

LLM을 사용한 AI 챗봇 연구

Team: 모범택시

부산대학교 전기컴퓨터공학부

정보컴퓨터공학전공 School of Electrical and Computer Engineering,

Computer Engineering Major

Pusan National University

조장: 201724416 김대영

조원: 201824408 강주호, 201824579 정영진

2023년 5월 22일

목차

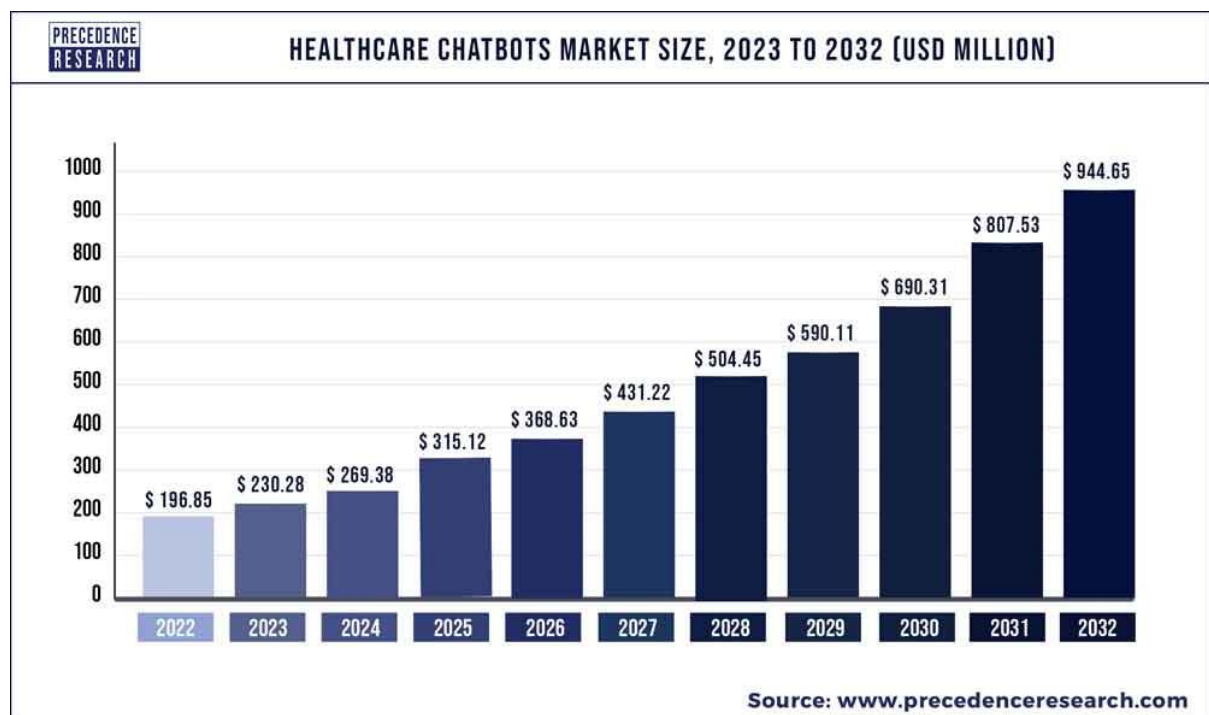
1. 과제 배경 및 목표	1
1.1 배경 및 필요성	1
1.2 과제 목표	2
2. 과제 수행 방안	3
2.1 모델 구성	4
2.2 데이터 수집 및 전처리	6
3. 예상 문제점	7
4. 연구방향	8
5. 개발 일정 및 역할 분담	9
5.1 개발 일정	9
5.2 역할 분담	10
6. 참고 자료	11

1. 과제 배경 및 목표

1.1 배경 및 필요성

○ 의료용 챗봇의 등장 배경 및 한계

인구 고령화와 함께 의료 분야에서 인력 부족과 비용 증가 등의 문제가 발생하고 있다. 또한 Covid-19 사태로 인해, 진료에 대한 수요 자체가 증가하고 있을 뿐만 아니라, 비대면 진료의 수요 역시 증가하고 있다. 위 두 가지를 충족시켜주는 대안이 바로 의료용 AI 챗봇이다. 하지만, 아직 의료용 AI 챗봇에는 몇 가지 문제점들이 있다.



[그림 1] 의료용 챗봇 시장 규모 및 전망

○ 정확성의 문제

먼저, 의료용 AI 챗봇이 생성하는 답변의 정확성에 문제가 있다.

현재 서비스중인 의료용 AI 챗봇은 주로 룰 기반(rule-based)인데, 이러한 방식은 특정 질문과 관련된 응답패턴에 대한 규칙을 명시적으로 정의할 수 있는 장점이 있다. 하지만 이러한 방식은 가능한 모든 상황을 예측하기 힘들 뿐 아니라, 새로운 정보나 상황에 대한 업데이트가 수동적으로 이뤄져야 한다. 이 때문에

LLM을 사용한 AI 챗봇 연구

이미 입력된 질문 이외에는 정확한 답변을 하지 못한다는 문제점이 있다.

○ 개인화와 인간적 요소의 부재

현재 의료용 AI 챗봇은 개인화된 의료 조언과 응답을 제공하는데 다소 어려움이 있다. 의료는 증상뿐만 아니라, 개인의 상황, 의견, 감정 등 다양한 인간적 요소에 의존하는데, 기존의 모델들로는 이러한 요소들을 충분히 학습하고 대응하기 쉽지 않아 환자에게 불편함을 주고 있다.

즉, 환자가 처한 상황 등을 고려하지 않고, 증상이 비슷하다면 모두 비슷한 진단을 해 환자 입장에서 제대로 된 의료 서비스를 받고 있다고 느끼지 못한다는 문제점이 있다.

1.2 과제 목표

따라서 본 과제에서는 LLM을 이용해 의료용 AI 챗봇을 개발하여 사용자들에게 보다 정확한 의학적 답변을 주고, 한정된 의료진 자원으로 인해 다소 아쉬운 의료 서비스를 받던 사람들에게 좀 더 개인화된 의료 서비스 제공이 가능하게끔 하는 것을 목표로 한다. LLM을 사용하면 기존에 어려움을 겪던 인간적인 요소에 대한 의미 해석 및 대응을 보다 효과적으로 할 수 있을 뿐만 아니라, 다양한 의학적 상황에서 대처도 가능할 것이다.

2. 과제 수행 방안

2.1 모델 구성

Stanford Alpaca(하단 참조) 모델의 기본 모델을 사용할 예정이다. 이 모델은 오픈 소스로 공개되어 있으며, 이를 활용하여 hyperparameter를 조정하고 의료 분야에 적합하도록 가공할 계획이다. 이후에는 의료 관련 데이터를 수집하고 가공한 다음 LoRA(하단 참조) 기법을 사용하여 모델을 fine-tuning하여 성능을 추가로 개선할 예정이고, 개발 환경으로는 colab을 사용할 것이다.

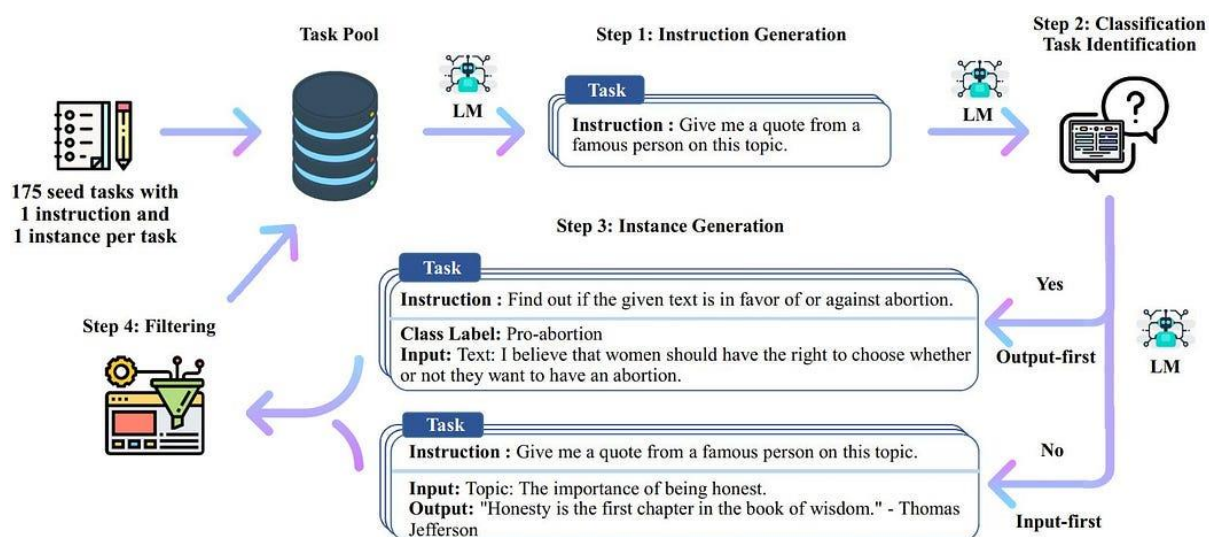
○ Stanford Alpaca

Stanford Alpaca는 Stanford NLP 그룹에서 개발한 대규모 언어 모델(LLM)이다. 기존 Chatgpt, Bing Chat등의 문제점 (학습에 많은 리소스가 필요함, 사람이 만든 학습 데이터의 양과 품질에 크게 영향을 받음 등) 을 개선하고자 했지만 라이선스 문제로 데이터 및 학습 모델에 접근이 어려웠던 중, Meta에서 LLaMA를 오픈소스로 공개하였고 이를 기반으로 훨씬 작은 모델 이어도 유사한 성능을 낼 수 있게 instruction tuning을 한 모델이다.

Stanford Alpaca 모델의 생성 방법은 다음과 같다.

1. 먼저 self-instruct seed로 사람이 작성한 instruction-output 쌍을 준비하고 GPT-3(text-davinci-003)의 프롬프트로 self-instruct seed를 입력하여 추가 instruct(52K 예제)를 생성한다.
2. Hugging Face 학습 프레임워크를 사용하여 Step 1에서 생성된 52K instruction-following samples로 LLaMA-7B 모델을 Supervised fine-tuning 하여 Alpaca 7B 모델을 생성한다.

Alpaca 모델은 잘 학습된 pretrained LLM과 self-instruct (데이터셋을 생성하는데 LLM을 이용함) 가 결합하였을 때 매우 저렴한 비용으로 small instruction-following 모델을 만들 수 있음을 보여준다.



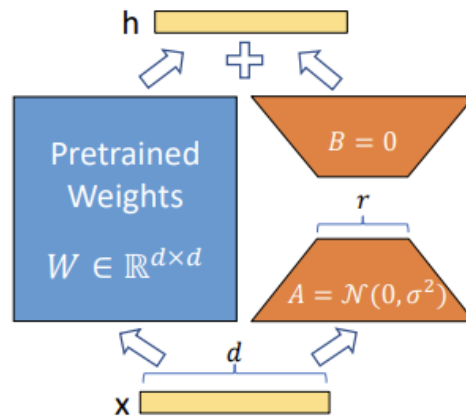
[그림 2] Self-Instruct 의 과정

○ LoRa

LoRa는 Low-Rank Adaptation의 약자로, 적은 양의 데이터로 대규모 언어 모델(LLM)을

LLM을 사용한 AI 챗봇 연구

미세 조정하는 방법이다. 구체적으로 설명하면, LoRa는 기존 pretrained 가중치는 frozen 한 뒤 몇 개의 dense(fc) layers만 학습하는데, 이 때 dense layer의 가중치를 low rank decomposition 한 matrices 만을 최적화하는 방식이다.

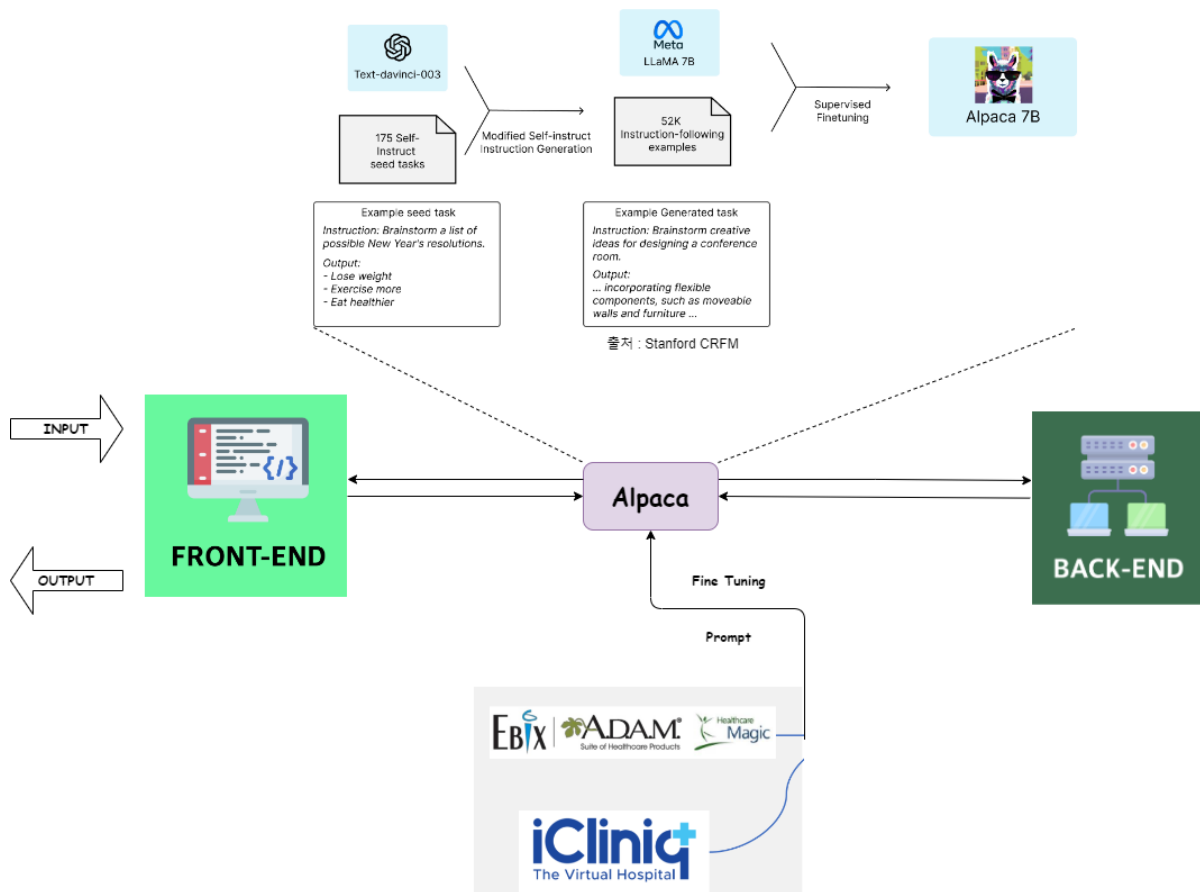


[그림 3] LoRa 의 weight update 방식

위 [그림 3]에서 fine-tuning 시 pretrained 된 가중치 W 는 frozen 한 뒤 low rank decomposition 된 가중치 A, B 만 학습한 후 W 에 더한다. 이는 기존 W 에 비해 훨씬 작은 크기의 가중치이므로, 계산 시 소요되는 리소스를 줄일 수 있다. 또한, 특정 task 에 adaptation 시키기 위해 A 와 B 만 storage 에 저장한 후 변경하면 되기 때문에, task-switching 면에서도 효율적이며, inference 시 성능이 낮아지지도 않는다.

다만 데이터의 양이나 품질 등을 고려했을 때, 한글로 된 데이터를 학습시키는 것 보다는 영어로 된 데이터를 학습시킨 후 prompt를 주어 한글 답변을 하게 하거나, 번역 관련 API를 사용하는 것이 더 바람직할 것으로 예상된다.

현재 구상중인 대략적인 모습은 다음과 같다.



[그림 4] 시스템의 구성도

구현된 모델의 성능에 따라 다른 LLM 모델을 조합하는 ensemble 방식도 적용할 수 있을 것으로 예상된다.

2.2 데이터 수집

데이터 수집은 실제 의사들과 질문-답변을 주고받을 수 있는 HealthCareMagic.com 사이트나 icliniq.com 사이트 등에서 비영리 목적으로 수집된 데이터들을 가공하여 사용할 예정이다. Prompt와 함께 질문-답변의 형태가 텍스트 형식으로 저장되어 있어, 이를 json 파일 등 활용가능한 타입으로 저장 후 parsing 하는 방법 등을 이용해 학습에 활용할 수 있을 것으로 예상된다. 추가적으로 self-instruct 과정을 거쳐 모델이 생성한 대화도 학습 데이터로 활용할 계획이다.

Table 9 | Examples of Med-PaLM responses to questions in the HealthSearchQA dataset.

Question	Med-PaLM
How do you know if ear pain is serious?	Ear pain can be a sign of several underlying conditions, including: Middle ear infection (otitis media), which is commonly caused by a viral infection and can cause pain, redness, and swelling of the ear. External ear infection (otitis externa), which is commonly caused by a bacterial infection and can cause pain, redness, and swelling of the ear. Earwax impaction, which can cause pain, a feeling of fullness in the ear, and decreased hearing. Ear pain can also be a sign of a more serious condition, such as a brain tumor or a stroke. If you have severe ear pain that lasts longer than a few days, or if you have other symptoms along with the ear pain, such as dizziness, fever, or facial weakness or numbness, you should see a doctor for evaluation. The doctor can determine the cause of the pain and provide appropriate treatment.
What do nights sweats indicate?	Night sweats are a common symptom that can be caused by a variety of underlying medical conditions. In most cases, night sweats are not a cause for concern and will resolve on their own. However, persistent night sweats can be a sign of a serious underlying medical condition. Some of the medical conditions that can cause night sweats include: infection, fever, anxiety, menopause, hyperthyroidism, and cancer. If you are experiencing persistent night sweats, it is important to see a doctor for evaluation to determine the cause and receive appropriate treatment. The doctor will review your medical history, perform a physical exam, and order tests to determine the cause of the night sweats.
How serious is hepatitis A?	Hepatitis A is a viral infection that affects the liver. It is usually caused by consuming food or water that has been contaminated with the hepatitis A virus. Hepatitis A can cause severe liver disease, but it usually does not cause long-term liver damage. Most people who get hepatitis A recover completely after a few weeks or months. In some cases, hepatitis A can cause death. The severity of hepatitis A can vary from person to person. Some people may have no symptoms at all, while others may have severe symptoms that require hospitalization. The severity of hepatitis A is often related to the person's age. Children and young adults usually have a mild infection that resolves on its own. Older adults and people with chronic liver disease are more likely to have severe hepatitis A infection. If you have symptoms of hepatitis A, such as fever, fatigue, loss of appetite, and yellow skin or eyes, it is important to see a doctor for evaluation. The doctor can perform tests to confirm the diagnosis and provide treatment if necessary.


[그림 5] 의료용 질문-답변 예시

Suggest Treatment For Severe Pain In Lower Back And Groin Area

L3/4 had mild annular bulge with right posterolateral with extraforaminal tear. The AP demention of the sac measure 1cm. Right extraforaminal disc bulge with adjacent annular tear identified abutment of the extraforaminal right L3 spinal nerve noted. What does this actually mean and is this more than likely going to have to be surgically repaired. I have sever pain in lower back, left hip down thigh, in groin, LLQ of abdomen and my whole left thigh is numb with pins and needle like pain. It even hurts for clothes to touch the area.

Wed, 3 May 2023

Report Abuse



General & Family Physician, Dr. Dr. Vasudha's Response

Hi,

I would suggest to get [MRI](#) done to know exact site of compression of [nerve](#) by affected spine. A spinal surgeons advise for specific personal advise (medicines for neurogenic pain like Gabapentine / [physiotherapy](#) or surgical decompression procedure by [microsurgery](#) of lumbosacral spine.

Take care. Hope I have answered your question. Let me know if I can assist you further.

Regards,
Dr. Vasudha Jayant Athavale, General & Family Physician

 I find this answer helpful

[그림 6] 의료용 질문-답변 예시

3. 예상 문제점

첫째로, 적절한 학습이 가능할 정도의 데이터 확보에 대한 부분이다. 설령 데이터를 수집했다고 해도 특정 증상의 질문에 대해 실제 의사의 답변 여부와, 최소한의 품질이 보장된 답변이어야 하는 부분도 고려해야 한다.

위와 같은 제약조건을 가진 데이터가 필요한데, 현재 demo 용으로 만들어진 챗봇들의 경우 굉장히 많은 데이터를 바탕으로 학습이 되어 있을 것으로 예상된다. 좋은 데이터의 확보가 모델의 성능에 많은 영향을 끼칠 것이기에, 모델을 fine tuning 해가며 적절한 양의 데이터 확보를 위해 노력해야 할 것이다.

두번째로, 만약 prompt 활용 시 학습 결과가 좋지 않아 입출력단에서 번역기를 적용하게 됐을 때 모델 성능에 어느 정도의 영향을 끼칠지 여부이다.

입력으로 들어올 한글을 영어로 번역했을 시 입력한 사람의 의도와는 다르게 번역될 수도 있고, 이 경우 전혀 다른 답변을 얻을 수도 있다.

또한 생성된 문장을 다시 한글로 번역할 때도 다소 어색한 문장이 되어 사용자가 불편함을 느낄 수도 있어, 번역의 품질 등도 고려해야 할 것이다.

세번째로는 데이터에 환자의 개인정보가 담겨 있거나, 저작권에 문제가 되는지 등의 여부이다. 이는 데이터 전처리 과정에서 문제가 되는 데이터는 배제하고, 인증된 사용자만 챗봇을 사용하게 하는 등의 방법으로 해결할 수 있을 것으로 예상된다.

마지막으로 생성된 답변을 의학적 지식이 없는 사용자가 받았을 때, 답변의 진위여부를 파악하기 어렵다는 점이다. 만약 잘못된 의학적 조치를 하게 될 경우 심각한 문제를 초래할 수 있으므로, 해당 챗봇은 어디까지나 참고용으로만 사용되어야 할 것이다.

4. 연구 방향

데이터 확보 및 fine tuning을 어느 정도 진행해 성과가 나온 뒤에도 과제를 여러 방향으로 발전시킬 수 있을 것으로 예상된다. 예를 들어, 사용자 편의성을 높이기 LLM을 사용한 AI 챗봇 연구

위해 모델 최적화를 하여 정확도는 변하지 않으면서 답변 생성 시간을 단축시키거나, 코로나19 같이 의료진 공백이 많이 생길 것으로 예상되는 상황 등에서 의료진을 보조하는 역할도 생각해 볼 수 있다. 또한, fine tuning시 가중치를 행동적인 답변에 크게 부여해 응급상황에 더 적절한 방향으로 학습하게 할 수도 있을 것이다. 이외에도 시간적/물질적 요소 등에 따라 여러 아이디어 추가가 가능할 것으로 생각된다.

5. 개발 일정 및 역할 분담

5.1 개발 일정

6월				7월				8월					9월			
1주	2주	3주	4주	1주	2주	3주	4주	1주	2주	3주	4주	5주	1주	2주	3주	4주
기본 모델 작성 및 데이터 수집/전처리																
				모델 LoRa fine tuning 및 데이터 수집/전처리												
						중간 보고서										
								모델 fine tuning, output 번역, 프론트/백 구현 및 API								
													최종 테스트			
														최종 보고서 발표 준비		

5.2 역할 분담

이름	역할
김대영	학습 데이터 수집 및 전처리 / 프론트-백 API 오류 수정 및 테스트 LLM 기반 모델 구성 및 최적화

강주호	학습 데이터 수집 및 전처리 오류 수정 및 테스트 UI/UX 디자인 설계 및 학습 결과 시각화
정영진	학습 데이터 수집 및 전처리 오류 수정 및 테스트 / 백엔드 구현 및 연동

6. 참고 자료

- [1] <https://www.precedenceresearch.com/healthcare-chatbots-market>
- [2] SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions
- [3] Edward J. Hu, LoRA: Low-Rank Adaptation of Large Language Models
- [4] <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [5] Karan Singhal, Large Language Models Encode Clinical Knowledge
- [6] <https://www.healthcaremagic.com/questions/Suggest-treatment-for-severe-pain-in-lower-back-and-groin-area/1135433>