Open Source SW Contribution

제출일	2023.04.04	담당교수	송인식 교수님
과목	오픈소스 SW 기여	프로젝트명	Dr.Bot
학번 / 이름			
32197256 박성우		32173575 이해주	

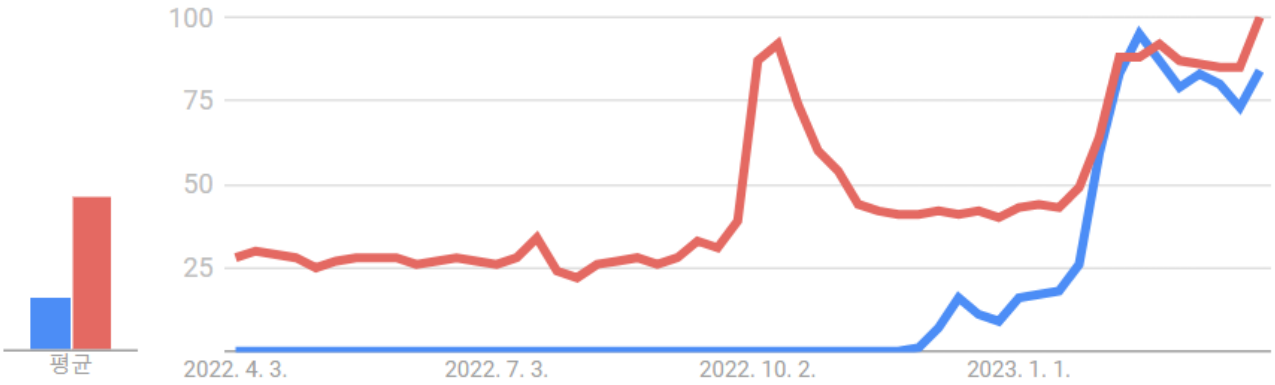
Contents

1. 개요

1.1. 배경, 목적 및 필요성

과학잡지뿐 아니라 문화 콘텐츠 관련 잡지에서도 ChatGPT 특집기사가 최근 몇 달간 쏟아져 나왔다. 잡지 기획안을 직접 ChatGPT 로 작성하거나 여성 모델 그림을 AI 로 그렸는데 사진과 구별이 안 될 정도로 높은 수준이라는 것이다. 상반기 대졸 신입사원 면접에서도 ChatGPT 에 대한 전망과 미래 등이 단골 예상 면접 문제로 거론될 만큼 ChatGPT 는 현재 전 세계에서 최고의 전성기를 맞이하고 있다.

OpenAI 에서 ChatGPT 를 출시함에 따라 AI 에 대한 관심이 이전보다 급속도록 높아진 것을 체감할 수 있다. 다음은 구글 트렌드를 통해 시간 흐름에 따른 관심도의 변화를 나타낸 그래프이다. 빨간선은 AI, 파란선은 ChatGPT 를 나타낸다.



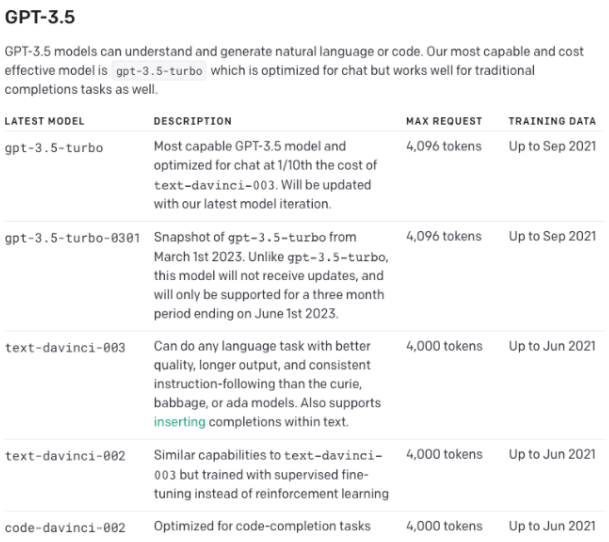
[그림 1] 시간 흐름에 따른 관심도 변화

그래프에서 나타나듯 지난 2022 년 8 월 미국에서 열린 미술대회에서 AI 로 그린 그림이 우승을 차지하며, AI 와 예술의 경계에 대한 논란으로 잠깐 뜨거워졌던 AI 에 대한 관심도가 2 월에 ChatGPT 출시와 함께 가파르게 높아진 것을 확인할 수 있다.

ChatGPT 는 GPT 3.5 라는 언어모델을 토대로 사용자와 대화를 주고받으며 질문에 답하는 대화형 AI 서비스를 의미한다. 해당 서비스는 출시 두 달 만에 월 이용자가 1 억 명을 돌파할 정도로 글로벌 메가히트를 기록했다. 국내에서도 대한상공회의소가 지난 2 월 20_60 대 성인을 대항으로 조사한 결과, 전체 응답자의 35.8%가 ChatGPT 를 사용했다고 답했다. ChatGPT 는 사용자와 대화를 나누며 대화 내용을 학습, 축적한 후 사용자의 질문을 듣고 더욱 최적화된 답변을 내놓는 강화학습(Reinforcement Learning) 알고리즘을 탑재했다. 1750 억 개의 매개변수를 토대로 전 세계 수억 명의 사용자가 공개한 텍스트를 실시간으로 학습한다. 사용자가 더 많이 활용할수록 ChatGPT 의 지능이 더 높게 올라간다는 의미이다.

3 월 2 일 OpenAI 사에서 ChatGPT API 를 공개했다. OpenAI 의 ChatGPT 자료[i]를 보면 GPT3.5 에서 fine-tuned 했다고 설명이 되어 있다. 다시 말해, GPT3.5 는 ChatGPT 의 일종의 전신이라고 표현할 수 있다. 따라서, ChatGPT 와 유사한 성능의 모델을 사용하려면 GPT3.5 기준의 모델을 사용해야 한다.

Text-davinci-003 모델을 사용하면 GPT3.5 를 사용해볼 수 있다는 설명이다. GPT3.5 계열의 모델을은 아래와 같이 존재하며, Text-davinci-003 외에도 text-curie-001, text-babbagte-001, text-ada-001 과 같은 모델들이 존재한다. Gpt-3.5-turbo 가 ChatGPT 모델이라고 할 수 있다.

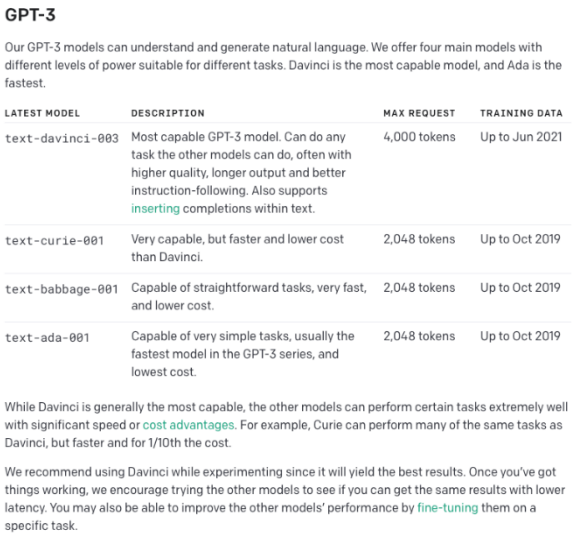


GPT-3.5

GPT-3.5 models can understand and generate natural language or code. Our most capable and cost effective model is `gpt-3.5-turbo` which is optimized for chat but works well for traditional completions tasks as well.

LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
`gpt-3.5-turbo`	Most capable GPT-3.5 model and optimized for chat at 1/10th the cost of `text-davinci-003`. Will be updated with our latest model iteration.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-0301`	Snapshot of `gpt-3.5-turbo` from March 1st 2023. Unlike `gpt-3.5-turbo`, this model will not receive updates, and will only be supported for a three month period ending on June 1st 2023.	4,096 tokens	Up to Sep 2021
`text-davinci-003`	Can do any language task with better quality, longer output, and consistent instruction-following than the curie, babbage, or ada models. Also supports inserting completions within text.	4,000 tokens	Up to Jun 2021
`text-davinci-002`	Similar capabilities to `text-davinci-003` but trained with supervised fine-tuning instead of reinforcement learning	4,000 tokens	Up to Jun 2021
`code-davinci-002`	Optimized for code-completion tasks	4,000 tokens	Up to Jun 2021

[그림 2]GPT-3.5



GPT-3

Our GPT-3 models can understand and generate natural language. We offer four main models with different levels of power suitable for different tasks. Davinci is the most capable model, and Ada is the fastest.

LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
`text-davinci-003`	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality, longer output and better instruction-following. Also supports inserting completions within text.	4,000 tokens	Up to Jun 2021
`text-curie-001`	Very capable, but faster and lower cost than Davinci.	2,048 tokens	Up to Oct 2019
`text-babbage-001`	Capable of straightforward tasks, very fast, and lower cost.	2,048 tokens	Up to Oct 2019
`text-ada-001`	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,048 tokens	Up to Oct 2019

While Davinci is generally the most capable, the other models can perform certain tasks extremely well with significant speed or cost advantages. For example, Curie can perform many of the same tasks as Davinci, but faster and for 1/10th the cost.

We recommend using Davinci while experimenting since it will yield the best results. Once you've got things working, we encourage trying the other models to see if you can get the same results with lower latency. You may also be able to improve the other models' performance by fine-tuning them on a specific task.

[그림 3] GPT-3

본 프로젝트에서는 OpenAI 에서 공개한 API 를 토대로 의료 챗봇을 만들어 봄으로써 거대 언어모델(Large Language Model, LLM)과 학습된 모델을 미세조정(Fine tuning)하는 방법에 대해 이해하고 평가하는 예시를 제공한다. 예상되는 기대효과로는 사용자가 간단한 키워드, 혹은 구체적인 질문으로 병의 증세에 대한 답변을 얻을 수 있다. 또한 복잡한 사용방법 없이 직관적인 사용이 가능하다.

1.2. 시나리오

서비스 시나리오는 초기 이 챗봇 구상할 때 생각했던 다양한 타깃으로 하여 짜보았다.

구성요소	설명
사용자 그룹	의료서비스 관련 전문적인 답변이 필요한 스마트폰 이용자

[i] https://openai.com/blog/chatgpt

시놉시스	1) A 씨는 원인 모를 복통으로 불편함을 호소하고 있으며 병원에 갈만큼 아프지만 현재 업무량이 많아 당장 병원을 갈 수 없는 상황이다. 따라서 간단하게라도 진단을 받고자하며 증상이 완화되기를 원한다. 2) B 씨는 장시간 두통으로 인해 약국 약을 복용하였으나 증상이 완화되지 않았다. 병원에 방문할 만큼 아픈 것은 아니지만 증상이 잘 낫지 않기 때문에 원인과 도움이 되는 해결법을 제공받고자 한다.
니즈	병원에 가지 않고 통증의 원인이 될 수 있는 요인들을 제공받고 증상 완화와 치유 촉진에 도움이 되는 방법들을 제안 받고자 한다.
불편사항	인터넷에서는 환자에게 주어진 상황에 맞는 통증 원인을 정확히 알아낼 수 없으며 해당 정보의 정확성을 알 수 없기 때문에 신뢰성이 떨어진다. 하지만 의료 챗봇을 이용한다면 사용자가 언급한 증상들을 기반으로 알맞는 정보를 제공할 수 있다.
대안마련	사용자가 알고자 하는 증상의 원인을 간단한 채팅(키워드)로 파악할 수 있으며 높은 신뢰성을 제공하며 시간이 없는 사람들은 병원에 가지 않고도 어느정도 자신의 문제를 파악하여 증상을 완화시킬 수 있다.

[표 1] 서비스 시나리오

2. 현황 분석

2.1. 국내외 사례 조사 및 분석

① 카카오 i 커넥트 톡 - 유통

코로나 19 이후 언택트 상담의 필요성 증가로 인해 고객센터를 통한 전화 문의가 급증하여 이 때문에 길어지는 대기 시간으로 이용자들은 상담 도중 전화를 끊어 버리는 사태가 빈번하게 발생하기 시작했다. 이는 고객의 불만으로 이어졌으며 고객센터만으로는 원활한 운영이 불가능한 상태에 이르게 되었다. 따라서 해당 유통 업체는 전화 상담을 일부 디지털로 변환하고자 하였으며, 챗봇을 통해 신제품이나 이벤트성 정보를 전달하고자 하였다. 이에 카카오 I 커넥트 측에서는 고개사의 브랜딩 요소와 디자인을 고려한 이미지를 도입하였으며, 실제 고객사와 챗봇 사이에 발생하는 이질감을 줄이는 작업을 진행했다.

‘카카오 싱크’는 한 번의 동의만으로 모든 서비스를 사용할 수 있게 하는 시스템으로 사용자들이 번거롭게 회원가입을 하지 않아도 즉각적으로 사용할 수 있도록 편리함을 제공했다.

‘Advanced ML 인공지능 엔진’은 대화형 인공지능으로 적은 학습 데이터로 사용자의 의도를 파악할 수 있는 카카오의 고유적이며 차별적인 기술을 사용했다. 챗봇에서 답변을 하지 못한 경우 바로 상담사와의 채팅으로 전환하도록 이끌어냈으며, 이 때 챗봇에 스몰톡을 할 수 있도록 학습하여 이용자의 의도를 파악하지 못했을 경우에도 사용자가 거부감이 들지 않도록 자연스러운 사용감을 유도했다.

② WhatsApp Chatbot – 헬스케어

코로나 팬데믹으로 인해 대부분의 의료기관들은 AI 기반의 챗봇을 의료 인프라와 융합하여 환자에게 최상의 치료를 제공하는 데에 도움을 받고, 도움이 되었다. WhatsApp 의 챗봇은 내부 데이터베이스와 추가로 통합되어 학습, 훈련으로 답변의 정확도가 높으며 NLP 로 구동되기 때문에 사용자 쿼리를 빠르게 이해하고 실시간으로 정확한 응답을 제공할 수 있다.

또한 헬스케어에서 가장 민감한 부분인 의료 정보에 액세스(접근)하는 것을 가능하게 하면서 증상과 병력을 원할하게 공유할 수 있도록 했다. 이를 통해 테스트와 결과를 주고받을 수 있으며, 문서를 종이형태로 들고 다니지 않고 암호화된 폴더에 안전하게 챗봇 자체에서 보관이 가능해지게 되었다. 간단하게는 예약, 의료 알림, 결제 등의 기능까지 지원을 했다.

③ 의료용 AI 챗봇 메드팜

메드팜은 구글과 딥마인드가 의료 전문가와 환자가 제기한 질문에 실용적이며 안전한 답변을 줄 수 있도록 설계되어 출시가 되었다. 사용자의 질문을 이해하여 일반 언어로 응답할 수 있도록 설계된 대규모 언어 모델로 설계 되었다. 이는 5400 억개의 매개변수로 구성된 구글의 LLM 인 ‘PaLM’을 기반으로 하며 전문 의료 검사, 연구 및 의료에 대한 소비자 질문을 다루는 7 개의 질문-응답 데이터 셋에 대한 훈련을 받았다.

메드팜의 가장 큰 장점은 정확도가 높아진 것인데 임상의의 답변과 92.6% 일치한다는 것이다. 이는 의료와 관련되어 민감한 부분을 더 안전하게 제공하여 안심하고 사용할 수 있다는 것을 의미한다.

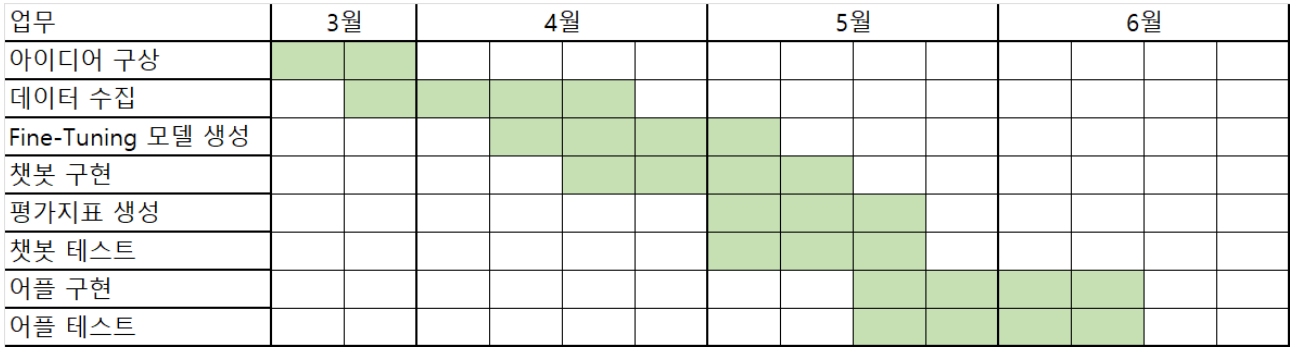
3. 플랫폼 추진 방법, 전략 추진 및 추진 일정

3.1. 추진 체계

직책	구성원
백엔드, PL, PE	박성우
프론트엔드, PM, PE	이해주

[표 2] 팀 멤버 및 역할 소개

3.2. 추진 일정



[그림 4] 플랫폼 추진일정 Short-term

4. 프로젝트 내용

4.1. 기능 및 용어

① LLM (Large Language Model)

대형언어모델은 GAN 과 함께 AI 분야에서 최근 폭발적인 발전을 이끈 신경망이다. 오픈 AI 의 GPT 시리즈, 구글의 팜(PaLM) 또는 메타의 라마(LLaMa) 등이 LLM 으로 방대한 텍스트 데이터를 사용해 훈련된다. LLM 은 고품질의 웹 문서, 책, 위키피디아의 기사들, 블로그 글과 깃허브의 오픈소스 코드 등 공개된 데이터를 활용해 학습한다. 이를 토대로 텍스트가 주어지면 다음에 어떤 텍스트가 올지를 확률적으로 예측해 내고 그 결과 더 긴 텍스트가 생성되면 또 다시 다음 텍스트를 예측하는 과정을 되풀이한다.

② 미세조정 (Fine tuning)

GPT 시리즈와 같은 대형 AI는 범용성이 뛰어난 기초(foundational) 모델이기 때문에 특정한 AI도구 개발에 쓸 때 이용할 수 있다. 이때 미세조정이 필요하다. 기초 모델에 특정 데이터를 추가로 학습시켜 적은 비용으로 매우 특화된 기술을 가진 모델을 개발할 수 있다.

Chat GPT 에서 사용하는 데이터는 2021 년 까지이다. 2022 년도부터의 질문에는 답하지 못한다. 따라서 원하는 format 으로 원하는 답이 나오도록 유도하기 위해 fine-tuning 과정이 필요하다.

Fine-tuning 절차로는 먼저 format 을 맞추기 위한 데이터를 준비해야 한다. 그 다음 준비한 데이터를 토대로 모델을 생성하고 생성된 모델을 활용하여 챗봇을 구축한다. 이때 openAI playground 에서 사용자가 직접 만든 모델을 업로드하거나 다른 사람이 업로드해 놓은 모델을 사용할 수 있다. Fine-tuing 을 위해 준비해야 할 데이터로는 JSONL 형태의 데이터와 파일 안에는 prompt, completion 의 형태로 구성해야 한다. Prompt 는 원하는 프로프트를 구성하고, completion 은 원하는 답의 형태를 구성한다.

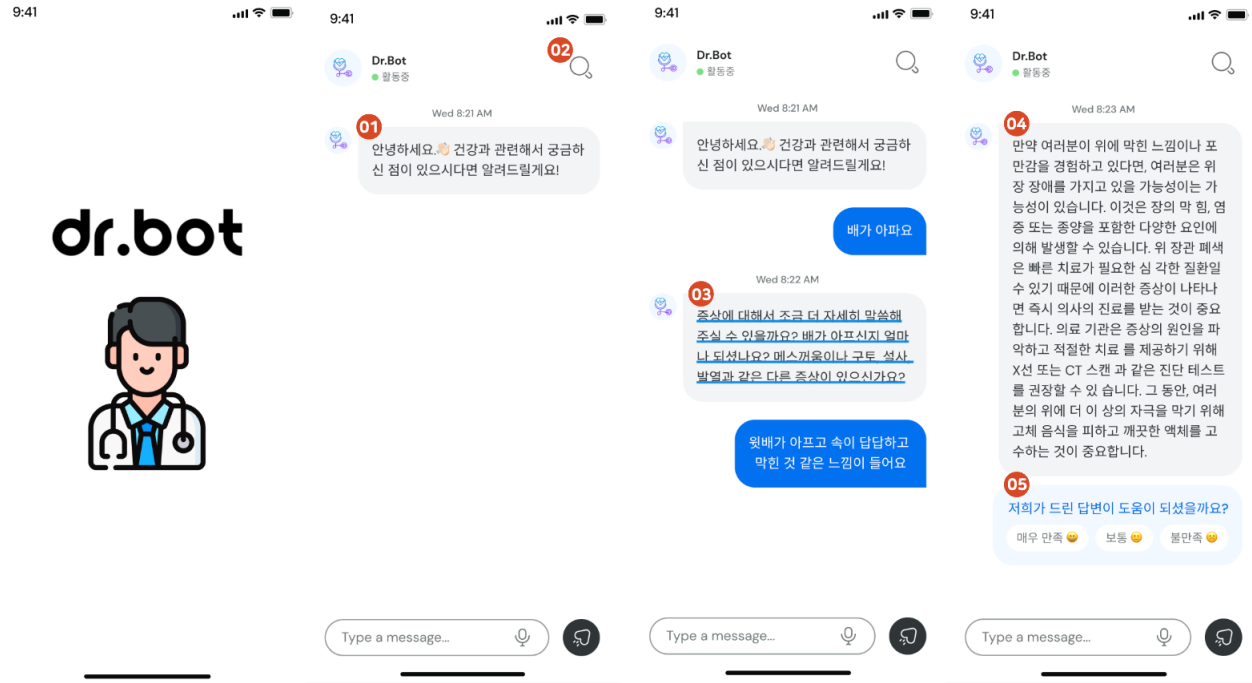
③ 퓨샷러닝 (Few-Shot Learning)

의료 분야와 같은 충분한 데이터를 구할 수 없거나, 레이블 데이터셋 생성에 많은 시간적, 비용적 부담이 요구되는 분야에서 Few-Shot Learning(FSL) 학습방식을 도입하여 인간의 추론과정을 모방하여 제한된 데이터로도 별도의 증강 없이 다양한 태스크에서 높은 일반화 성능을 달성하여 한계점을 극복하고자 하는 학습방식이다.

④ 분기계수 (Perplexity)

이전 단어로 다음 단어를 예측할 때 몇 개의 단어 후보를 고려하는지를 의미한다. 고려해야 할 단어 후보가 많다는 것은 그만큼 언어 모델이 쉽게 정답을 못 내고 있다고 해석할 수 있다. 즉, Perplexity 값이 낮을수록 언어 모델이 쉽게 정답을 찾아내는 것이므로 성능이 우수하다고 평가한다.

4.2. 요구사항



[그림 5] 요구사항 UI 피그마

챗봇의 특성상을 앱을 따로 설치하는 것이 아닌 보통 부가적인 서비스 형태로 이용자들에게 제공된다. 그 이유는 이용자들이 해당 앱 및 분야와 관련해서 간단하고 편리하게 질문을 빠르게 응답을 받고자 하기 때문이다. 앞서 언급한 사례들도 마찬가지로 독립적인 앱을 개발한 것이 아닌 카카오톡, WhatsApp 과 같은 메신저 앱을 통해 서비스를 제공한다. 하지만 해당 프로젝트는 해당 챗봇의 구현된 모습을 보여주고자 앱으로 개발하였다.

앱의 로고와 함께 메인 화면을 지난 후 사용자는 즉시 채팅을 할 수 있도록 이루어져 있다.

① 챗봇이 사용자에게 인사말을 건내는 것으로 채팅이 시작된다.

② 검색기능 : 대부분의 챗봇은 사용자와 상담한 내용이 휘발성으로 사라지게 된다. 하지만 이 챗봇은 사용자의 과거 채팅 이력을 보관하여 사용자가 이전에 어떠한 질문을 했었는지 검색이 가능하도록 설계하였다.

③ 사용자의 질문을 통해 알아들을 수 없거나 정보가 부족한 경우 사용자가 언급한 키워드를 이용해 증상을 유추하여 자세한 답변을 유도한다.

④ 데이터 셋에 있는 정보를 가공하여 사용자에게 답변을 한다.

⑤ 마지막으로 채팅이 끝난 후 해당 챗봇의 유지보수 및 보완을 위해 사용자에게 설문을 하여 데이터를 축적한다.

4.3. 위험 평가 및 대책

예상되는 위험 요인으로는 잘못된 정보를 제공함으로써 사용자의 병세를 키울 수 있다. 이에 대한 대책으로는 응급처치 방법, 증상 완화에 도움이 되는 간단한 조치 등을 제시하고 "증상이 지속되거나 악화될 시 반드시 병원 진단을 받을 것"을 권고한다. 또한 다른 사람의 개인 정보를 잘못 출력하는 등의 악용될 우려가 있다.

해외 인터넷 보안업체 노드 VPN 은 14 일 오픈 AI 에서 만든 챗 GPT 를 해킹 수단으로 악용하려는 악성 해커들이 급격하게 늘고 있다고 경고했다. 이 업체에 따르면, 해커들이 주로 이용하는 인터넷 게시판 다크웹에 챗 GPT 관련 문의가 올해 1 월 120 건에서 2 월 870 건으로 625% 증가했다. 게시물은 주로 챗 GPT 를 해킹하거나 유인 도구로 활용하는 방법, 챗 GPT 로 악성코드를 퍼뜨리는 방법을 공유하는 내용들이다. 이처럼 다른 사람의 개인정보를 침해하거나 잘못 출력하도록 유도하며 악용될 우려가 있다. 따라서 회원가입 등의 개인정보를 입력하는 기능을 추가하지 않는 것으로 대처할 예정이다.

5. 참고문헌

- 정병일. 2023.03.03. AI 이해에 필요한 개념 7 가지. AI 타임스.
- 박서형, 임유한, 이상철. (2022). Temporal Iterative EPNet : 퓨샷 러닝을 이용한 시계열 분류 모델. 한국정보과학회 학술발표논문집, (), 699-701.
- 신운섭, 송상헌. (2022). 언어 인공지능의 상식추론과 평가 체계 현황. 인문사회과학연구, 23(3), 133-166.
- 우상명. 2022.06.28. 자연어 생성과 평가 방법 (NLP generation evaluation). Testworks.
- Tony Park. 2022.04.05. [NLP] 언어모델의 평가지표 'Perplexity' 개념 및 계산 방법. https://heytech.tistory.com/344
- 강정석, 권성재, 서원진, & 장소진. (2017). 챗봇을 활용한 전문의료 상담 및 예약플랫폼 "헬스챗". 한국정보처리학회 학술대회논문집, 24(2), 778-780.