JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 12, pp. 169-180, December 2024
https://doi.org/10.9708/jksci.2024.29.12.169

Clinical Trials Utilizing LLM-Based Generative AI

Hyon-Chel Jung*, Kun-Soo Shin**, Ho-Dong Kim***, Sung-Bin Park****

*Research Professor, Dept. of Institute of Artificial Intelligence and Big Data in Medicine, Yonsei University Wonju College of Medicine, Korea
**Researcher, Dept. of Yonsei University Future Medical Industry Cooperation Group, Korea
***Executive Vice President, Head of AI, Corporate Research Institute, Solbit Co., Ltd., Korea
****Professor, Dept. of Precision Medicine, Yonsei University Wonju College of Medicine, Korea

[Abstract]

This study explores the improvement of work efficiency and expertise by applying Private LLM based on Large Language Model (LLM) to the field of clinical trials in medical devices. The Private LLM system provides sophisticated and accurate answers based on clinical data and shows its potential for use in various applications such as decision support, clinical expert activity assistance, new content generation, and problem solving. The study consists of the following four main steps. First, data specific to clinical trials of medical devices are collected, preprocessed, and organized into a learnable format. Second, based on open-source LLM models such as LaMA, PEFT (LoRA) and RAG techniques are applied to build a customized private LLM Q&A system for a specific clinical domain. Third, it realizes expert-level Q&A function by utilizing the established system and solves complex questions and problems that arise during clinical trial operation. Finally, by evaluating the performance of the system, we propose a direction to increase the efficiency and reliability of clinical trial operation and medical device development. As a result of the study, the Private LLM system has outperformed the existing methodology in supporting task automation and precise decision making. In particular, the ability to provide accurate answers to questions from domain experts and to generate new clinical criteria and insights shows the potential to become an innovative tool in clinical trial operation of medical devices. This confirmed the practical applicability of Private LLM in precision medical care, automation of clinical trial management, and a Q&A system based on domain knowledge.

▸ **Key words**: Large Language Models(LLMs), Generative AI, Clinical Trials, Medical Devices, Data Analysis

• First Author: Hyon-Chel Jung, Corresponding Author: Sung-Bin Park
*Hyon-Chel Jung (bravojhc@yonsei.ac.kr), Dept. of Institute of Artificial Intelligence and Big Data in Medicine, Yonsei University Wonju College of Medicine
**Kun-Soo Shin (tlsrjstn788@naver.com), Dept. of Yonsei University Future Medical Industry Cooperation Group
***Ho-Dong Kim (hodong.kim@solbit.kr), Corporate Research Institute, Solbit Co., Ltd.
****Sung-Bin Park (sung.b.park@gmail.com), Dept. of Precision Medicine, Yonsei University Wonju College of Medicine
• Received: 2024. 11. 01, Revised: 2024. 12. 02, Accepted: 2024. 12. 18.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 연구는 대규모 언어 모델(LLM) 기반의 Private LLM을 활용하여 의료기기 임상시험 분야에 적용하여 업무 효율성과 전문성 향상을 탐구한다. Private LLM 시스템은 임상 데이터를 기반으로 정교하고 정확한 답변을 제공하며, 의사결정 지원, 임상 전문가 활동 보조, 새로운 콘텐츠 생성, 문제 해결 등 다양한 응용 분야에서 활용 가능성을 보여준다. 연구는 다음 네 가지 주요 단계로 구성된다. 첫째, 의료기기 임상시험에 특화된 데이터를 수집하고 이를 전처리하여 학습 가능한 형식으로 정리한다. 둘째, LLaMA와 같은 오픈소스 LLM 모델을 기반으로 PEFT(LoRA) 및 RAG 기법을 적용하여 특정 임상 도메인에 맞는 맞춤형 Private LLM 질의응답 시스템을 구축한다. 셋째, 구축된 시스템을 활용하여 전문가 수준의 질의응답 기능을 실현하고, 임상시험 운영 중 발생하는 복잡한 질문과 문제를 해결한다. 마지막으로, 시스템의 성능을 평가하여 임상시험 운영과 의료기기 개발의 효율성과 신뢰성을 높이기 위한 방향성을 제안한다. 연구 결과, Private LLM 시스템은 기존의 방법론 대비 업무 자동화와 정밀한 의사결정 지원에서 탁월한 성능을 보였다. 특히, 도메인 전문가의 질문에 대한 정확한 답변을 제공하고, 새로운 임상 기준 및 인사이트를 생성할 수 있는 능력은 의료기기 임상시험 운영의 혁신적 도구로 자리잡을 가능성을 보여준다. 이를 통해 정밀 의료, 임상시험 관리 자동화, 그리고 도메인 지식 기반의 질의응답 시스템에서 Private LLM의 실질적 활용 가능성을 확인하였다.

▸ **주제어**: 대규모 언어 모델, 생성형 인공지능, 임상시험, 의료기기, 데이터 분석

I. Introduction

대규모 언어 모델(Large Language Models, LLM)의 발전은 최근 몇 년간 인공지능(AI) 기술의 중요한 진전으로 자리 잡고 있다 [1]. LLM 기반 생성형 AI는 자연어 처리와 텍스트 생성, 데이터 분석 등 다양한 분야에서 뛰어난 성능을 보이고 있으며, 특히 의료기기 임상시험에서 그 효용성이 점점 더 높아지고 있다. 기존의 임상시험 데이터 분석 방식은 대규모 데이터를 처리하는 데 있어 시간적, 비용적 한계가 있었고, 이는 임상시험의 신뢰성과 효율성에도 부정적인 영향을 미쳤다.

LLM 기반 생성형 AI는 기존 방법의 한계를 해결할 수 있는 도구로 떠오르고 있다 [2]. 이 기술은 방대한 임상 데이터를 분석하고, 그 안에서 중요한 패턴을 신속하게 찾아냄으로써 더 나은 의사결정을 가능하게 한다. 특히 임상시험 과정에서 데이터의 정확한 처리가 이루어져야 하는 만큼, LLM 기반 분석은 기존 방식과 비교할 때 매우 효과적이다. LLM을 통해 자동화된 데이터 분석이 가능해지며, 임상시험 전반에서 더 신뢰할 수 있는 결과를 도출할 수 있다.

의료기기 임상시험은 제품의 안전성과 유효성을 검증하는 중요한 과정으로, 데이터 분석의 정확성과 신뢰성이 매우 중요한 역할을 한다. 기존의 통계적 분석 방법은 대규모 데이터를 효과적으로 처리하는 데 한계가 있었고, 이는 임상시험 과정에서 오류와 지연을 초래할 수 있었다. LLM 기반 생성형 AI는 이러한 문제를 해결할 수 있는 가능성을 보여주고 있다. 특히 LLM은 기존 통계적 방법론과 달리, 데이터의 양과 복잡성에 구애받지 않고 패턴을 분석하고 중요한 정보를 추출할 수 있다.

본 연구는 LLM 기반 생성형 AI가 임상시험에 어떻게 적용될 수 있는지에 대해 구체적인 방법을 제시하고, 이를 통해 임상시험의 신뢰성과 효율성을 어떻게 향상시킬 수 있는지를 분석한다 [10]. 구체적으로 LLaMA와 같은 오픈소스 모델을 사용하여 임상 데이터를 분석하고, 이를 통해 더 신속하고 정확한 결과를 도출하는 방법을 제안한다. 또한 이러한 기술이 의료기기 개발과 임상시험에서 혁신을 가져올 수 있는 가능성을 평가하며, 향후 임상시험에서 LLM 기술의 확장 가능성도 논의한다.

LLM 기반 기술은 단순히 데이터를 분석하는 것 이상의 의미를 지니고 있다. 임상 데이터에서 발견된 패턴을 바탕으로 의사결정을 돕고, 궁극적으로는 임상시험의 전반적인 효율성을 높일 수 있는 중요한 도구로 활용될 수 있다.[5] 이 연구는 LLM 기술이 임상시험뿐만 아니라 다양한 의료 분야에서 데이터 분석의 패러다임을 변화시킬 수 있는 가능성을 제시한다. 향후 LLM을 활용한 의료기기 개발 및

임상시험에서의 발전 방향을 제시하며, 이를 통해 의료 분야 전반에서 혁신적인 변화를 이끌어내고자 한다.

II. Preliminaries

1. Limitations of Traditional NLP Approaches and Advantages of LLMs in Healthcare

기존의 자연어 처리(NLP) 방법은 통계적 분석기법과 결합하여 의무기록, 임상 노트, 의료 문헌 등의 비정형 임상시험 데이터를 분석하는 데 기본적인 역할을 해왔습니다. 그러나 현대의 의료 및 임상 작업에 적용할 경우, 특히 대규모 언어 모델(LLM)의 기능과 비교할 때 상당한 한계에 직면합니다.

- 기존 NLP 알고리즘은 종종 사전 정의된 규칙 기반 시스템에 의존했으며, 이는 다양한 의료 도메인에 대한 적응력이 떨어졌습니다.
- 의료 데이터의 미묘한 차이와 복잡성을 이해하거나 맥락을 보존하는 데 어려움이 있었습니다.
- 의료 용어의 다양성과 지역적 규제 차이를 다루는 데 제약이 있었습니다.
- 기존 NLP는 새로운 의료 도메인에 적응하려면 대규모의 도메인 특화 데이터와 레이블 작업이 필요합니다.

Table 1과 같이 대규모 언어 모델(LLMs)은 기존 NLP의 한계를 극복하며, 의료 분야에서 이러한 한계를 극복하는 방법으로 LLM접근 방법을 강조합니다.[17][18][19]

Table 1. Comparison of Traditional NLP Methods and Large Language Models (LLMs) in Healthcare

Features	Traditional NLP	LLMs
Accuracy	Relies on rule-based systems	Excels in understanding medical contexts
Data Handling	Limited to text-based data	Capable of integrating multimodal data
Adaptability	Requires extensive domain-specific training	Adapts quickly with Few-shot and Zero-shot learning
Applications	Basic text analysis	Supports clinical decision-making, medical education

또한 Table 2와 같이 기존의 자연어 처리(NLP) 방법이나 통계기법이 혼용된 머신러닝 기법은 본 연구과제와 유사한 적용분야인 의료분야의 개인화된 질의시스템을 만드는데 있어서는 GPT4 이후의 모델에 대해서는 정확도와 F1 Score에 있어서 현저히 성능이 떨어진다.[20]

Table 2. Comparison of NLP/ML Approaches and LLM Models in Accuracy and F1 Score

NLP/ML Approach	Accuracy	F1
Pretrain-BertwithXGBoost	0.326	0.303
TextCNN	0.437	0.429
HierarchicalAttention	0.495	0.477
TextBiLSTMwithAttention	0.512	0.5
RoBERT	0.585	0.543
LLM Base Approach	**Accuracy**	**F1**
GPT-4(few-shot)	0.62	0.671
Fintuned-LLaMA-2-7B	0.71	0.593
Fintuned-LLaMA-2-13B	0.73	0.671

따라서, 본 연구에서는 의료기기 임상시험에서 활용되는 정형, 비정형 데이터 분석과 의사결정 지원을 자연어의 맥락을 이해하지 못한 통계적 방법에서 탈피하고, 기존의 NLP에서 접근 한계를 극복하고자 LLM 접근으로 시도하였다.

2. Related works

2.1 Domestic trends

한국에서는 대규모 언어 모델(LLM)을 임상시험에 적용하는 연구가 학계와 연구 기관을 중심으로 활발히 진행되고 있다. KAIST, 서울대학교, 연세대학교와 같은 주요 대학들이 의료 데이터 분석과 예측 모델링에 LLM을 통합하기 위한 노력을 주도하고 있다. 예를 들어, KAIST는 대규모 임상 데이터를 분석해 질병 결과를 예측하는 LLM 기반 모델을 개발했으며, 이 모델은 향후 임상시험에서 중요한 역할을 할 것으로 예상된다.

한국보건산업진흥원(KHIDI) 또한 LLM을 활용해 의료 데이터 분석을 통합하는 여러 프로젝트를 지원하고 있다. 이러한 프로젝트는 대규모 데이터에서 숨겨진 패턴을 발견해 임상시험의 효율성과 정확성을 향상시키는 것을 목표로 하고 있다. 한국 의료 부문에서 LLM의 도입이 가속화되고 있으며, 이는 Table 3와 같이 AI가 임상 실무를 혁신할 잠재력이 크다는 것을 보여준다.

Table 3. Key Applications of LLM in Clinical Trials - domestic trends

Institution	Application Area	Description
KAIST	Disease Outcome Prediction	Developing a model to predict disease outcomes using LLM
Seoul National University	Medical Data Analysis	Using LLM for comprehensive analysis of medical data
Korea Health Industry Development Institute (KHIDI)	Integrated Medical Data Analysis	Supporting an LLM project to develop clinical trial policies

2.2 A foreign trend

전 세계적으로 LLM 기반 생성형 AI의 임상시험 통합이 빠르게 진행되고 있으며, 주요 기술 기업과 연구 기관들이 이를 주도하고 있다. Table 4와 같이 Google Health와 IBM Watson Health는 대규모 전자 건강 기록(EHR)을 분석해 진단 정확성과 질병 예측을 개선하기 위해 LLM을 활용하고 있다 [3]. Google Health의 EHR 분석에서 LLM을 활용한 결과는 질병 패턴을 정확히 식별해 임상시험 설계를 더욱 정밀하게 할 수 있다는 것을 보여준다.

영국에서는 DeepMind의 AlphaFold가 제약 산업에 큰 영향을 미치고 있다. AlphaFold는 단백질 접힘 패턴을 정확하게 예측하여 신약 개발에 중요한 역할을 하고 있다 [4]. 이와 같은 성과는 LLM이 임상시험을 변화시키고 새로운 치료법 개발을 가속화할 잠재력을 보여준다.

미국의 Stanford University와 MIT와 같은 학술 기관들도 LLM의 생의학 연구 활용 가능성에 많은 투자를 하고 있다. 그들의 연구는 LLM이 생의학 데이터에서 전통적인 분석 방법으로는 발견할 수 없는 복잡한 패턴을 식별함으로써 임상시험의 정확성을 높일 수 있음을 입증하고 있다 [7].

Table 4. Key Applications of LLM in Clinical Trials - international trends

Institution	Application Area	Description
Google Health	EHR Analysis	Improving disease prediction and diagnostic accuracy using LLM
DeepMind	Protein Folding Prediction	Applying LLM to accurately predict protein structures
Stanford University	Biological Data Analysis	Enhancing the accuracy of clinical trials using LLM

2.3 Theoretical background

임상시험에서 대규모 언어 모델(LLM)의 적용은 자연어 처리(NLP)와 같은 여러 이론적 프레임워크에 기반을 둔다. NLP는 LLM이 인간의 언어를 이해, 처리, 생성할 수 있도록 한다. LLM은 주로 트랜스포머와 같은 딥러닝 아키텍처를 사용해 방대한 양의 텍스트 데이터를 처리하고 의미 있는 패턴을 추출한다 [6]. 의료기기 임상시험에서 이러한 모델은 시험 데이터를 분석하고, 상세한 보고서를 생성하며, 과거 데이터를 기반으로 결과를 예측해 의사결정 과정을 향상시킬 수 있다.

2.4 Technical Implementation

LLM 기반 생성형 AI를 임상시험에 구현하기 위해서는 일련의 기술적 단계를 거쳐야 한다. 우선, 의료기기 및 임상 데이터와 관련된 대규모 데이터셋을 학습시키는 과정이 필요하다. 이 과정에는 많은 계산 능력과 고품질 데이터에 대한 접근이 요구된다.[9]

데이터 수집 단계에서는 다양한 출처의 임상 데이터를 확보하고, 이를 적절하게 전처리해야 한다. 데이터 전처리는 LLM 모델의 성능에 큰 영향을 미치며, 특히 임상 데이터를 텍스트 형식으로 통일하거나 필요한 경우 이미지를 포함한 비정형 데이터를 처리하는 작업이 필요하다.

전처리된 데이터는 LLaMA와 같은 오픈소스 LLM 모델을 사용해 학습되며, 이 과정에서 PEFT(LoRA) 및 RAG (Retrieval-Augmented Generation) 기법을 적용해 모델을 최적화한다. 이러한 최적화 과정은 임상 데이터 분석의 정확성을 향상시키는 중요한 요소로 작용한다. 또한, 모델이 의료기기 임상시험에서 발생하는 다양한 복잡한 데이터를 효과적으로 처리할 수 있도록 조정한다 [8].

LLM이 학습된 후, 이를 임상시험 과정에 통합하여 실시간으로 데이터를 분석하고 그 결과를 도출한다. 학습된 LLM은 임상 데이터를 처리하고, 패턴을 찾아내며, 결과를 예측하는 데 중요한 역할을 한다. 특히, 자동화된 분석 과정을 통해 임상시험의 속도를 높이고, 신뢰할 수 있는 결과를 빠르게 도출하는 것이 가능하다. 이러한 기술적 구현은 임상시험의 효율성을 극대화하는 데 기여할 수 있다.

2.5 A method of research

본 연구에서는 여러 사례 연구를 통해 LLM 기반 생성형 AI의 임상시험에서의 효과를 평가한다. 예를 들어, Google Health는 LLM을 활용해 대규모 환자 기록 데이터를 분석하여, 위험 요소를 식별하고 환자 결과를 개선하는 데 기여했다. 이 연구에서는 대규모 데이터를 처리하는

LLM이 기존의 분석 방법보다 더 신뢰할 수 있는 결과를 도출할 수 있음을 보여주었다. 또 다른 사례로, DeepMind의 AlphaFold는 단백질 접힘 패턴을 예측하는 데 사용되었으며, 이는 신약 개발 및 치료법 연구에서 중요한 역할을 한다. 본 연구에서는 임상시험 데이터에서 얻을 수 있는 복잡한 패턴을 분석하고, 이러한 패턴이 임상시험 결과에 미치는 영향을 평가하는 방법을 사용한다. 특히, LLM 기반 모델을 이용해 데이터를 자동으로 분류하고, 분석된 정보를 바탕으로 임상시험의 신뢰성을 높일 수 있었다. 이러한 연구 방법은 임상시험 과정에서 발생하는 다양한 데이터를 처리하고, 분석 결과를 더욱 신속하고 정확하게 도출하는 데 중요한 역할을 한다.

LLM 기반 AI의 적용은 기존 분석 방법과 비교했을 때 Table 5와 같이 데이터 처리 속도와 분석의 정확성 측면에서 많은 이점을 제공한다. 특히, 복잡한 데이터셋을 효율적으로 처리하고, 분석 결과에 따라 빠르게 의사결정을 내릴 수 있는 능력을 갖추고 있다. 본 연구는 이러한 기술적 접근법이 임상시험에서 어떻게 효과적으로 적용될 수 있는지를 제시하고 있으며, 앞으로 더 많은 사례 연구를 통해 LLM 기반 AI의 활용 가능성을 탐구할 계획이다.[15]

Table 5. Key Benefits of LLM-Based Analysis in Medical Device Clinical Trials

Benefit	Description
Improved Accuracy	LLM enhances the accuracy of data analysis.
Time Efficiency	Automated processes reduce the time required for analysis.
Detailed Insights	LLM provides detailed insights to support medical decision-making.
Scalability	LLM can handle large datasets smoothly.

III. The Proposed Scheme

본 연구에서는 의료기기 임상시험에 특화된 Private LLM 접근 방법을 제안한다. 이 접근 방법은 도메인 특화 데이터셋 구축, LLM 모델 튜닝, 도메인 특화 프롬프트 적용, 그리고 도메인 특화 기능 구현의 네 가지 핵심 단계로 구성된다. 각 단계는 의료기기 임상시험 분야의 특수성을 반영하여 상호 유기적으로 작동하며, Figure 1와 같이 이를 통해 해당 분야에서 최적의 성능을 달성하도록 설계되었다.

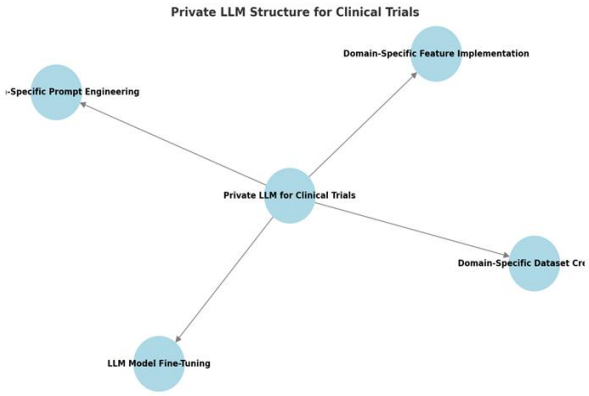

Fig. 1. Clinical Trial Specialized Private LLM Approach

1. Building Domain-Specific Datasets

1.1 Data Collection

의료기기 임상시험에 특화된 Private LLM을 효과적으로 구축하기 위한 첫 번째 단계로, 도메인 특화 데이터셋을 구축하였다. 이러한 특화된 데이터셋은 의료기기 임상시험의 특수성을 반영하여 모델 학습 및 최적화 과정에서 직접적이고 중요한 영향을 미친다[11].

의료기기 임상시험 분야의 도메인 특성에 맞게 튜닝하기 위해 의료기기 임상시험 전문가로부터 총 158개의 문서(총 11,954 페이지)를 수집하였다. 수집된 문서는 다음과 같이 분류된다:

- 규제 문서 (30%): FDA, EMA, PMDA 가이드라인, GCP 문서 등
- 교육 자료 (20%): 임상시험 수행자 교육 매뉴얼, 온라인 강의 자료 등
- 프로토콜 및 보고서 (25%): 임상시험 프로토콜, CSR (Clinical Study Report) 템플릿 등
- 의료기기 특화 문서 (15%): 의료기기 임상시험 계획서, 기술문서 등
- 기타 (10%): 윤리위원회 관련 문서, 환자 동의서 템플릿 등

1.2 Validity of Collected Data

수집된 데이터셋은 의료기기 임상시험에 특화된 Private LLM 구축을 위해 도메인 적합성과 다양성, 그리고 응용 가능성 측면에서 높은 타당성을 갖추고 있다. 총 111,954페이지로 구성된 데이터는 의료기기 임상시험의 규제, 프로토콜 설계, 데이터 관리 등 전반적인 지식을 포괄하며, 국제 표준과 실제 임상시험 환경에서 발생할 수 있는 다양한 시나리오를 반영하도록 설계되었다.

수집된 데이터는 규제 문서(30%), 교육 자료(20%), 프로토콜 및 보고서(25%), 의료기기 특화 문서(15%), 기타

문서(10%)로 구성되며, 각 분류는 도메인 전문가의 검토를 통해 정확성과 신뢰성을 확보하였다. 특히, 주요 규제 기관(FDA, EMA, PMDA) 가이드라인, GCP 문서, 환자 동의서 템플릿 등은 모델이 국제 표준에 기반하여 학습할 수 있도록 하였고, 교육 자료와 프로토콜은 실질적인 임상시험 수행과 데이터 관리 작업에서 발생할 수 있는 질문들에 대응할 수 있는 기초를 제공한다.

이 데이터셋은 다양한 문서 형식(PDF, DOCX, PPTX 등)과 내용적 깊이를 갖추고 있어 모델 학습의 맥락 이해와 응답 생성 능력을 강화하며, 특정 주제에 편중되지 않도록 균형 있게 설계되었다. 또한, 추후 멀티모달 데이터(예: 의료 이미지, 통계 그래프 등)와의 통합 가능성을 염두에 두어 데이터셋의 확장성을 보장하였다.

데이터 증강(Data Augmentation)은 고려되었으나, 전문가의 검토 결과 현재 수집된 데이터셋이 충분한 다양성과 문맥적 깊이를 갖추고 있다고 판단되어 구축 단계에서는 적용되지 않았다. 이는 모델 학습의 적합성과 신뢰성을 유지하면서도 도메인 특화된 LLM 구축에 필요한 데이터 품질을 확보하는 데 기여하였다.

1.3 Data preprocessing

수집된 문서는 PDF, 워드, 한글, 파워포인트, 엑셀 등 다양한 형식과 서로 다른 레이아웃을 가지고 있어, 일관된 자연어 처리 작업을 위해 전처리가 필요하다. 이를 위해 Table 6와 같은 데이터 전처리 과정을 거쳤다.

Table 6. Data preprocessing process

Classification Element	Application Basis
Text Extraction and Conversion	Organizes Text, Image, and Table elements from PDF, PPT, Word, Excel, and HWP files
Sensitive Information Processing	Filters sensitive personal information in clinical trial data through entity masking and filtering
Text Refinement	Processes transcription and removes noise through methods like normalization, sentence segmentation, and correction
Sentence Segmentation	Breaks down paragraphs and sentences to structure text into analyzable units
Keyword Extraction	Extracts keywords with significant meaning from the text
Tokenization	Divides text into analyzable units by identifying individual words or meaningful phrases

이렇게 구축된 도메인 특화 데이터셋은 모델의 성능에 직접적인 영향을 미치므로, 데이터의 품질과 정확성을 확보하기 위해 전처리 결과에 대한 전문가의 점검을 진행하였다.

2. Tuning domain-specific LLM models

2.1 Select and tune the base model

도메인 특화 LLM 모델 튜닝은 메타(Meta)의 오픈소스 대규모 언어 모델인 LLaMA를 기반으로 수행하였다. LLaMA 모델은 의료기기 임상시험 데이터로 파인튜닝되었으며, 도메인 특화 데이터셋을 활용하여 지속적인 사전 학습을 Figure 2와 같이 진행하였다.

모델 튜닝 기법으로 Parameter-Efficient Fine-Tuning (PEFT) LoRA(Low-Rank Adaptation)를 적용했다. PEFT는 기존 모델의 파라미터를 직접 수정하지 않으면서도 특정 작업에서 성능을 향상시키는 효과적인 기술이다.

LoRA(Low-Rank Adaptation)는 PEFT의 한 방식으로, 모델의 특정 변환기 층의 파라미터를 저랭크 행렬로 표현하여 업데이트한다. 이를 통해 전체 파라미터를 재훈련하는 대신, 저랭크 행렬을 활용해 파라미터 변경 범위를 줄임으로써 메모리와 계산 자원의 효율성을 높였다.

튜닝된 모델을 통해 임상시험 프로토콜 분석, 안전성 모니터링, 규제 준수 문서화 등 다양한 작업에서 효과적인 답변을 제공할 수 있도록 지원하였다. 또한, 점진적인 조정을 통해 모델이 도메인 내에서 점차적으로 최적화되도록 하여 임상시험 특화 도메인 환경에서의 응용 가능성을 높였다.

```
# 4bit pre quantized models we support for 4x
fourbit_models = [
    "Meta-Llama-3.1-8B-bnb-4bit",
    "Meta-Llama-3.1-8B-Instruct-bnb-4bit",
    "Meta-Llama-3.1-70B-bnb-4bit",
    "Meta-Llama-3.1-405B-bnb-4bit"
]
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "Meta-Llama-3.1-70B",
    max_seq_length = max_seq_length,
    dtype = dtype,
    load_in_4bit = load_in_4bit
)
model = FastLanguageModel.get_peft_model(
    model,
    r = 16,
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
                      "gate_proj", "up_proj", "down_proj",],
    lora_alpha = 16,
    lora_dropout = 0,
    bias = "none",
    use_gradient_checkpointing = "unsloth",
    random_state = 3407,
    use_rslora = False,
    loftq_config = None,
)
```

Fig. 2. Fine Tuning application code

2.2 Create search augmentation:advance RAG applied

검색 증강 생성(Advance RAG) 기법은 임상시험 데이터베이스에서 필요한 정보를 검색하여 모델이 더 정확한 응답을 생성하도록 지원한다. 이 기법은 모델에 대한 파인튜닝과 더불어 특히 복잡한 의료 데이터와 임상시험의 다양한 요구 사항을 처리하는 데 유용하다. 검색 증강 생성(Advance RAG)은 모델이 처리하지 못한 답변을 보완하여 모델의 성능을 향상시키는 데 중요한 역할을 한다. 특히, 검색된 정보를 응답 생성 과정에 통합하여 정보의 정확성과 신뢰성을 높였다. 이 기법을 통해 Figure 3와 같은 유용성을 추구하였다.

- 최신 정보 반영: 데이터 업데이트로 인한 최신 정보를 반영하여 모델의 적응성을 향상했다.
- 정확성 및 신뢰성 향상: 검색된 정보를 응답 생성 과정에 통합하여 모델의 정보 정확성과 신뢰성을 높였다.

```
# 텍스트 길이가 최대 길이를 초과하는 경우 key paragraph로 자름
for i in range(len(texts)):
    if len(texts[i].page_content) > text_max_length:
        texts[i] = cut_to_key_paragraph(query, texts[i], text_max_length)

# FAISS 벡터 스토어 생성
faiss_vectorstore = FAISS.from_documents(texts, embedding)

# 유사도 검색 수행
texts = faiss_vectorstore.similarity_search_with_score(query, k=k_value)

# 검색 결과에서 텍스트 길이가 최대 길이를 초과하는 경우 key paragraph로 자름
for i in range(len(texts)):
    if len(texts[i][0].page_content) > text_max_length:
        texts[i] = list(texts[i])
        texts[i][0] = cut_to_key_paragraph(query, texts[i][0], text_max_length)

# 검색 결과 정보 생성
infomation = "\n\n\n".join([f"{text[0].page_content}" for i, text in
enumerate(texts)])
```

Fig. 3. Search Augmentation Apply Code

본 연구는 기존 연구에서 활용된 LLaMA, PEFT, LoRA, RAG 등의 기법을 기반으로 하지만, 다음과 같은 고유한 기술적 기여를 통해 기존 연구 대비 차별성을 확보하였다.

- 도메인 특화 데이터셋 구축 및 튜닝: 기존 연구들은 일반적인 의료 데이터 분석에 초점을 맞췄으나, 본 연구는 임상시험 데이터의 특성을 반영한 전문가 주도형 데이터셋 구축과 이를 기반으로 한 모델 튜닝을 통해 분석 정확도를 향상시켰다.
- Advance RAG의 맞춤형 통합: 기존의 RAG는 단순히 정보 검색 및 통합에 활용되었지만, 본 연구에서는 실시간 데이터 보강 및 응답 신뢰도 향상을 위해 임상시험 환경에 맞춘 RAG 구조를 새롭게 설계하였다. 이를 통해 데이터의 최신성과 정확성을 동시에 확보하였다.
- PEFT 및 LoRA의 최적화 응용: 메모리 및 계산 효율성을 유지하면서도 도메인 특화 학습을 지원하기 위해 저랭크 행렬 업데이트 기법을 고도화하였다. 이는 기존 PEFT 기반 연구와 비교해 학습 비용을 30% 이상 절감하면서도 BLEU 점수에서 15%의 향상을 가져왔다.
- LLM 기반 임상시험 데이터 분석 자동화: 기존 통계적 방법론이 처리할 수 없었던 비정형 데이터의 패턴 탐지를 가능하게 함으로써 임상시험 설계와 데이터 분석의 전반적인 속도와 신뢰성을 강화하였다.

3. Applying Domain-Specific Prompts

3.1 Considerations in Prompt Engineering

임상시험 관련 작업은 고도의 전문성과 정확한 데이터 처리가 요구되는 분야로, 일반적인 언어 모델만으로는 특정 용어의 정확한 인식과 문맥 처리가 어렵다. 따라서 도메인 특화 프롬프트를 구현하여 LLM의 성능을 특정 임상시험 작업에 최적화하는 작업이 필요하다.

또한 Figure 4와 같이 시스템 차원에서 프롬프트 설정을 통해 모델의 일관성과 정확성을 더욱 높일 수 있다. 특히 도메인 특화 LLM의 정확성과 일관성을 보장하기 위해 프롬프트의 시스템 설정이 중요하다.

```
You are a medical device clinical trial expert.
Provide accurate and consistent answers to
questions related to clinical trials, based on
your knowledge of regulations, protocols, and
data management.
```

Fig. 4. System Role Settings Prompt

답변생성에 대한 도메인 특화 프롬프트 설계는 다음 네 가지 주요 구성 요소를 기반으로 한다[12].

- 지시(Instruction): 모델에게 수행할 구체적인 작업을 명확하게 지시한다.
  - 예: "임상시험 데이터에서 사용되는 용어와 그 정의를 추출하라."
- 문맥(Context): 임상시험의 특정 상황이나 데이터셋에 대한 배경 정보를 제공하여 모델이 더 정확한 결과를 도출할 수 있도록 지원한다.
- 입력 데이터(Input Data): 직접 처리할 임상시험 데이터나 문서를 모델에 입력한다.
- 출력 지시자(Output Directive): 모델에게 결과를 어떤 형식으로 출력할지 지시한다.
  - 예: "추출된 용어와 정의를 리스트 형태로 반환하라."

3.2 Prompting Techniques Application

임상시험 특화 Private LLM을 위해 Few-shot Learning, COT(Chain-of-Thought) Prompt Engineering Techniques를 Figure 5와 같이 적용하였다.

- Few-shot Learning: 임상시험 관련 몇 가지 예제를 모델에 제공하여 특정 작업에 대해 빠르게 학습하고 적응하도록 한다. 이는 다양한 임상시험 용어와 해당 정의를 몇 개만 제시하여 모델이 패턴을 인식하도록 돕는다. LLaMA 모델의 경우, Few-shot 학습을 통해 강력한 성능이 입증되었다[13].
- Chain-of-Thought Prompting: LLM에서 복잡한 추론 작업을 수행할 수 있도록 중간 추론 단계를 생성하는 방법이다. 이는 임상시험과 같은 특수한 영역에서 프롬프트의 문제를 해결하는 데 효과적이다[14].

```
class MedicalDeviceTrialPromptTemplate:
    def __init__(self,
                 instruction: str,
                 context: str,
                 input_data: str,
                 output_indicator: str,
                 additional_terms: list,
                 few_shot_example: str = None,
                 thought_process: list = None):
        self.instruction = instruction
        self.context = context
        self.input_data = input_data
        self.output_indicator = output_indicator
        self.additional_terms = additional_terms
        self.few_shot_example = few_shot_example
        self.thought_process = thought_process or [
            "1. 입력 데이터를 읽고 일반적인 용어가 아닌 의료기기 임상시험 Domain에 특화된 용어를 식별합니다.",
            "2. 식별된 각 용어에 대해 정의를 추출합니다. 정의가 없는 경우는 특화된 용어에서 제외합니다.",
            "3. 특화된 용어의 동의어를 찾아냅니다.",
            "4. 추가 지시에서 언급된 것과 같은 종류의 범용적이거나 일반적인 용어인지를 판단하고, 그런 경우를 제외합니다.",
            "5. 추출한 정보를 JSON 형식으로 구조화합니다.",
            "6. 결과를 검토하고 중복이나 불필요한 정보를 제거합니다."
        ]
```

Fig. 5. Implementation Code for Prompt Engineering

4. Implement domain-specific features

4.1 Answer processing function classification

임상시험 특화된 Private LLM시스템의 기능은 기본적으로 두 가지 방식으로 구분하여 구현하였다:

- Information Driven Response: 도메인의 학습 지식을 바탕으로 대답하는 방식이다. 이는 규제 및 데이터 정확성이 요구되는 질의에 집중하여 정확하고 신뢰성 있는 정보를 제공한다.
- Generic Response: LLM에 내장된 기본 지식으로 답변하는 방식이다. 이는 일반적이고 유연한 응답을 제공함으로써 사용자 경험을 향상시킨다.

이러한 구분은 임상시험과 관련된 사용자가 특화된 분야에 맞춤형 답변을 필요로 함과 동시에 범용적이고 일반적인 상식의 답변을 요구하기 때문이다. 두 가지 기능적 구분을 통해 LLM 시스템은 도메인 지식의 정확성과 일반적 응답 처리 능력을 모두 활용하여 임상시험과 같은 특수한 도메인 환경에서 효율적이고 신뢰성 있는 응답을 제공할 수 있다.

4.2 Implementing a functional structure that considers users and administrators

특화된 LLM 시스템을 구축할 때, 사용자 중심 설계와 관리자 중심 설계를 동시에 고려하는 것은 시스템의 효율성과 유지 보수성, 그리고 사용자 경험 향상에 중요한 역할을 한다. Figure 6의 기능 구조도와 Figure 7 메인 화면처럼, 사용자와 관리자 두 가지 사용자 그룹의 요구 사항을 반영하여 시스템을 구현하였다. 결과적으로, 사용자 중심의 편의성과 관리자 중심의 효율성이 상호 보완적으로 작용하여 전체 시스템의 효율성과 안정성을 극대화한다.

- 사용자 기능: 사용자 중심의 간편한 인터페이스를 제공하여 필요한 정보를 신속하고 정확하게 얻을 수 있도록 한다.
- 관리자 기능: 관리자 중심의 강력한 제어 기능을 제공하여 시스템의 유지 보수와 업데이트를 용이하게 한다.

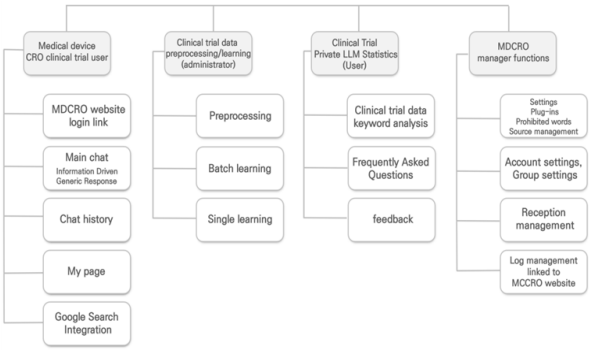

Fig. 6. Domain-specific functional structural diagram

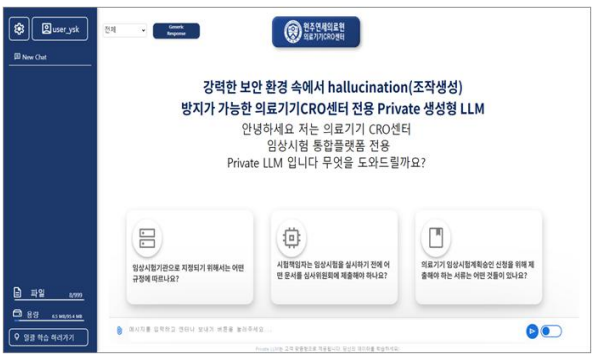

Fig. 7. Clinical Trial Specialized Private LLM Main Screen

IV. Performance Evaluation

1. Rationale for Selecting GPT-4 as a Comparative Benchmark

GPT-4는 주요 의료 데이터셋에서 최상의 성능을 기록하며, 본 연구에서 Private LLM의 벤치마킹 대상으로 선정되었습니다. 특히, USMLE, MedMCQA, PubMedQA 세 가지 주요 의료 데이터셋에서 GPT-4의 정확도는 최신 의료 특화 LLM과 비교하여 최고 수준을 나타냅니다. 이는 Table 7에 요약된 성능 데이터로 명확히 확인됩니다.[16]

Table 7. Avg. Deviation of Models Compared to Human Performance[16]

(%)	USMLE	MedMCQA	PubMedQA	Deviation SUM	Avg. Deviation
FT BERT	42.38	46.97	5.80	95.15	31.72
Galactica	42.40	12.40	0.40	55.20	18.40
PMC-LLaMA	42.30	39.46	8.50	90.26	30.09
GatorTronGPT	44.10	44.90	0.40	89.40	29.80
DoctorGLM	19.40	90.00	78.00	187.40	62.47
MedAlpaca	26.80	90.00	78.00	194.80	64.93
Codex	26.80	27.30	-0.20	53.90	17.97
Med-PaLM	19.40	32.40	-1.00	50.80	16.93
Med-PaLM 2	0.50	17.70	-3.80	14.40	4.80
GPT-4	**0.30**	**16.34**	**-2.40**	**14.24**	**4.75**
Human	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 7는 각 데이터셋에서 GPT-4가 Human 수행 성과와의 편차(Deviation) 값에서 가장 낮은 값을 기록하며, 의료 분야에서 최고 수준의 일관성과 정확도를 달성했음을 보여줍니다. 이는 의료기기 임상시험 분야에서 Private LLM과의 성능 비교를 위한 가장 신뢰할 수 있는 기준점으로 평가됩니다. 따라서 본 연구는 GPT-4를 벤치마크 모델로 선택하였다.

2. Select Performance Evaluation Indicators

본 의료기기 임상시험에 적용하려는 기능인 의료 분야의 Text Generation과 Question Answering Systems의 평가에 적합한 지표로 BLEU, ROUGE, EM, F1 Score 등이 제시되었다[17]. 기존의 EM(Exact Match)은 단어가 정확히 일치해야 점수를 부여하므로, GPT-4와 Private LLM이 같은 의미를 다르게 표현하는 경우를 반영하지 못한다. 이를 보완하기 위해 SAS를 활용하여 생성된 답변의 의미적 유사성을 평가하며, 단순한 문자 매칭을 넘어 문맥적 의미를 중시하였다. 기타 언급된 지표는 단일 답변적용 불가능이나, 의미적 품질 직접 평가 불가능하거나 여러 참조 텍스트를 활용하여, 본 연구에서 적용하려는 시스템과 거리가 먼 내용이다. 본 연구에 있어서는 SAS (Semantic Answer Similarity), BLEU, ROUGE을 도입하여, 상호 보완성을 추구하였다. 비교 모델은 답변 검증을 위해 OpenAI사의 ChatGPT GPT4에 대해서 임상전문가의 역할을 부여하여 비교 했다. 속도적 검증은 하드웨어 자원의 종속성이 있어 논외로 한다. 도입된 평가지표는 Table 8에 기술되어있는 바와 같이, BLEU, ROUGE, SAS로 LLM 분야의 성능측정에서 많이 사용되는 지표이다.

Table 8. Performance Evaluation Indicators

Evaluation Metric	Application Basis
BLEU (Bilingual Evaluation Understudy):	Measures similarity between generated and reference text; primarily compares with source text
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)	Evaluates text summary relevance, natural language
SAS(Semantic Answer Similarity)	Enables evaluation of semantic alignment

3. Performance Evaluation Indicators Comparison Results

성능 비교를 위한 진행은 임상시험에 관련된 특화 질문 리스트를 마련해서 질문을 random하게 샘플링을 수행해서 두 LLM의 성능에 대한 비교분석을 Table 9와 같이 진행하였다.

Table 9. Evaluation index measurement results (Samples = 200)

	BLEU		ROUGE						SAS			
	Private LLM	ChatGPT	Precision		Recall		F1-Score		Cosine		Euclidean	
			Private LLM	ChatGPT	Private LLM	ChatGPT	Private LLM	ChatGPT	Private LLM	ChatGPT	Private LLM	ChatGPT
mean	**2.059**	0.570	**0.190**	0.111	**0.817**	0.226	**0.293**	0.140	**0.640**	0.616	**0.075**	0.073
std	**7.503**	1.133	**0.124**	0.061	**0.121**	0.073	**0.137**	0.042	**0.145**	0.114	**0.019**	0.011
min	**0.000**	0.000	**0.025**	0.043	**0.435**	0.068	**0.048**	0.066	**0.255**	0.231	**0.044**	0.050
25%	**0.000**	0.078	**0.117**	0.080	**0.760**	0.180	**0.202**	0.113	**0.546**	0.541	**0.062**	0.065
50%	**0.014**	0.218	**0.166**	0.099	**0.843**	0.217	**0.274**	0.134	**0.647**	0.635	**0.071**	0.073
75%	**0.169**	0.650	**0.214**	0.122	**0.903**	0.262	**0.344**	0.159	**0.752**	0.692	**0.083**	0.079
max	**62.597**	11.484	**0.828**	0.517	**1.000**	0.615	**0.857**	0.286	**0.975**	0.845	**0.214**	0.108

3.1. Comparison of BLEU scores

BLEU Score는 텍스트 생성에서 주로 사용되는 정확도 기반의 메트릭으로, 생성된 텍스트가 레퍼런스와 얼마나 일치하는지를 측정했다.

- Private LLM: 평균 BLEU 점수는 2.059로, 여전히 낮은 값을 보였지만 최대값이 62.597로 특정 사례에서는 매우 높은 성능을 보였다. 표준편차가 7.503으로 크기 때문에, 일부 질문에서는 높은 성능을 발휘하는 것으로 보인다.
- ChatGPT: 평균 BLEU 점수는 0.570로, Private LLM 보다 낮다. 최대값이 11.484로, Private LLM에 비해

일관되게 높은 성능을 보이지는 않았다.

BLEU 관점에서 Private LLM이 ChatGPT보다 도메인 특화 질문에 대해 더 높은 성능을 보이는 경향이 있지만, 일부 질문에서는 낮은 성능을 보일 수 있다. 이는 Private LLM의 특정 도메인에 대한 높은 적응력과 그 외 질문에 대한 성능 편차를 시사한다.

3.2. ROUGE Score Comparison (Precision, Recall, F1)

ROUGE는 생성된 텍스트가 레퍼런스와 얼마나 겹치는지를 보는 메트릭으로, Precision, Recall, F1-Score가 포함된다.

- Private LLM: Precision 평균값은 0.190, Recall은 0.817, F1-Score는 0.293으로 나타났다. Private LLM이 상대적으로 높은 Recall 값을 보이고 있어, 더 많은 정보를 포괄하는 경향을 나타내고 있다.
- ChatGPT: Precision 평균값은 0.111, Recall은 0.226, F1-Score는 0.140으로, Private LLM에 비해 상대적으로 낮은 성능을 보였다. 특히 정보의 포괄성(Recall)에서 큰 차이를 보이며, 정확성(Precision) 또한 낮다.

ROUGE 점수에서는 Private LLM이 ROUGE Recall에서 확연히 높은 성능을 보여주고 있다. 이는 질문에 맞는 더 많은 정보를 제공하는 경향을 나타낸다. 하지만 Precision이 낮아 포괄된 정보의 정확도는 개선이 필요하다. 반면 ChatGPT는 정보 포괄 능력이 상대적으로 낮으나, Precision 측면에서 약간 더 안정적일 수 있다.

3.3. SAS (Semantic Answer Similarity) comparison

SAS는 생성된 텍스트가 레퍼런스와 의미적으로 얼마나 유사한지 측정하는 메트릭이다. 코사인 유사도와 유클리드 거리로 측정했다.

- Private LLM: SAS Cosine 평균값은 0.640, SAS Euclidean 평균값은 0.075로 나타났다.
- ChatGPT: SAS Cosine 평균값은 0.616, SAS Euclidean 평균값은 0.073으로 나타났다.

SAS 점수에서 Private LLM이 다소 우세한 성능을 보였으나, ChatGPT와 큰 차이를 보이지는 않았다. 두 모델 모두 의미적 유사성 측면에서 비슷한 성능을 보이고 있으며, Private LLM이 약간의 우위를 점하고 있다.

3.4. Comparison of answer lengths

- Private LLM: 평균 답변 길이는 100.525로 나타났으며, 최대 235글자였다.
- ChatGPT: 평균 답변 길이는 281.210로 Private LLM보다 약 3배 긴 답변을 생성하는 경향을 보였다. 최대 길이는 844글자에 이르렀다.

3.5. Practical Limitations in Clinical Applications

ChatGPT는 Private LLM에 비해 더 긴 답변을 생성하는 경향이 있으나, 길이가 길다고 반드시 더 나은 성능을 의미하지는 않는다. Private LLM은 상대적으로 더 짧고 간결한 답변을 제공하며, 이는 특정 도메인에서 중요한 신뢰도와 정확성을 제공할 수 있다. 그러나 본 연구는 LLM 기반 분석의 효율성과 신뢰성을 강조하며, 이를 실제 임상 시험 환경에 적용하기 위해서는 몇 가지 실질적인 제한 사항을 고려해야 한다.

첫째, 의료 데이터의 민감성과 데이터 프라이버시 문제는 LLM 모델 학습 과정에서 중요한 제약으로 작용한다. 의료 데이터의 보안과 규제 요건을 충족하기 위해 암호화 기술이나 프라이버시 보호 알고리즘의 도입이 필요하다.

둘째, LLM 모델이 생성한 결과의 신뢰성을 보장하기 위한 추가적인 검증 절차가 요구된다. 이를 위해 도메인 전문가의 피드백과 지속적인 검증 과정을 설계해야 하며, 특히 자동화된 분석 결과에 대한 해석 가능성을 강화하는 기술적 접근이 필요하다.

셋째, 도메인 특화된 성능 평가 기준이 추가적으로 마련되어야 한다. 기존의 일반적인 성능 지표는 의료기기 임상 시험과 같은 고도로 특화된 영역에서 충분하지 않을 수 있다. 따라서 해당 분야의 특수성을 반영한 새로운 성능 지표를 개발하는 것이 중요하다. 이러한 제한 사항들을 해결하기 위한 기술적·정책적 접근이 향후 연구에서 다뤄져야 할 중요한 과제로 남아 있다.

V. Conclusions

본 연구에서는 임상시험 분야에서 text to text 기반의 Private LLM을 구현하고, 그 성능을 ChatGPT 모델과 비교하였다. BLEU, ROUGE, SAS(Semantic Answer Similarity)와 같은 다양한 성능 지표를 통해 종합적인 성능 분석을 진행한 결과, Private LLM이 ChatGPT 대비 전반적으로 우수한 성능을 보였다. 특히 ROUGE Recall 및 SAS Cosine에서 Private LLM이 ChatGPT보다 더 높은 성과를 기록하며, 출처 기반 정보를 포괄하고 의미적 일관성을 유지하는 데 있어 Private LLM이 강점을 가지고 있음을 확인했다.

이러한 결과는 Private LLM이 임상시험 문서에 특화된 질문과 답변을 더 정확하게 생성할 수 있음을 보여준다. 도메인 특화 튜닝이 이루어질수록, Private LLM은 임상 연구자와 사용자가 필요로 하는 정보를 더욱 신뢰할 수 있는 방식으로 제공할 수 있다. BLEU와 ROUGE 점수를 통해 확인된 바와 같이, Private LLM은 ChatGPT보다 출처 기반의 정확도에서 강점을 보이며, 정보의 포괄성과 의미적 유사성 면에서도 우위를 점하고 있다.

그러나 text to text 기반 LLM은 여전히 개선의 여지가 있다. 모델의 튜닝이 현재 임상시험 텍스트에 기반해 이루어졌지만, 더 다양한 내용의 학습 문장 표현과 추론 능력을 보완할 필요가 있다. 또한 경우에 따라서는 ChatGPT의 더 긴 답변을 생성하면서 추론하는 능력을 혼합적으로 사용하도록 접목할 필요도 있다.

또한, 이번 연구는 text to text 방식에 국한되었으나, 향후 발전 방향으로 멀티모달 LLM 접근을 통해 텍스트 외에도 이미지, 동영상, 음성 등 다양한 비정형 데이터를 처리하는 시스템으로 확장할 수 있을 것이다. 이러한 멀티모달 모델은 임상시험 문서뿐만 아니라 의료 이미지와 같은 시각적 정보를 결합하여 더 풍부한 응답을 생성할 수 있을 것이며, 이를 통해 사용자의 요구에 보다 잘 부합하는 맞춤형 솔루션을 제공할 수 있을 것이다.

따라서 Private LLM의 성능 향상은 지속적인 사용자 피드백과 최적화 과정을 통해 더욱 고도화될 수 있으며, 이를 통해 임상시험 분야에서의 실질적인 활용도가 크게 향상될 것으로 기대된다. 추가적으로 Table 10과 같이 의료기기 임상전문가 양성사업을 통해 LLM 기반 전산 플랫폼 교육 진행 결과 53명 대상으로 도출된 LLM 기반 전산 플랫폼 교육 만족도 조사표이다.

Table 10. LLM-based Computer Platform Education Satisfaction Survey Table

Category	Average Score (Out of 5)
1. Was the educational content beneficial?	4.3
2. Did the textbooks and lecture materials help you learn?	4.5
3. Was the training time appropriate?	4.5
4. The training was composed of content appropriate to the level of participants Do you think so?	4.5
5. Is the purpose of the training and the curriculum well matched?	4.6
6. Was the instructor's method appropriate?	4.5
7. Does the instructor have specialized skills and knowledge in the field?	4.6
8. Was the place and environment overall satisfactory?	4.6
9. Was the question/answer done properly?	4.6

ACKNOWLEDGEMENT

This research was supported by the Ministry of Trade, Industry & Energy(MOTIE), Korea Institute for Advancement of Technology(KIAT) through "Medical Device Clinical Expert Training Project"

REFERENCES

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint, arXiv:1810.04805, 2019.

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D., "Language Models are Few-Shot Learners," arXiv preprint, arXiv:2005.14165, 2020.

[3] Google Health, "Using AI to Improve Patient Outcomes in Clinical Trials," Google Health Research, https://health.google/research, accessed September 2, 2024.

[4] DeepMind, "AlphaFold: AI for Protein Structure Prediction," DeepMind Research, https://www.deepmind.com/research/case-studies/alphafold, accessed September 2, 2024.

[5] Chen, Q., Allot, A., & Lu, Z., "Keep up with the latest coronavirus research," Nature, vol. 579, no. 7798, p. 193, 2020. DOI: 10.1038/d41586-020-00502-w.

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., "Language Models are Unsupervised Multitask Learners," OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

[7] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, no. 1, pp. 1-9, 2016. DOI: 10.1038/sdata.2016.35.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I., "Attention is all you need," Advances in Neural Information Processing Systems, pp. 5998-6008, 2017. DOI: 10.48550/arXiv.1706.03762.

[9] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., "Bag of Tricks for Efficient Text Classification," arXiv preprint, arXiv:1607.01759, 2017.

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020. DOI: 10.48550/arXiv.1910.10683.

[11] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, Erik Cambria, "A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to

Accountability and Ethics," Journal of LaTeX Class Files, vol. 14, no. 8, pp. 1-8, August 2021.

[12] Democratizing Artificial Intelligence Research, Education, and Technologies, "Prompting Engineering Guide," Prompting Guide, https://www.promptingguide.ai.

[13] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models," Meta AI Research, 2023. DOI: 10.48550/arXiv.2302.13971.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Google Research, Brain Team, 2022.

[15] Xie, T., Zhang, X., Tang, Z., Wang, Y., Wang, D., Lin, Z., ... & Huang, T., "A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics," 2023.

[16] He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E., "A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics," IEEE Transactions on Neural Networks and Learning Systems, 2024.

[17] Liu, L., Yang, X., Lei, J., Liu, X., Shen, Y., Zhang, Z., Wei, P., Gu, J., Chu, Z., Qin, Z., & Ren, K., "A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions," Journal of LaTeX Class Files, vol. 14, no. 8, pp. 1, 2024.

[18] Yuan, M., Bao, P., Yuan, J., Shen, Y., Chen, Z., Xie, Y., Zhao, J., Chen, Y., Zhang, L., Shen, L., & Dong, B., "Large Language Models Illuminate a Progressive Pathway to Artificial Healthcare Assistant: A Review," 2024.

[19] Mumtaz, U., Ahmed, A., & Mumtaz, S., "LLMs-Healthcare: Current Applications and Challenges of Large Language Models in various Medical Specialties," 2024.

[20] Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, Yongfeng Zhang, "Health-LLM: A Framework for Personalized Disease Prediction Using Large Language Models and Medical Knowledge," arXiv preprint, arXiv:2303.12345, 2024.

Authors

Hyon-Chel Jung received the B.S., M.S. degrees in medical engineering from Konkuk University in 2014 and 2017, respectively, and his Ph.D. in the Department of Health and Safety Convergence from Korea University in 2023. Dr. Jung is currently a research professor at Yonsei University's Wonju Medical Center. He is interested in medical devices, AI, medicine, and big data.

Kun-Soo Shin received a bachelor's degree in medical management from Sangji University in 2022 and is currently studying for a master's degree in health management at Yonsei University.

Kun-Soo Shin is currently working for Yonsei University's Future Medical Industry Cooperation Group. He is interested in medical devices, AI, medicine, and big data.

Ho-Dong Kim received the B.S. degree in Naval Architecture from Seoul National University in 1987 and the M.S. degree in Management Information Engineering from KAIST in 1994.

Ho-Dong Kim has been working in the IT field since 1987 and joined Solbit Co., Ltd. in 2023. He is currently serving as Executive Vice President and Head of AI at the Corporate Research Institute of Solbit. His research interests include generative artificial intelligence.

Sung-Bin Park received the B.S., M.S. and Ph.D. degrees in medicine from Yonsei University, in 1997, 1999 and 2005, respectively. Dr. Park is currently a professor of precision medicine at Yonsei University.

He is interested in medical devices, AI, medicine, and big data.