LLM을 활용한 질환 예측 방법

이하은[1], 신혜지[1], 한다경[1]
[1]성신여자대학교 컴퓨터공학과 학부생
[1]20211145@sungshin.ac.kr, 20211135@sungshin.ac.kr, 20211153@sungshin.ac.kr

Disease Prediction Method Using LLM

Ha-Eun Lee[1], Hyeji Shin[1], Da-Gyeong Han[1]
[1]Dept. of Computer Engineering, Sungshin Women's University

요 약

본 논문에서는 대규모언어모델(LLM)을 활용한 질환 예측 방법을 연구하고, 프롬프트 엔지니어링을 최적화하여 진단 정확도를 향상시키는 방법을 모색하였다. 사용자 프롬프트에 제공되는 정보 유형(예: 통증 위치, 통증 유발 행동)이 진단 성능에 미치는 영향과 프롬프트 기법이 모델의 일관성 및 신뢰성을 개선하는 효과를 분석하였다. 실험 결과, 특정 정보 유형이 포함될 시 진단 정확도가 높은 것을 밝혀냈으며, 프롬프트 기법을 통해 4.83점(5점 만점)의 진단 정확도와 93.3% 이상의 일관된 결과를 도출할 수 있었다. 이 연구는 LLM의 의료 진단 활용 가능성을 제시하며 향후 연구 방향에 대한 기초를 제공한다.

1. 서론

최근 생성형 인공지능과 대규모언어모델의 성능이 뛰어난 성과를 보임에 따라 마케팅, 법무, 고객서비스, 의료 등 다양한 응용 분야에서 활용되고 있다. 기술이 발달하면서 대규모언어모델을 의료 목적으로 사용하고자 하는 연구와 서비스가 소개되고 있다.

2. 배경

대규모언어모델(LLM, Large Language Model)은 방대한 양의 텍스트 데이터를 학습하여 텍스트를 이해하고 생성하는 언어 모델이다. 대표적인 모델로 GPT, Bard, BERT, PaLM, LLaMA 등이 있다. Med-PaLM, Meditron 등의 의료분야 활용에 적합한 모델도 고안되었으며, LLM의 의료적 성능 테스트를 위해 미국 의사 면허 시험(USMLE)을 풀게 하는 등 의료분야에서 활용하기 위한 시도가 활발히 이루어지고 있다. 특히, 진료 기록을 입력하면 의료 진단을 해주는 DxGPT와 같은 서비스의 등장으로 진료 성능을 판단하기 위한 연구도 주목을 받고 있다. Kanjee Z et al.[1]에서는 의사들도 어려워하는 질병 케이스에 대해서 ChatGPT가 효과적인 감별 진단 성능을 보임을 입증하였다.

본 논문에서는 Kanjee Z et al.의 연구 결과를 기반으로, 의료 진단에서 LLM의 활용도를 높이기 위해 프롬프트 엔지니어링을 최적화하는 방안을 모색한다. AI에게 질문하거나 명령을 내리는 User role과 AI의 행동 방식을 설정하는 System role을 엔지니어링 하였다. User 프롬프트 테스트를 통해 질환 진단을 위해 어떤 정보가 유용한지 제안하며, 다양한 프롬프팅 기법을 적용해 보며 질환 진단에 뛰어난 정확도를 보이는 기법을 소개한다. 해당 과정에서 LLM 중 일반적으로 우수한 성능을 나타낸 [2] OpenAI의 GPT-4 Turbo 모델을 사용하여 실험을 수행하였다.

3. 연구 방법

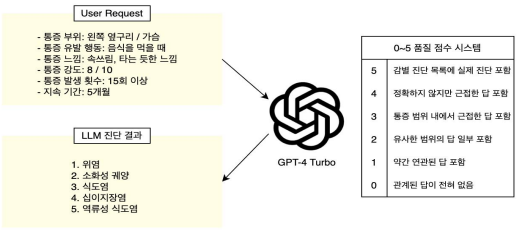
프롬프트 엔지니어링을 시작하기 전에 프롬프트에 들어갈 테스트 데이터를 구축하였다. User 프롬프트에서 사용자가 어떤 정보를 넘겨주는 것이 유용한지 탐구하기 위해 의사와의 인터뷰를 통하여 통증 부위, 통증 유발 행동, 통증 느낌, 통증 강도, 통증 발생 횟수, 지속 시간을 후보로 선정하였다. 각 질환에 대한 후보 정보는 실제 질병 케이스와 전문 병원에서 제공하는 의료 정보를 수집하여 사용하였다.

질환에 따른 진단의 편향성을 줄이기 위하여 실

제 진단명에 해당하는 정답지는 상이한 발병 신체 위치, 증상과 복잡성을 가진 질환으로 구성하였다. 최종 선정된 진단할 질환의 정답지는 위암, 디스크 탈출증, 손목터널증후군으로, 국내 20~50대가 자주 걸리는 대표 질병이다.

LLM이 가능성 높은 진단명을 결과로 반환하도록 프롬프팅 테스트를 진행하고, 모델이 도출한 질환의 정확도와 품질을 판단하기 위해 Kanjee Z et al.의 0~5 품질 점수 시스템을 적용하였다.

(그림 1) 실험 예시



4. 연구 결과

4.1 질환 정보 제공의 유용성

User 프롬프트의 6가지 항목별 데이터의 유무에 따라 총 $63(=2^6-1)$가지 테스트를 세 가지 질환에 맞춰 실시했다. 3명의 평가자가 교차검증을 통해 테스트 결과의 품질을 평가한 결과, 최고 평균 4.67의 결과를 보이며 통증 부위, 통증 유발 행동, 통증 느낌에 대한 정보가 들어있을 때 일반적으로 목표 진단명을 정확히 진단하였다.

<표 1> 사용자 정보에 따른 진단 품질 점수 결과

	통증 부위	통증 유발행동	통증 느낌	통증 강도	통증 발생횟수	지속 기간	정확도(진단 품질 점수) 위암	허리 디스크	손목터널 증후군	총 평균
1	o	o	o	x	x	x	4	5	5	**4.67**
2	o	o	x	x	x	x	4	5	5	**4.67**
3	o	x	o	x	x	x	4	5	5	**4.67**
4	o	x	x	x	x	x	3	4	5	4
5	x	o	o	x	x	x	4	5	5	**4.67**
6	x	o	x	x	x	x	2	4	5	3.67
7	x	x	o	x	x	x	4	4	4	4
…	…						…			
63	o	o	o	o	o	o	4	5	5	**4.67**

4.2 프롬프팅 엔지니어링 기법

의료 분야 질의응답에서 구체화 과정이 응답의 질을 높인다는 연구 결과를 바탕으로[3] 다양한 프롬프팅 기법을 적용하여 모델의 일관성과 정확도를 향상시켰다. 예시를 제공하는 퓨샷 프롬프팅과 명확한 조건 및 형식을 설정하는 조건 기반 프롬프팅을 적용한 결과, 진단 성능이 평균 4.83점으로 향상되었으며, 모델이 93.3% 이상의 일관된 진단을 하였다.

5. 결론

본 논문에서는 LLM을 활용한 질환 예측 방법을 제안하고, 프롬프트 엔지니어링을 통해 질환 진단의 정확도 및 일관성을 높이는 방법을 모색하였다. 실험 결과, 프롬프트에서 제공하는 정보의 구체성이 질환 진단의 성능에 영향을 미친다는 것을 확인하였다. 특히, 통증 부위, 통증 유발 행동, 통증 느낌 정보가 포함된 경우, 질환 진단의 정확도가 유의미하게 향상되었다. 또한, 프롬프트에서 적용한 일부 기법들이 LLM의 진단 결과의 일관성을 높일 수 있었다. 이러한 연구 결과는 LLM을 활용한 의료 진단의 가능성을 제시하며, 프롬프트 엔지니어링을 통해 LLM의 진단 성능을 향상시킬 수 있음을 보여준다.

그러나 본 연구에는 몇 가지 한계점이 존재한다. 첫째, 국내 20~50대에서 자주 발생하는 질병을 예측하도록 평가하였기 때문에 다른 연령대나 지역에 따른 질병의 다양성을 충분히 반영하지 못했다. 둘째, 질환 진단의 정확성을 평가하는 데 사용된 품질 점수 시스템이 다소 주관적일 수 있다. 셋째, 현재 연구는 LLM의 진단 성능에 대한 초기 평가 단계에 머물러 있으며, 실제 의료 환경에서의 적용 가능성에 대한 심층적인 검토가 필요하다.

향후 연구로 다양한 질환군과 연령대, 지역별 데이터로 확장하여 모델의 일반화 성능을 평가할 필요가 있다. 또한, 프롬프트 엔지니어링 기법을 더욱 세분화하여 프롬프트 설계 방안을 모색해야 한다. 이러한 연구는 LLM을 활용한 질병 예측 시스템의 상용화 가능성을 높이는 데 기여할 것이다.

본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330(1):78-80. doi:10.1001/jama.2023.8288

[2] 한동근, 최승빈, 김보성, 배성민, 송성민, 신병철, 박세진*. 대규모 언어 모델(LLM)의 종류별 프로그래밍 성능 비교. 대한전자공학회 하계학술대회. 2024.2553-2557

[3] 강혜란, 박성식, 김학수. 의료 질문 응답 시스템의 성능 향상을 위한 프롬프트 엔지니어링 연구. 한국컴퓨터종합학술대회. 2024. 513-515