

DAT 3기 캡스톤 프로젝트 보고서

RAG모델과 LangChain프레임워크를 활용한

의료 챗봇 시스템 구현

: 증상 기반 질환명 정보 전달을 중심으로

DAT 3기

연어

이주은

이다빈

황서진



Data Analysis & Technology



DAT 3기 캡스톤 프로젝트 보고서

RAG모델과 LangChain프레임워크를 활용한

의료 챗봇 시스템 구현

: 증상 기반 질환명 정보 전달을 중심으로

Medical Chatbot System implementation

using RAG model and LangChain framework

: Focusing on Disease Information Based on Symptoms

이 보고서를 캡스톤 프로젝트 보고서로 제출합니다.

2024년 05월 30일

DAT 3기

연어

이주은

이다빈

황서진



DAT 3기 캡스톤 프로젝트 보고서

RAG모델과 LangChain프레임워크를 활용한

의료 챗봇 시스템 구현

: 증상 기반 질환명 정보 전달을 중심으로

국문 요약

본 연구는 RAG(Retrieval-Augmented Generation) 모델과 LangChain 프레임워크를 활용하여 증상 기반 질환명을 제공하는 의료 챗봇 시스템을 구현하는 것을 목표로 한다. 인간과 기계 간 상호작용의 중요성이 증가하면서 대형 언어 모델(LLM)의 한계, 특히 환각 현상을 극복하기 위해 RAG 모델을 도입하였다. 서울 아산병원의 질환 백과 데이터를 크롤링하여 전처리 후, RAG 기반 접근 방식을 구현하였다. RAG 기반 챗봇은 사용자의 증상 입력에 대해 보다 정확한 질환 정보를 제공하였으며, 의료 조치의 신속성을 높였다. 본 연구는 데이터 다양성과 실시간 업데이트의 필요성을 강조하며, 향후 다양한 의료기관의 데이터 통합과 사용자 인터페이스 개선을 제안한다.

Abstract

This study aims to implement a medical chatbot system that provides disease information based on symptoms using the Retrieval-Augmented Generation (RAG) model and LangChain framework. With the increasing importance of human-machine interaction, the RAG model was introduced to overcome the limitations of large language models (LLMs), particularly hallucination issues. After crawling and preprocessing data from Seoul Asan Hospital's disease encyclopedia, the RAG-based approach was implemented. The RAG-based chatbot provided more accurate disease information based on user input symptoms and improved the speed of medical response. This study emphasizes the need for data diversity and real-time updates, proposing the integration of data from various medical institutions and improvements in the user interface for future research.



목 차

제 1장 서론	1
제 1절 연구 배경	1
제 2절 연구 목적 및 논문 구성	1
제 2장 이론적 배경	2
제 1절 생성형 AI 주요 요소 이해	2
제 2절 RAG 모델	3
제 3절 벡터 데이터베이스	4
제 3장 챗봇 시스템 구현	5
제 1절 데이터 전처리	5
제 2절 코사인 유사도 기반 구현	6
제 3절 RAG 기반 구현	8
제 4장 결론	9
제 1절 연구 결과	9
제 2절 한계점 및 제언	12
참고문헌	13



표 목차

[표 3-1] 코사인 유사도로 학습한 상위 단락, 유사도	7
[표 3-2] gpt-3.5-turbo를 사용한 추가 답변 생성	8
[표 3-3] RAG 기반 최종 답변 생성	9
[표 4-1] RAG 기반 챗봇 구현 결과	10
[표 4-2] 코사인 유사도 기반 챗봇 구현 결과	11



제 1 장 서론

제 1 절 연구 배경

인간과 기계간의 상호작용이 점차 발전하면서, LLM(Large Language Models, 대형 언어 모델)을 기반으로 한 대화형 챗봇 시스템이 우리의 일상생활에서 더욱 중요한 역할을 하고 있다. LLM 은 방대한 텍스트 데이터 학습을 통해 복잡한 자연어 작업을 수행할 수 있으며, 이를 통해 고객 서비스, 창의적 콘텐츠 생성, 질의응답 등 다양한 분야에 적용되고 있다.

그러나, 많은 챗봇 시스템은 사용자의 요구에 부합하는 정보를 제공하는 데 한계가 있고, 사용자 맞춤형 질의응답을 제공하는 데 거짓 정보, 최신 정보 처리의 어려움을 겪고 있다. 또한 AI 의 학습 및 생성 과정에서 생겨나는 환각 현상(Hallucination)은 중요한 문제 중 하나다. 즉, 자신이 모르는 사실을 적당히 지어내서 얘기하는 것이다. 따라서 언뜻 그럴듯해 보이지만 실제로는 잘못된 정보일 경우가 있다. 최신 정보와 상충되는 답변을 생성할 수도 있으며 이러한 환각은 대표적인 생성형 AI 의 한계이고, 이전에 학습한 내용을 사용하므로 새로운 문제나 도메인에 대한 이해도가 낮을 수 있다.

이러한 문제점을 해결하기 위해 검색 증강 생성(Retrieval Augmented Generation, RAG) 서비스 아키텍처를 사용한다. RAG 는 LLM 에게 미리 질문과 관련된 참고자료를 알려줌으로써 환각을 줄이고 보다 정확하게 답변을 생성할 수 있도록 한다. 이렇게 RAG 모델은 사용자의 질의를 더 정확하고 적절한 답변을 제공하여 실제 비즈니스 환경에서도 유용하게 활용될 수 있는 기술로 다양한 도메인에 적용할 수 있다.

제 2 절 연구 목적 및 논문 구성

인공지능 챗봇 서비스는 이전부터 다양한 영역에서 활용되고 있지만, 최근 신종 코로나바이러스 감염증(코로나 19)으로 인해 비대면 서비스가 늘어나며 그 역할은 더욱 활발해졌다. 특히 환자 급증, 진료 시스템 수용 능력 등의 문제 해결을 위해 의료계에서도 인공지능 챗봇 서비스 도입을 적극적으로



추진하고 있다. 코로나 19 팬데믹 이후, 많은 사람들이 증상이 나타났을 때, 이것이 코로나 19 때문인지 아니면 단순한 감기인지 혼란스러워하는 상황이 자주 발생했다. 코로나 19 와 감기는 증상이 유사할 수 있기 때문에, 사람들이 자신이 겪고 있는 증상을 정확하게 이해하고 올바른 조치를 취하는 것이 매우 중요해졌다. 본 연구는 이를 위해 부위별 증상 기반 의료 챗봇 시스템을 통해 사용자가 자신의 건강 상태에 대해 신속하게 이해하고 적절한 의료 조치를 빠르게 취할 수 있도록 돕고자 한다.

본 보고서의 구성은 다음과 같다. 2 장에서 생성형 AI 의 이론적 배경을 설명하고, RAG 모델의 주요 단계에 대해 알아본다. 또 벡터 DB 를 활용하여 정보 저장 및 검색 방법에 대해 논의한다. 3 장에서는 구체적인 구현 방법으로 데이터 전처리 과정, 코사인 유사도 기반과 RAG 모델 기반으로 구현하는 방법을 설명하고 LangChain 프레임워크를 활용하여 사용자의 질의와 답변을 생성한다. 마지막으로 결론은 4 장에서는 연구 결과 및 한계점, 제언에 대해 기술하고 있다.

제 2 장 이론적 연구

제 1 절 생성형 AI 주요 요소 이해

2.1.1 파운데이션 모델

현재는 이미지와 텍스트를 동시에 학습하는 멀티모달 모델들의 성능과 기능이 업그레이드되고 있고 기초모델로 자리를 잡아 가고 있다. 파운데이션 모델은 Bommasani등 스탠포드대 여러 연구자들이 파운데이션 모델들의 기회와 리스크라는 논문을 발표하면서 처음 제안하였다. 파운데이션 모델의 데이터는 텍스트, 이미지, 음성, 정형데이터, 3D 시그널 등 구분하지 않고 학습에 이용되며, 인간의 창의력과 추론력을 포함한 일을 수행하고 이러한 기초모델을 AI 패러다임의 변화와 파운데이션 모델이라고 논하고 있다. 여기서 많은 양의 데이터는 비지도 학습을 통해 모델을 학습시킨 후 배포되어 사용자가 원하는 목적에 맞게 다운스트림 작업에 대해 파인튜닝이나 문맥 내 학습 등과 같은 과정을 거쳐 완성되는 것이 파운데이션 모델이라고 볼 수 있다.



2.1.2 LLM

생성형 AI 모델은 대규모 언어모델로 텍스트와 같은 언어 데이터를 학습하여 결과를 제공하여 LLM은 최근 인공지능 분야에서 큰 주목을 받고 있다. OpenAI의 GPT 시리즈 및 Google의 BERT등이 대표적인 LLM 모델이다. LLM은 NLP, NLG 작업에서 딥러닝을 활용하는 기반 모델로서, 언어의 복잡성과 연결성을 학습할 수 있도록 돕기 위해 대규모 언어 모델은 방대한 양의 데이터에 대해 사전학습되며 파인튜닝, In-context learning, Zero/one/few-shot learning 같은 기술을 사용한다. LLM의 학습단계는 데이터 수집 및 전처리, 모델 선택 및 구성, 모델학습, 평가 및 파인튜닝으로 진행된다.

2.1.3 프롬프트 엔지니어링

프롬프트 엔지니어링은 LLM을 사용하여 창의적인 콘텐츠를 생성하는 방법이다. 또한, LLM의 창의적인 텍스트 생성 능력을 향상시키기 위해 LLM에 제공되는 프롬프트의 품질을 개선하는 데 중점을 준다. 프롬프트의 지시문에 따라 포함시키는 예시가 없는 경우인 Zero-Shot, 예시가 1개인 One-shot, 2개 이상인 Few-shot Learning으로 구분되며 다양한 예시를 제공할수록 일정 수준 이상의 답변을 얻을 수 있다.

제 2 절 RAG 모델

대규모 언어모델(Large Language Models, LLMs)은 자연어 처리(NLP) 분야에서 획기적인 발전을 이루었으며, 다양한 응용 분야에서 놀라운 성과를 보여주고 있다. 그러나 이러한 모델들은 몇 가지 중요한 한계점을 가지고 있고 그 중에서도 가장 두드러진 문제는 '환각(hallucination)' 현상이다. 환각 현상은 모델이 실제로 존재하지 않거나 부정확한 정보를 생성하는 것을 의미한다. 이는 모델이 입력 데이터나 질문에 대한 올바른 답변을 제공하는 대신, 그럴듯해 보이지만 실제로는 잘못된 정보를 만들어내는 경우를 말한다. 이러한 문제는 특히 신뢰성과 정확성이 중요한 분야에서 큰 문제를 야기할 수 있다.

이 문제를 해결하기 위해 생성형 AI의 뇌 역할을 할 수 있는 RAG(Retrieval-Augmented Generation) 기법이 대두되고 있다. RAG는 대규모 언어모델에 외부 지식베이스를 통합하여 실시간으로 정보의 정확성을 확인하고 최신 데이터를 제공할 수 있도록 하는 접근 방식이다. 이를 통해 모델이 생성하는 정보의 신뢰성과 정확성을 크게 향상시킬 수 있다. RAG 모델은 텍스트 생성 과정에서 외부 지식베이스를 활용하여 보다 정확한 정보를 제공하는 것을 목표로 한다. 이 모델은 입력 쿼리에 대해 관련 문서를 검색하는 검색기와 이 문서들을 바탕으로 텍스트를 생성하는 생성기로 구성된 질의응답 시스템 구조이다.



RAG 모델은 먼저 문서, 웹 페이지, 뉴스 기사 등 다양한 형태의 소스 데이터를 수집하여 이를 처리한다. 이 과정에서 데이터를 작은 조각(청크)으로 분할하고, 각 청크를 벡터 형태로 변환하는 임베딩(embedding)을 수행한다. 생성된 청크는 텍스트를 고정된 크기의 벡터로 변환하여 벡터 데이터베이스에 저장된다. 이러한 벡터 데이터베이스는 검색 및 유사성 계산을 효율적으로 수행할 수 있게 한다. 사용자가 질문을 입력하면, 해당 질문을 임베딩하여 벡터 데이터베이스에서 유사한 벡터를 검색하고 검색된 청크들은 원래 텍스트 형태로 디코딩되어 정보가 추출된다. 이렇게 검색된 정보를 기반으로 텍스트를 생성하는 과정에서 생성할 텍스트의 종류, 길이, 언어적 스타일 등을 지정할 수 있다. RAG 모델은 정보 검색과 생성 과정을 통합하여 보다 정확하고 의미 있는 텍스트를 생성할 수 있도록 설계되었다. 최종적으로, 검색된 청크를 조합하여 응답을 생성한다. 이 과정에서 시스템 프롬프트를 사용하여 보다 정확한 답변을 생성하도록 구성할 수 있다. 예를 들어, 질문에 대해 "모른다"고 답변하도록 프롬프트를 설정하면, 모델은 없는 정보를 지어내지 않고 정확한 답변을 제공할 수 있다. 이를 통해 LLM 의 정보 부족 및 환각 문제를 효과적으로 해결할 수 있고, 신뢰성 있고 정확한 답변을 제공한다는 장점을 가지고 있다.

제 3 절 벡터 데이터베이스

벡터 데이터베이스는 LLM 의 장기 기억 부족 문제를 해결하기 위해 개발된 새로운 유형의 데이터베이스이다. LLM 은 입력된 텍스트를 벡터로 변환한 후, 이 벡터를 기반으로 예측하거나 생성 작업을 수행한다. 벡터 임베딩은 고차원 벡터로 표현되는데, 고차원 공간에서 유사한 벡터를 빠르게 찾기 위해서는 특수한 인덱싱 기법이 필요하다. 하지만 전통적인 데이터베이스는 이러한 인덱싱 기법을 제공하지 않기 때문에 검색 성능이 떨어진다. 따라서 벡터 데이터베이스는 벡터 임베딩을 효율적으로 관리하고 비정형 및 반정형 데이터 관리를 위한 완벽한 솔루션을 제공한다

벡터 데이터베이스에서는 유사도 메트릭(Similarity Metrics)을 적용하여 가장 유사한 벡터를 찾는다. 일반적인 벡터 데이터베이스의 파이프라인은 다음과 같다.

- ① 색인(indexing): PQ, LSH, HNSW 와 같은 알고리즘을 사용하여 벡터를 색인한다. 이 단계에서는 더 빠른 검색을 가능하게 하는 데이터 구조에 벡터를 매핑한다.
- ② 쿼리(Querying): 유사도 메트릭(코사인 유사도, 유클리디안 거리, 내적 등)을 적용하여 가장 가까운 벡터를 찾는다
- ③ 후처리(Post Processing): 데이터셋에서 최종 최근접 이웃을 검색하고 이를 후처리하여 최종 결과를 반환한다. 일부 벡터 데이터베이스는 벡터 검색을 하기 전에 필터를 적용할 수 있다.

벡터 데이터베이스는 벡터 임베딩을 저장하고 색인화하며, 거리 및 유사성 검색 등의 다양한 기능을 제공하기 때문에 필수적이다. 즉, 벡터 데이터베이스는 비정형 데이터와 반정형 데이터를 효과적으로 관리하도록 설계되어 있으며 이러한 특성 덕분에 벡터 데이터베이스는 머신 러닝 및 AI 기반 디지털 환경에서 중요한 역할을 한다.

제 3 장 챗봇 시스템 구현

제 1 절 데이터 전처리

3.1.1 자료수집

분석에 사용된 자료는 서울 아산병원에서 제공하는 질환백과에서 부위별 질환명, 증상, 원인, 관련질환, 진료과를 크롤링하여 데이터를 수집하였다. 두 가지 기법을 기반으로 챗봇을 구현하여 전처리도 각각 다른 방식으로 진행하였다.

3.1.2 코사인유사도 기반 구현

데이터에 존재하는 결측치는 '-'로 대체하고 각 행의 모든 값을 문자열로 변환하여 결합한 후, 텍스트의 온점을 제외한 특수문자를 제거하는 전처리 과정을 진행했다. 이 때, 불용어 제거는 수행하지 않았다. 이는 데이터를 크롤링하는 과정에서 필요한 내용만 선별하여 수집했기 때문에, 수집된 데이터에 불필요한 단어들이 거의 없다고 판단하여 불용어를 제거하지 않았다. 그 후, SentenceTransformer 모델을 사용하여 문장의 임베딩을 생성하였다.

3.1.3 RAG 기반 구현

데이터셋 내 결측값들은 fillna() 메서드를 사용하여 '-'로 대체하였다. 각 텍스트 컬럼을 문장 단위로 분리하여 가장 긴 문장을 추출하였으며, 이 과정에서 텍스트 데이터의 정제와 탐색이 이루어졌다. 이후, RecursiveCharacterTextSplitter 를 이용해 텍스트를 일정 크기의 청크로 분할하여 모델 학습에 적합한 형태로 변환하였다. 마지막으로, OpenAI 임베딩을 사용하여 텍스트를 벡터화하고, 이를 Chroma 벡터 스토어에 저장하여 질의 응답 시스템을 구축하였다.

제 2 절 코사인 유사도 기반 구현

RAG 기반으로 챗봇 시스템을 구현하기에 앞서 문장 유사도를 기반으로 구현해보기 위해 코사인 유사도를 사용했다.

코사인 유사도는 두 벡터 간의 코사인 각도를 측정하여 유사도를 계산하는 방법이다. 본 연구에서는 코사인 유사도를 활용하여 사용자의 질문 벡터와 임베딩 컬럼의 각 벡터들 간의 각도를 계산하여 상위 1 개의 단락을 추출한다. 코사인 유사도는 아래 식을 따른다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

이 식에서 두 벡터가 완전히 같은 방향을 가리킬 경우 코사인 유사도는 1, 90도의 각을 이룰 경우 0, 반대 방향을 가리킬 경우 -1 이 된다. 코사인 유사도는 벡터의 크기가 아닌 방향에만 의존하게 되므로, 벡터의 크기에 영향을 받지 않고 유사도를 계산할 수 있다는 장점이 있다. 유사도의 값은 -1 에서 1 까지의 범위를 가지며, 1 에 가까울수록 두 벡터는 서로 유사하다고 판단하며, -1 에 가까울수록 서로 다르다고 판단한다.

코사인 유사도를 활용하여 '무기력감, 37.5 도 이상의 고열, 기침, 인후통, 가래, 근육통, 두통, 호흡곤란, 폐렴 등의 증상이 발생해.'라는 질문을 통해 상위 1 개의 단락을 추출했다. 그 결과, 코로나바이러스 증상을 인용해서 질문을 했지만 가슴 부위의 '폐부종(Pulmonary edema)'에 대해 정제되지 않은 문장으로 답변하는 것을 확인할 수 있다.

질문	무기력감, 37.5 도 이상의 고열, 기침, 인후통, 가래, 근육통, 두통, 호흡곤란, 폐렴 등의 증상이 발생해.
답변	가슴. 폐부종(Pulmonary edema) 폐부종은 저산소증과 심한 호흡 곤란을 야기하는 상태로 질환이라기보다는 증상에 가깝습니다. 폐부종을 일으키는 원인은 다양합니다. 이들 원인에 의해 폐정맥 및 모세 혈관에서 폐의 간질 조직과 폐포로 체액이 빠져나갑니다. 이로 인해 폐포와 기도가 침범당하여 가스 교환이 악화됩니다. 이러한 증상이 저산소증과 심한 호흡 곤란을 야기하는 것입니다. 폐부종 원인은 심인성과 비심인성으로 구분할 수 있습니다. 심인성은 심장에 의한 폐부종을 말하며 승모판 협착증 좌심실 부전 등이 이에 해당합니다. 비심인성은 심장에 의한 폐부종이 아닌 경우를 말하며 저 알부민혈증 림프계의 차단 또는 성인 호흡 곤란 증후군처럼 여러 기전에 의해 폐포-모세혈관막의 투과성이 증가된 경우를 말합니다. 폐부종의 증상으로는 기침 객담 호흡 곤란 등이 있습니다. 호흡 곤란은 특히 누웠을 때 악화될 수 있습니다. 심한 경우에 청색증과 말초 부종이 동반되기도 합니다. 증상이 경미한 경우 감기로 보일 수 있으나 증상이 지속되고 호흡 곤란의 악화 말초 부종 등의 증상이 나타난다면 병원을 방문해야 합니다.
유사도	0.694480836391449

[표 3-1. 코사인 유사도로 학습한 상위 단락, 유사도]

상위 1 개의 단락의 내용을 바탕으로 대화형 거대언어모델인 OpenAI 의 ChatGPT 의 API 를 활용하여 최종 답변을 생성했다. 모델을 활용할 때, 모델에 역할을 부여하고 질문과 관련된 문서 단락을 알려줌으로써 모델이 목적에 맞게 답변을 생성할 수 있도록 필요한 정보와 지침을 명시한 프롬프트(prompt)를 제공한다. 프롬프트에 system, user, assistant 세 가지 역할을 부여할 수 있다.

- system: 초기 설정이나 지침을 제공하여 대화의 흐름과 내용을 목적에 맞게 이끌어줌
- user: 사용자의 질문이나 명령을 제공
- assistant: 사용자와 이전 대화 내용을 저장하여 모델이 문맥을 이해할 수 있도록 함

본 논문은 system 기능에 역할을 부여하여 “다음 내용을 참조하여 사용자의 질문에 간단히 답변할 것. 문서와 질문에 대한 요약(summaries)이 주어지면 참조('SORUCES')를 사용하여 최종 답변을 작성할 것. 답을 모르면 '모른다'고 말하고 답을 만들어 내려고 하지 말것.” 라는 문구를 추가하였다. 이전의 대화 내용을 기억할 수 있는 assistant 의 기능은 사용하지 않았다.

모델은 'gpt-3.5-turbo'를 사용하였고 질문은 [표 3-1]과 동일하게 코로나바이러스 증상을 작성하였다. 그 결과, 코사인 유사도를 계산하여 추출한 답변보다 더 정제되고 완전한 문장을 구사하는 것을 확인할 수 있다.

질문	무기력감, 37.5도 이상의 고열, 기침, 인후통, 가래, 근육통, 두통, 호흡곤란, 폐렴 등의 증상이 발생해.
답변	가슴 폐부종은 저산소증과 호흡 곤란을 일으키는 상태이며 주로 폐정맥 및 모세 혈관에서 체액이 빠져나가 폐의 간질 조직과 폐포를 침범합니다. 이로 인해 가스 교환이 악화되어 호흡 곤란이 발생할 수 있습니다. 이러한 증상이 나타나면 즉시 병원을 방문하는 것이 좋습니다.

[표 3-2. gpt-3.5-turbo 를 사용한 추가 답변 생성]

제 3 절 RAG 기반 구현

RAG 모델은 제 2 장 2 절에서 살펴본 바와 같이 주어진 질문이나 주제에 관련된 정보를 검색하여 가져오고, 이를 기반으로 응답을 생성하는 데 활용되는 검색 증강 생성 모델이다. 코사인 유사도 기반

구현과 동일하게 프롬프트를 구성하고 코로나바이러스 증상 중 일부를 인용하여 최종 답변을 생성하였다. 결과는 [표 3-3]과 같다.

질문	무기력감, 37.5도 이상의 고열, 기침, 인후통, 가래, 근육통, 두통, 호흡곤란, 폐렴 등의 증상이 발생해.
답변	위 증상은 호흡기 감염의 가능성이 있습니다. 이러한 증상은 코로나바이러스같은 호흡기 감염병의 일부 증상일 수 있습니다. 따라서 의료진과 상담하여 검사를 받는 것이 좋습니다.

[표 3-3. RAG 기반 최종 답변 생성]

문장 구사는 코사인 유사도 기반과 비슷하나 질문에 대한 답변을 코로나바이러스로 정확하게 응답한 것을 확인할 수 있었다. 이는 검색된 정보를 바탕으로 답변을 생성하는 RAG 기법의 장점을 활용하여 더 나은 답변을 만들어 낸 유의미한 결과라고 볼 수 있다.

제 4 장 결론

제 1 절 연구 결과

서울아산병원 홈페이지의 질환백과에서 부위별로 총 1600 개의 질환을 크롤링하여 데이터를 구성하였다. 전처리, 토큰화, 임베딩 단계를 거친 후 RAG 기반 챗봇과 코사인 유사도 기반 챗봇을 구현하였다. 챗봇은 사용자가 증상을 입력하면 적절한 질환명과 정보를 반환하도록 구성했다.

질문	답변
무기력감, 37.5도 이상의 고열, 기침, 인후통, 가래, 근육통, 두통, 호흡곤란, 폐렴 등의 증상이 발생해	위 증상은 호흡기 감염의 가능성이 있습니다. 이러한 증상은 감기나 독감과 관련이 있을 수 있지만, 폐렴과 같은 심각한 합병증을 유발할 수도 있습니다. 증상이 심해지면 의료진과 상담하여 적절한 치료를 받는 것이 중요합니다
눈 가려움, 충혈, 눈부심, 눈곱등의 증상이 발생해	이러한 증상은 거의 항문성 충수염이나 눈 건조증과 관련이 있을 수 있습니다. 하지만 정확한 진단을 위해서 안과 전문의를 방문하시는 것이 좋습니다.
입술에 물집, 가려움, 따가운 증상이 발생해	입술에 물집, 가려움, 따가운 증상이 발생한다면 헤르페스 바이러스로 인한 단순포진이나 피부염 등의 피부 질환의 가능성이 있습니다. 이러한 증상이 지속되거나 심해진다면 피부과 전문의를 방문하여 진단과 치료를 받는 것이 좋습니다.
귀에 통증, 귀울림, 먹먹함이 느껴져요.	귀에 통증, 귀울림, 그리고 먹먹함이 느껴진다면 이는 이명(티락스)이나 이명과 관련된 이명성 난청의 가능성이 있습니다. 이러한 증상이 지속된다면 이비인후과 전문의를 방문하여 상세한 진단과 치료를 받는 것이 좋습니다.
숨을 들이쉬고 내쉴 때 휘파람 소리가 나.	이러한 증상은 기관지 천식과 관련이 있을 수 있습니다. 기관지 천식은 기관지가 염증으로 인해 좁아지고 점막이 붓는 상황으로, 호흡 시에 휘파람 소리가 발생할 수 있습니다. 이는 의사와 상담하여 정확한 진단과 치료를 받는 것이 중요합니다.

[표 4-1. RAG 기반 챗봇 구현 결과]

질문	답변
무릎의 통증입니다. 특히 계단을 오르내릴 때 통증이 더 심합니다. 통증은 대개 오전보다 오후에 더 심합니다.	무릎의 통증은 계단을 오르내릴 때 더 심해지고, 대개 오후에 더 심해진다고 하네요. 이러한 증상은 슬관절염과 관련이 있을 수 있습니다. 슬관절염은 슬개골과 대퇴골 사이의 연골이 마모되거나 염증이 발생할 때 발생하는 질환입니다. 이로 인해 통증과 관절가동이 제한될 수 있습니다. 증상이 계속되면 전문가의 진료를 받는 것이 좋을 것입니다.
가래에 혈액이 묻어 나와요.	혈토(Coughing up blood)는 가래에 혈액이 섞인 상태로 나오는 증상을 말합니다. 이는 다양한 요인에 의해 발생할 수 있으며, 즉각적인 치료가 필요할 수 있습니다. 이러한 증상이 나타날 경우 즉시 의료진과 상담하시는 것이 중요합니다.
설사, 구토를 하며 복부가 팽만한 느낌이 있고 기운이 없어요.	복부 팽만과 기운 부족은 여러 가지 이유로 발생할 수 있습니다. 설사와 구토가 동반되면서 이러한 증상이 나타나는 경우 식중독, 위장 통증, 소화불량 등이 의심될 수 있습니다. 이러한 증상이 계속되거나 심해진다면 즉시 의료진 상담이 필요합니다.
귀에 통증, 귀울림, 먹먹함이 느껴져요.	귀에 통증, 귀울림, 먹먹함과 같은 증상은 이명(티너이러스)의 가능성이 있습니다. 귀에 이상 증상이 지속된다면 정확한 진단을 위해 이비인후과 전문의를 방문하시는 것이 좋습니다.
숨을 들이쉬고 내쉴 때 휘파람 소리가 나	숨을 들이쉬고 내쉴 때 휘파람 소리가 나는 증상은 기도나 천연동기 손상으로 인해 발생할 수 있는 호흡기 질환을 의심해볼 수 있습니다. 이러한 증상은 만성 기관지염, 천식, 폐렴, 폐색전증 등 여러 호흡기 질환과 관련이 있을 수 있습니다. 이러한 증상이 지속된다면 호흡기내과 전문의를 방문하여 적절한 진단과 치료를 받는 것이 중요합니다.

[표 4-2. 코사인 유사도 기반 챗봇 구현 결과]

RAG 기반 구현 모델과 코사인 유사도 기반 구현 모델 모두 사용자가 입력한 증상에 대해 적절한 질환명과 정보를 전달하고 진료과를 추천했다.

RAG 기반 모델은 외부 지식베이스와의 연동을 통해 질환을 예측하기 때문에, 비교적 복잡한 증상 조합

에 대해 더 정확한 진단을 내릴 수 있지만 응답 속도가 다소 느릴 수 있다. 반면, 코사인 유사도 기반 모델은 입력된 증상 벡터와 기존 데이터의 유사성을 측정하여 질환을 예측하기 때문에 단순하고 빠르지만 증상 조합이 복잡해 질수록 예측의 정확도가 떨어질 수 있다.

‘숨을 들이쉬고 내쉴 때 휘파람 소리가 나’라는 동일한 질문에 RAG 기반 구현 모델은 기관지 천식이라는 구체적인 진단을 제시했다. 반면, 코사인 기반 구현 모델은 만성 기관지염, 천식, 폐렴, 폐색전증 등 여러 호흡기 질환을 제시한 것을 확인할 수 있다.

제 2 절 연구 한계점 및 제언

본 연구는 사용자가 입력한 증상을 기반으로 질환을 예측하고 정보를 전달하기 위해 RAG 기반 챗봇과 코사인 유사도 기반 챗봇을 구현했다. 본 연구의 한계점은 다음과 같다. 첫 번째로 사용된 데이터셋이 서울아산병원 질환백과에서 가져온 것으로, 특정 의료기관의 자료에 한정되어 있어 다양한 의료기관의 데이터를 충분히 반영하지 못했다. 이는 증상 및 질환 패턴을 충분히 고려하지 못해 결과의 일반화에 한계를 초래할 수 있다. 두 번째로 질병 정보와 증상에 대한 최신 데이터를 실시간으로 반영하기 어렵기 때문에 두 모델 모두 드문 질병이나 비정형적 증상에 대해 적절히 대응하지 못할 수 있다. 마지막으로 사용자가 증상을 정확히 입력하지 못할 경우, 모델이 성능이 저하될 수 있다.

이러한 한계점을 보완하기 위해 다양한 의료기관의 데이터를 통합하여 데이터의 다양성과 신뢰성을 높이는 것이 필요하다. 또한, 데이터의 최신성을 유지하기 위해 정기적으로 데이터를 업데이트하고 새로운 질병이나 변종에 대한 정보를 반영하는 시스템을 구축해야 한다. 또한 사용자가 증상을 정확하게 입력하지 못할 경우를 대비하여 사용자 인터페이스를 개선하고 사용자가 증상을 보다 쉽게 입력할 수 있도록 증상 입력을 도와주는 자동 완성 기능을 도입할 수 있다. 사용자 경험을 향상시키기 위해 사용자 피드백을 체계적으로 수집하고 이를 반영하는 시스템을 구축하는 것도 향후 연구의 과제로 남아 있다. 이를 통해 챗봇의 성능을 개선하고 사용자 만족도를 향상시킬 수 있을 것이다.

참고문헌

1. 조찬영, 강성준, 정현준. (2023-11-23). RAG기반 랭체인을 이용한 생성형 AI 챗봇 구현. Proceedings of KIIT Conference, 제주.
2. 정천수. (2023). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG모델과 LangChain 프레임워크 기반. 지능정보연구, 29(4), 129-164.
3. 추승연. (2021). 의료 인공지능 챗봇 상담 서비스의 가치가 이용의도에 미치는 영향. 국내석사학위논문 이화여자대학교 대학원, 서울
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.
5. 정천수. (2023). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG 모델과 LangChain. *지능정보시스템학회 논문지*, 29(4), 129-164.
6. Elastic. What is a Vector DB. <https://www.elastic.co/kr/what-is/vector-database>
7. 김홍비, 유용균. (2023). 거대언어모델과 문서검색 알고리즘을 활용한 한국원자력연구원 규정 질의응답 시스템 개발. 한국산업정보학회논문지, 28(5), 31-39.

