려대학교).

김슬기(2022). *인공지능 기반 행정서비스 수용 영향 요인 연구 : 챗봇 기반 민원행정서비스를 중심으로* (석사학위논문, 서울대학교).

지침서(2023.01.31). *국가사이버안보센터,* 국가정보보안기본지침.

해설서(2020.12). *개인정보위원회,* 개인정보의 기술적, 관리적 보호조치 기준 해설서.

보도자료(2024.03.28). *개인정보위원회,* 주요 인공지능(AI) 서비스 사전 실태점검 결과 발표.

조대협(2024.02.02). Agent/Tool 을 이용하여 ChatGPT와 구글 검색엔진 연동하기. *조대협의 블로그,* Retrieved from https://bcho.tistory.com/1426

plusha(2023.10.26.). LangChain:Tool들을 사용하는 Agent. *PKGPL(Pukyong Geophysical Prospecting Laboratory),* Retrieved from https://pkgpl.org/2023/10/26/langchain-tool들을-사용하는-agent/

쪼변(2020.06.23). [망분리] 망분리의 모든 것. 쪼변의 라이브러리, Retrieved from https://blog.naver.com/mint860703/222009286392

mingTato(2023.09.19). Neo4j 소개.

mingTato Medium, Retrieved from https://medium.com/@minji.sql/neo4j-소개-5a2c1b789190

Cobus Greyling(2023.09.27.). RAG & Fine-Tuning. Cobus Greyling Medium, Retrieved from https://cobusgreyling.medium.com/rag-fine-tuning-e541512e9601

최재철(2024.07.13.). 기존RAG 한계극복. GraphRAG란?, *brunchstory,* Retrieved from https://brunch.co.kr/@b2439ea8fc654b8/38

Agents Concepts. (2024). *LangChain,* Retrieved from https://python.langchain.com/v0.1/docs/modules/agents/concepts/

Ram(2017). Conversational AI: The Science Behind the Alexa Prize, *Alexa Prize.*

Zaib(2022). Conversational question answering: a survey, *Knowledge and Information Systems* 64(12).

Clark(2019). What makes a good conversation? Challenges in designing truly conversational agents, *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1-12.

Sarmah(2024). *HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction,* arXiv:2408.04948v1.

Nikhil(2024.08.12.). HybridRAG: A Hybrid AI System Formed by Integrating Knowledge Graphs and Vector Retrieval Augmented Generation Outperforming both Individually. MARKTECHPOST, Retrieved from https://www.marktechpost.com/2024/08/12/hybridrag-a-hybrid-ai-system-formed-by-integrating-knowledge-graphs-and-vector-retrieval-augmented-generation-outperforming-both-individually/