

Digital Insight 2023

ChatGPT의 파급 효과와 기관의 LLM 도입 전략

Contents

1

LLM이란 무엇인가?

- A 기존 AI의 문제점 • 01
- B 전이 학습(Transfer Learning)과 기초 모델(Foundation Model) • 02
- C GPT-3의 독특한 전이 학습 • 04
- D GPT-3로 인한 변화 • 06
- E LLM 의 확장 • 07
- F 개발된 LLM 현황 • 07

2

ChatGPT 유료화가 시장에 미치는 영향

- A 유료 서비스로 인한 AI 발전 가속화 • 10
- B 유료화의 다른 이름은 글로벌 경쟁 • 11
- C 학문적인 AI에서 상업적인 AI로 • 12
- D AI 연구자에서 AI 사용자로 • 14
- E 보편적 서비스에서 산업별 서비스로 • 15
- F 일반적인 용도에서 기업에 특화된 LLM으로 • 16
- G 주니어 여러 명 대신 ChatGPT를 잘 쓰는 시니어 한 명 • 17

3

검색 시장의 변화를 비롯한 ChatGPT 파급 효과 분석

- A 일하는 방식의 변화 • 18
- B 생성형 AI를 활용한 사용자 인터페이스의 전환 • 19
- C 소비자의 구매 의사 결정 과정 중 ChatGPT의 역할 • 20

4

공공기관의 LLM 활용을 위한 제언

공공기관의 LLM 활용을 위한 제언 • 25

5

공공기관의 LLM 도입을 위한 5개의 아키텍처

- A 타입A: GPT-4 액세스를 API를 통해서 구현 • 28
- B 타입B: 오픈 소스 LLM을 사용 • 29
- C 타입C: 파인 튜닝 방식 • 30
- D 타입D: 오픈AI의 파인 튜닝 방식 • 31
- E 타입E: 오픈AI의 Enterprise ChatGPT • 32
- F 공공기관에 적합한 모델 • 32

그림 목차

[그림 1] 공간 컴퓨팅 등장 배경	• 03
[그림 1] 전이 학습의 장점과 단점	• 02
[그림 2] GPT-3의 독특한 사용방식	• 05
[그림 3] 현재 LLM의 현황	• 08
[그림 4] 오픈AI의 Technical Report	• 13
[그림 5] GPT-4의 답변	• 15
[그림 6] ChatGPT에서 “Chat history & training”을 해제하면 입력된 데이터를 학습에 사용하지 않음	• 16
[그림 7] 기존의 검색에 의한 문서 작성 방법과 ChatGPT를 활용한 문서 작성 비교	• 18
[그림 8] 구매의사결정과정의 변화	• 21
[그림 9] 러닝화에 대한 GPT-4의 답변	• 22
[그림 10] 러닝화에 대한 네이버 Cue:의 답변	• 23
[그림 11] 타입A: GPT-4 직접 액세스를 오픈AI API로 구현	• 28
[그림 12] 타입B: 오픈 소스 LLM 사용	• 29
[그림 13] 타입C: 파인 튜닝 방식	• 30
[그림 14] 타입D: 오픈AI의 파인 튜닝	• 31
[그림 15] 타입E: Enterprise ChatGPT 형태	• 32
[그림 16] 공공기관의 LLM 모델	• 33

1

LLM이란 무엇인가?

LLM은 'Large Language Model'의 약자로, 그동안 한국어로는 '초거대 AI'라고 번역되었다. 그런데 '초거대 AI'라고 하면 덩치가 큰 AI' 정도로만 인식될 수 있기 때문에, 여기서는 국제적으로 통용되는 용어인 'LLM'으로 지칭한다.

LLM은 2020년 오픈AI(OpenAI)에서 만든 GPT-3가 1750억 개의 파라미터(Parameter)를 갖고 있다고 해서 붙여진 이름이다. 이후 ChatGPT(GPT-3.5)가 발표되자 그 혁신적인 기능에 사람들이 주목하기 시작했고, 이에 ChatGPT와 같은 유형의 거대 언어 모델을 LLM으로 지칭하게 되었다. 이런 배경을 토대로 현재의 LLM이 앞으로 어떻게 나아갈 것인지 살펴보고자 한다.

A 기존 AI의 문제점

AI의 발전에 따라 점점 더 많은 사람이 딥러닝을 활용하고 있다. 그러자 이전에는 미처 생각하지 못한 새로운 난관들이 생겨나기 시작했다.

첫째, 너무나 많은 데이터를 필요로 한다. 많은 데이터를 학습시킬수록 AI의 성능이 좋아져 더 정확한 결과를 만들 수 있지만, 필요한 데이터가 상상 이상으로 방대하다는 것이 문제다. 게다가 각 데이터에 이름을 붙이는 데이터 라벨링(Data Labeling) 작업에도 엄청난 시간과 비용이 든다. 사람이 일일이 이름을 붙이는 작업이기 때문이다. 또한 라벨링은 단순 작업처럼 보여도 어떤 목적의 AI 모델을 개발하느냐에 따라 그 규칙이 달라지므로 해당 분야의 전문적인 지식도 필요하다.

둘째, 딥러닝 모델을 만드는 일이 쉽지 않다. 원하는 목적에 꼭 맞는 딥러닝 모델을 만드는 것은 생각보다 어렵다. 이는 소수의 AI 전문가만이 가능한 일이어서 그들의 연봉 또한 매우 높을뿐더러 책임자를 구하기도 힘들다.

셋째, 장비를 갖추기 위한 비용이 많이 든다. AI 모델을 컴퓨터에 학습시키려면 방대한 데이터를 처리하는 GPU를 장착한 컴퓨터가 대량으로 필요하며, 이러한 장비를 갖추기 위한 인프라 비용은 당연히 만만치 않다.

B 전이 학습(Transfer Learning)과 기초 모델(Foundation Model)

기업이나 기관에서 AI를 활용하려면 하루빨리 앞서 언급한 문제들을 해결해야 하는데, 이를 위해 등장한 것이 바로 ‘전이 학습(Transfer Learning)’이다.



[그림 1] 전이 학습의 장점과 단점

예를 들어 어떤 사람이 짬뽕 전문점을 개업한다고 할 때, 맛있는 짬뽕 레시피를 직접 처음부터 개발하려면 시간과 비용이 많이 들 수밖에 없다. 하지만 최고의 중식 셰프의 지도를 통해 레시피를 ‘전수’받으면 훨씬 빠르고 효율적으로 개업까지 할 수 있다.

전이 학습이 바로 이와 같다. 어떤 AI 모델을 개발하기 전에 먼저 그 분야에서 검증되고 널리 알려진 모델을 이용하면 개발 시간을 단축하고 성능도 빠르게 향상시킬 수 있다.

전이 학습은 딥러닝에서 매우 중요하다. 기존에 존재하는 완성된 AI 모델에 자신이 원하는 모델과 데이터를 학습시키면 보다 수월하게 모델링이 가능하고 정확도도 높아지며, 기업에서는 처음부터 AI 모델을 새로 개발해야 하는 수고를 덜 수도 있다.

한편 전이 학습에서는 기존에 학습된 모델을 사전 학습 모델(Pretrained Model) 또는 파운데이션 모델(Foundation Model), 즉 기초 모델이라고 부른다. 또한 이를 활용해 자신의 모델과 데이터로 학습하는 것을 미세 조정, 즉 파인 튜닝(Fine Tuning)이라고 지칭한다.

앞서 짚뽕 전문점의 사례를 보면, 짚뽕 분야 최고 전문가의 기술 지도가 기초 모델이고 이 기술 지도를 받아 자신만의 레시피를 만드는 것이 파인 튜닝이다. 이 방법은 컴퓨터 비전 및 자연어 처리 작업에도 적용할 수 있다.

대표적인 기초 모델로는 이미지 인식 분야의 VGGNet, ResNet, MobileNet, EfficientNet과 자연어 처리 분야의 BERT, GPT, GPT-2, RoBERTa 등이 있다. 이는 모두 대규모 데이터셋에서 사전 학습된 언어 모델로 다양한 자연어 처리 분야에서 높은 성능을 발휘한다.

예를 들어 의료 분야의 MRI 판독 시스템을 만든다고 하면, 먼저 마이크로소프트 연구진이 개발한 ResNet이나 옥스퍼드대학의 연구팀이 만든 VGGNet을 기초 모델로 선택한다. 그리고 MRI 판독 모델의 실제 판정 부분만 추가로 개발한 후 여기에 학습시킬 MRI 데이터를 준비한다. 이렇게 학습시키면 훨씬 적은 양의 MRI 데이터를 가지고도 높은 정확도와 빠른 학습 속도를 가진 시스템을 만들어 낼 수 있다.

결국 딥러닝에서 전이 학습이 중요한 이유를 정리하면 다음과 같다.

첫째, 학습 시간 및 컴퓨팅 자원을 절약할 수 있다. 처음부터 딥러닝 모델을 개발하고 학습시키는 것은 시간과 비용이 많이 소요된다. 이때 기초 모델을 전이 학습하면 훨씬 적은 자원으로 빠른 학습이 가능하다.

둘째, 성능을 대폭 개선할 수 있다. 기초 모델은 새로운 작업이 들어왔을 때 적용할 수 있는 기능을 이미 학습한 상태이기 때문에 처음부터 학습을 진행하는 것에 비해 대폭 향상된 성능을 제공한다.

셋째, 적은 양의 라벨링된 데이터로도 학습이 가능하다. 라벨링된 데이터는 매우 한정적인 데다, 앞서 말했듯 라벨링한 데이터를 얻는 데는 비용이 많이 든다. 하지만 전이 학습을 하면 적은 양의 라벨링된 데이터로 원하는 목표를 빠르게 달성할 수 있다.

넷째, 적은 양의 데이터셋으로도 학습이 가능하다. 전이 학습은 학습 데이터가 제한된 작업에서 특히 유용하다. 기초 모델은 이미 일반적인 특징을 학습한 상태이기 때문에 적은 양의 데이터셋으로 새로운 작업에 맞게 조정하기에 수월하다.

하지만 전이 학습에도 다음과 같은 몇 가지 문제가 있다.

첫째, 자신이 원하는 목적에 맞는 모델과 데이터를 별도로 준비해야 한다. 기초 모델이 있다고 해서 학습이 모두 해결되는 것은 아니며, 성공적인 학습을 위한 데이터는 사용자들이 준비해야 한다. 또한 이를 학습하는 과정에서 GPU가 탑재된 컴퓨터도 필요하다. 이는 마치 아무리 최고의 짬뽕 레시피를 전수받았다고 해도 영업을 시작하려면 가게와 식자재, 도구가 필요한 것과 같다.

둘째, 기초 모델이 있다고 모든 과정이 단번에 해결되는 것은 아니다. 기초 모델은 모든 일을 할 수 있지만, 이를 파인 튜닝해서 만든 모델은 그중 일부만 할 수 있다. 이는 마치 전문 중식 셰프는 모든 중국 요리가 가능하지만, 그로부터 짬뽕 레시피를 전수받은 사람은 짬뽕만 잘 만들 수 있는 것과 같다.

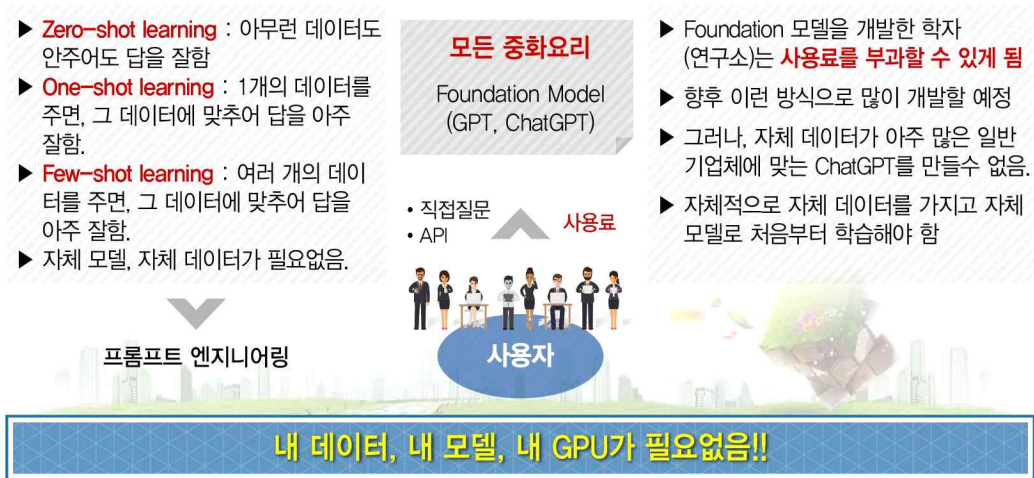
셋째, 기초 모델을 최초로 만든 사람이나 기업에게 보상이 별로 없다. 현재는 힘들게 좋은 AI모델을 만들어도 사람들이 완성된 모델을 별다른 비용 지불 없이 손쉽게 다운받아 사용할 수 있으며, 막상 그것을 만든 주체에는 큰 혜택이 없다. 이는 연구자들의 의욕을 꺾는 일이 될 수 있다.

C GPT-3의 독특한 전이 학습

GPT-3는 2020년 오픈AI에서 개발한 AI 언어 모델로, 기존 언어 모델보다 크기가 100배 이상 되는 거대 언어 모델이다. 1750억 개의 파라미터를 가진 덕분에 요약물 잘하는 것은 물론 Q&A, 번역, 글쓰기 등 언어로 할 수 있는 다양한 기능을 제공해 이전의 GPT-2와 기타 언어 모델들에 비해 월등한 성능을 자랑한다.

다만 그만큼 학습하기 위해 하드웨어에 많은 비용을 투자할 수밖에 없다. 예를 들어 최고의 중식 셰프가 직접 중식당을 차린다고 가정해 보면, 셰프 입장에서서는 수많은 메뉴의 음식을 제공하기 위해 복잡한 레시피를 개발하는 것은 물론 식당 규모를 키우고 종업원 수도 늘려야 한다. 하지만 식당을 이용하는 고객 입장에서서는 편리하고 좋을 것이다. 원하는 음식이 다양하게 있기 때문이다. 당연히 짬뽕만 전문으로 하는 식당보다는 다양한 메뉴가 있는 중식당을 이용할 가능성이 높을 것이고, 짬뽕을 먹으러 갔다가 다양한 메뉴를 보고 다른 음식을 고르거나 추가 주문할 수도 있다.

GPT-3도 마찬가지이다. 사용자가 원하는 기능을 다양하게 제공하기 때문에 이용자 입장에서서는 보다 많은 이점이 있다.



[그림 2] GPT-3의 독특한 사용방식

이러한 GPT-3의 특징 중 하나는 입력하는 데이터의 양에 따라 기초 모델의 학습 능력이 다르다는 것이다. 제로샷 러닝(Zero-shot Learning)은 추가 훈련이나 데이터 없이도 바로 작업을 수행할 수 있고, 원샷 러닝(One-shot Learning)은 한 개의 데이터, 퓨샷 러닝(Few-shot Learning)은 매우 적은 양의 데이터만 가지고 새로운 작업을 수행할 수 있다. GPT-3의 경우 이전의 GPT-2나 다른 언어 모델들에 비해 학습 능력이 탁월하기 때문에 이 가운데 제로샷 러닝만으로 대부분의 다양한 언어 처리 작업을 수행할 수 있다.

반대로 생각하면 제로샷보다는 원샷이, 원샷보다는 퓨샷 러닝이 훨씬 더 정확한 결과를 낼 수 있다는 이야기가 된다. GPT-3의 업그레이드 버전인 ChatGPT에서 질문할 때 예시 답변을 함께 주면 좀 더 양질의 답변을 얻을 수 있는 이유가 바로 이것이다. 그러므로 ChatGPT에 질문할 때는 프롬프트를 상세히 써 주는 것이 좋다. 이로 인해 프롬프트를 전문적으로 써주는 프롬프트 엔지니어라는 직업도 새롭게 등장했다.

GPT-3의 또 다른 특징 중 하나는 자연어 생성에 있어서도 놀라운 성능을 보여준다는 것이다. 이전의 언어 모델들에 비해 문장 구조, 문맥 이해, 문장 간 일관성 등을 더욱 정확하게 파악하고 사용자가 원하는 답변을 정교하게 생성하기 때문이다. 이는 GPT-3가 인간 수준의 자연어 생성 능력에 한 걸음 더 다가갔음을 의미한다. 따라서 다양한 분야에서 GPT-3를 활용할 수 있는 가능성도 매우 높아졌다.

뿐만 아니라 GPT-3는 기존의 전이 학습과 달리 소형의 모델을 만들 필요가 없고 학습시킬 필요도 없다. 단지 기초 모델만 잘 사용해도 요약, 번역, 질의 응답, 분류 등과 같이 다양한 언어 기능을 가뿐히 처리할 수 있다.

한편 이로 인해 사용자는 GPT-3에 더 의존하게 된다. 이에 GPT-3를 만든 오픈AI는 사용에 따른 토큰 수만큼 사용료를 부과하는 수익 모델을 내놓았다. 이는 향후AI 모델 및 산업의 발전에도 큰 영향을 줄 것으로 보인다.

D GPT-3로 인한 변화

이제 GPT-3가 가져온 세 가지의 큰 변화를 살펴보자.

첫째, AI 모델의 대중화가 시작되었다. 지금까지는 AI 모델을 활용하려면 반드시 코딩을 할 줄 알아야 했다. 그러나 GPT-3는 평소에 사용하는 말로 명령만 하면 바로 답변을 하기 때문에 사용성이 매우 높아져 코딩을 몰라도 얼마든지 AI 모델을 활용할 수 있다.

둘째, AI 모델 사용량에 따른 과금제가 시작됐다. 기업 입장에서 보면 AI 모델 개발에 들어가는 어마어마한 비용을 감당하기 위한 수익 구조가 반드시 있어야 한다. GPT-3 이전의 AI 모델들은 과금 자체가 불가능했지만 오픈AI는 GPT-3에 사용료를 부과하면서 기업의 안정성을 더욱 높였다. 이는 더 나은 기술 발전을 위한 투자 비용을 안정적으로 마련한 것과 같다.

셋째, AI 모델의 소스 코드는 더 이상 일반 대중에게 공개되지 않는다. GPT-3가 유료 서비스로 전환되면서 소스 코드도 공개할 수 없게 되었다. 지금까지 GPT-3에 대한 논문은 발표됐어도 소스 코드나 실제 학습 데이터는 공개되지 않았다. 단지 오픈AI에 10억 달러를 투자한 마이크로소프트만이 이를 볼 수 있다.

그렇다면 GPT-3는 파인 튜닝을 할 수 없을까. 그렇지 않다. 오픈AI가 GPT-3를 유료 라이선스 형태로 제공하면서도 전이 학습이 가능하도록 했기 때문이다. 또한 GPT-3 계열의 다른 네 가지 기초 모델(Ada, Babbage, Curie, Davinci)도 제공하고 있다. 이것은 모두 오픈AI API를 활용해 파인 튜닝이 가능하다. 하지만 그 이후의 모델, 즉 GPT-3.5는 자체 API를 사용해서 파인 튜닝을 가능하게 하고 GPT-4는 현재 파인 튜닝이 불가능하도록 막아 놓았다.

E LLM 의 확장

앞서 GPT-3와 같이 거대한 크기의 언어 모델을 LLM로 지칭한다고 설명했다. GPT-3 뿐 아니라 그 이후에 나온 언어 모델, 즉 오픈AI의 GPT-3.5와 GPT-4, 구글의 PaLM(Pathways Language Model) 같은 것들이 모두 LLM이다. 그리고 보통 LLM은 기초 모델이며, 기초 모델을 활용해서 파인 튜닝을 통해 작게 만든 LLM을 sLLM(small Large Language Model)이라고 한다.

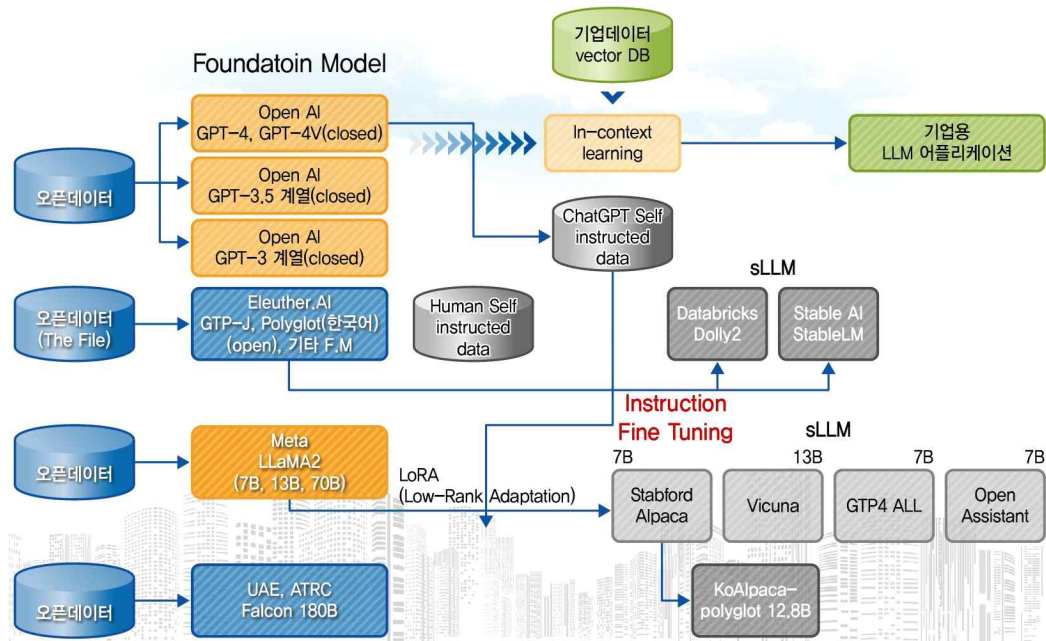
이러한 초대형 언어 모델들의 특성은 질문에 대해서 답을 만든다는 것이다. 이에 ‘만든다’는 의미에서 ‘생성형 AI(Generative AI)’라고 한다. 생성형 AI에는 글을 만들어 주는 LLM뿐 아니라 그림을 만들어 주는 Midjourney, Stable Diffusion, Dalle-3, 음악을 작곡해 주는 AIVA, Amper Music, 동영상을 만들어 주는 Runway 등이 있다.

한편 기존의 LLM에 그림을 인식하고 소리를 인식할 수 있는 기능도 추가되고 있다. 그래서 이를 멀티모달(Multi-modal) LLM이라고 지칭한다. 최근에 나오는 GPT-4V는 멀티모달 LLM 또는 이를 줄여서 LMM(Large Multi-modal Model)이라고 부르기도 한다.

이 외에 앞으로 나올 구글의 Gemini 라는 LLM도 멀티모달 형태로 개발되고 있다. 또한 오픈 소스 진영의 LMM으로 LLaVA(Large Language and Vision Assistant) 도 있다.

F 개발된 LLM 현황

ChatGPT 외에도 현재 다양한 LLM이 나와 있다. 다음 그림은 현재까지 출시된 각종 LLM 및 기업용 애플리케이션을 흐름대로 정리한 것이다.



[그림 3] 현재 LLM의 현황

대표적인 LLM 모델로는 오픈AI가 만든 GPT-3, GPT-3.5, GPT-4가 있다. 일반적으로 ChatGPT는 GPT-3 이후에 나온 LLM들을 통칭한다. 즉 GPT-4, GPT-4V 그리고 앞으로 나올 GPT-5까지 모두 ChatGPT다.

오픈AI는 GPT 모델 하나에 다양한 버전을 출시한다. 여기에 오픈AI API 및 랭체인(Langchain)을 활용하면 기업용 LLM 애플리케이션을 따로 파인 튜닝하지 않고도 원하는 기능을 만들 수 있다. 이러한 방식을 인 콘텍스트 러닝(In-context Learning)이라고 하는데, 프롬프트 튜닝과 유사하게 기업 데이터를 벡터로 바꾼 다음 프롬프트에 넣어 LLM으로 보내는 방식이다. 이렇게 하면 기업이 보유한 데이터로 ChatGPT와 같은 형태의 챗봇을 만들 수 있다. 다만 이 방식은 주로 기업 데이터 양이 적을 때 적용 가능하다.

기업 데이터 양이 매우 많다면 Dolly2, StableLM, Alpaca, KoAlpaca, Vicuna 등의 LLM을 바탕으로 파인 튜닝을 할 수 있다. 이때 학습 비용을 줄이기 위해 LoRA(Low Rate Adaption)라는 기법을 사용하는데, 파인 튜닝할 때 파라미터 계산을 대폭 줄여 좀 더 적은 GPU를 사용하는 효과가 있다.

오픈AI 이외의 LLM은 크게 두 가지로 구분할 수 있다. 하나는 원천이 되는 기초 모델을 메타의 LLaMA(Large Language Model Meta AI)로 만든 것이고 다른 하나는 비영리 AI 연구 기관인 일루서AI(ElleutherAI)에서 만든 것으로 GPT-J, GPT-Neo 등이다.

일루서AI는 2020년 오픈AI가 GPT-3를 출시하면서 소스 코드를 공개하지 않았을 때 오픈 소스를 표방하면서 생긴 연구소다. 주로 자원 봉사자들로 구성된 연구원, 엔지니어 및 개발자의 풀뿌리 집단이다. 이 연구소는 연구 결과를 학술지에 게시하고 아이디어와 방법론을 블로그에 게시하며 학습된 모델을 무료로 제공한다.

이러한 일루서AI의 뛰어난 성과 중 하나가 바로 GPT-J 모델이다. GPT-J는 6B 파라미터의 LLM으로, GPT-3와 유사한 성능을 보이는 오픈 소스 모델이다. 한국어 데이터셋으로 학습된 PolyGlut-Ko도 있는데 1.3B, 3.8B, 5.8B, 12.8B 등 네 가지 형태로 구성되어 있다. 또한 이는 한국어로 된 LLM 소스가 공개되었다는 점에서 특별한 의미를 갖는다. 특히 최근에 공개된 12.8B 버전 소스 코드는 공식 사이트(<https://huggingface.co/beomi/KoAlpaca-Polyglut-12.8B>)에서 받을 수 있다.

이 외에도 일루서AI의 눈에 띄는 성과 중 하나는 ‘더 파일(The Pile)’이라고 불리는 오픈 데이터 세트다. 이는 기초 모델을 만들 때 사용되며 완전히 공개된 상태이기 때문에 사용에 제한이 없다. 현재 많은 벤처회사들이 GPT-J를 근간으로 LLM 제품을 만들고 있다. 데이터브릭스의 Dolly나 스테이블AI의 StableLM 역시 LLaMA의 상업적 사용 제한을 극복하기 위해 일루서AI의 GPT-J를 사용했다.

LLM의 발전에 있어 또 한 가지 중요한 부분이 자가 학습 데이터(Self-Instructed Data)이다. 이는 스탠퍼드대학에서 공개한 Alpaca에서 사용한 방법으로, GPT-4처럼 발달된 LLM에 질문한 다음 그 답변을 모은 데이터세트다. 이를 이용하면 빠르게 데이터세트를 구축할 수 있을 뿐만 아니라 사람이 직접 질문을 평가하거나 수정할 필요가 없기 때문에 인건비를 상당히 줄일 수 있다. 한편 사람이 직접 자가 학습 데이터를 만드는 경우도 있는데, Dolly나 StableLM이 그 예다.

2

ChatGPT 유료화가 시장에 미치는 영향

ChatGPT(GPT-3.5)가 2022년 11월 30일 발표된 후 생성형 AI는 우리 삶에 엄청난 영향을 미치고 있다. 많은 사람들이 체감하고 있듯이 ChatGPT는 기술, 학문, 교육, 문화, 사회, 서비스, 기업 문화 등 사회 전반을 조금씩 바꾸고 있다. 특히 유료 버전인 GPT-4의 경우 상당히 많은 기능들이 있다. 마이크로소프트의 검색 엔진 Bing을 참조하거나 ADA(Advanced Data Analysis)를 통해 데이터를 분석하고 Dalle-3를 활용해 이미지를 생성할 수도 있다. 이처럼 유료화된 ChatGPT가 세상을 어떻게 바꾸고 있는지 구체적으로 살펴보자.

A 유료 서비스로 인한 AI 발전 가속화

예전에는 AI 모델들을 대부분 무료로 사용할 수 있었다. AI 전문가 또는 AI 기업들이 AI 모델을 개발한 후 결과물을 모두 깃허브(Github)에 올려서 누구나 소스 코드를 볼 수 있게 했기 때문이다. 그런데 오픈AI의 GPT-4는 다르다. 오픈AI는 ChatGPT(GPT-3.5)의 엄청난 성공 이후 발표한 GPT-4를 유료화했다. 이를 통해 오픈AI가 창출한 매출은 2023년 2억 달러, 2024년 약 10억 달러로 예상하고 있다. 게다가 이미 ChatGPT가 제공하는 언어의 수는 95개에 달해 조만간 거의 모든 나라에서 ChatGPT를 사용하게 될 것으로 보인다.

사람들이 유료 ChatGPT 서비스에 가입하는 이유는 다음과 같다.

- 유료 사용자는 사용량이 많은 시간대에도 서비스를 우선적으로 사용할 수 있는 권한을 갖는다.
- 유료 사용자는 새로운 기능 및 개선 사항을 반영한 모델을 우선적으로 사용할 수 있다.
- 유료 버전에서는 플러그인, 실시간 검색 기능, ADA, DALL-E-3를 사용할 수 있다.
- 출시된 되었으나 서비스는 하지 않고 있는 멀티모달이 가능한 GPT-4V도 유료 버전에서만 사용할 수 있다.

여기서 플러그인은 특정 기능을 실행할 수 있는 일종의 확장 프로그램이다. ChatGPT 플러그인은 최신 정보에 취약한 ChatGPT의 약점을 보완하고 연결된 다른 프로그램에서 사용자가 할 일을 대신 수행한다. GPT-4에는 2023년 10월 현재 1000여 개의 플러그인이 있으며, 실시간 검색은 마이크로소프트 Bing을 통해 가능하다.

ChatGPT는 기업에서도 많이 사용하고 있다. 유료 버전의 사용료는 월 20달러로 기업 입장에서는 비용 부담이 크지 않다. 이런 이유로 앞으로 유료 서비스 가입자 수는 상당히 늘어날 것으로 예상된다.

물론 유료 버전의 ChatGPT가 아니더라도 무료로 사용할 수 있는 마이크로소프트의 Bing Chat이나 구글 Bard 등에 많은 사람들이 관심을 갖고 있고, 우리나라에서는 카카오톡 기반의 야속업(Askup), 뽀빠(wrtn.ai) 등도 사용할 수도 있다. 하지만 인터넷 실시간 검색과 플러그인 기능이 제공되는 ChatGPT는 Bing Chat이나 Bard보다 답변의 깊이나 내용이 훌륭하기 때문에, 수요 측면에서 앞으로 AI 기술의 유료화는 계속될 것으로 전망된다.

B 유료화의 다른 이름은 글로벌 경쟁

실제로 ChatGPT 유료 서비스, 즉 유료 AI 챗봇 서비스는 모든 기업이 눈독 들이는 비즈니스 모델이다. 적정한 시스템과 콘텐츠를 보유하고 있으면 고객이 알아서 가입하고 사용료를 매달 지불하는 방식이기 때문이다. 기업 입장에서는 OTT와 같은 플랫폼 비즈니스보다 오히려 더 좋은 비즈니스 모델이라 할 수 있다. 플랫폼 비즈니스는 콘텐츠를 소비할 고객과 서비스를 제공할 콘텐츠 업체를 모으는 데 상당한 마케팅 비용이 들지만 ChatGPT 유료 서비스는 시스템만 안정적으로 확장하면 큰 비용이 들지 않는다.

이는 게임 분야와 비슷한 비즈니스 모델이라 할 수 있다. 즉 서비스를 만들 때는 많은 투자 비용이 들더라도 일단 서비스를 개시하면 유료 회원이 폭주하는 구조다. 물론 현재는 초기 단계로 서비스를 제공할 때 많은 GPU가 필요하고 GPU 가격이 폭등하면서 서비스를 제공하는 비용이 증가하고는 있지만, 향후 GPU 분야의 경쟁 체제가 성립되면 가격은 자연적으로 안정화될 것이다.

당연히 많은 글로벌 기업들이 이런 꿈같은 비즈니스 모델에 참여하고 싶어한다. 그러나 여기에는 캐즘(Chasm)이 존재한다. 이는 비즈니스가 일정 단계를 지나면 빠지기 쉽고 뛰어넘기도 어려운 죽음의 절벽과도 같다. 2020년 우리나라에서 벌어진 이루다 사태나 하루 만에 서비스를 접은 챗봇 테이(Tay)의 경우가 캐즘을 넘지 못한 대표적 사례다. 아무리 거대 글로벌 기업이라고 해도 서비스 초기에 치명적인 결함이 생기면 하루아침에 나락으로 떨어질 수 있는 것이다.

지금도 사람들은 ChatGPT를 향해 비판을 한다. 틀린 답을 너무 많이 제시하기 때문이다. 이것을 환각(Hallucination)이라고 하는데 우리나라 사용자들은 ‘아무 말 대잔치’라고 비아냥거리기도 한다.

GPT-4에서 많이 개선되긴 했지만 오답 논란은 앞으로도 계속될 것이다. 오픈AI는 GPT-5에서 이런 환각 현상이 상당 부분 해결될 것이라고 장담하지만, 사용자들은 내용의 정확성 검증을 위해 여전히 구글링을 계속해야 할지 모른다.

그럼에도 불구하고 오픈AI는 성공적으로 ChatGPT를 발표했고, 실시간 검색 기능이나 ChatGPT 플러그인, ADA, DALL-E-3 등을 계속 추가하고 있다. 현재의 흐름이라면 앞으로 ChatGPT에 대항할 만한 서비스나 비즈니스 모델이 얼마나 나올지는 사실 의문이다. ChatGPT의 유료 서비스 수입은 오픈AI의 서비스 완성도를 계속 높여 경쟁력을 강화하는 결과를 가져올 것이고, 이는 결국 오픈AI가 최후의 승자가 될 가능성이 높음을 의미한다.

ChatGPT의 유료 서비스로 인한 자본의 유입은 오픈AI가 구글과의 경쟁에서 유리한 위치를 선점하는 것은 물론 서비스의 품질을 높이는 데 기여할 것이다. 이런 측면에서 고도의 기술과 막대한 비용을 투자해야 하는 AI 산업에서는 글로벌한 몇 개의 서비스나 기업만 살아남을 가능성이 높다. 결국 승자만이 전 세계 시장을 차지할 것이다.

C 학문적인 AI에서 상업적인 AI로

얼마 전까지는 대학에서 새로운 AI 모델이나 기술을 개발하면 일반 기업에서 그것을 응용해 사용했다. 저명한 AI 리더라고 하면 대부분 대학교수였다. 스탠퍼드, 카네기멜론, MIT, 토론토, 옥스퍼드, 카이스트 등 전 세계 유수 대학들이 앞장서서 AI 분야를 이끌어 왔다. 국내외 AI 학회에서 새로운 논문을 발표하는 것은 연구자로서 영예로운 업적 중 하나였고, 논문을 발표하면 연구에 사용된 소스 코드도 함께 공개하는 것이 미덕이었다.

그런데 AI 모델이 점점 거대해지고 이를 학습시킬 GPU도 증가함에 따라 AI 연구에는 이전보다 몇 배의 비용이 들게 되었다. GPT-3를 학습하는 데만 최소 수백억 원의 비용이 들어갔다고 하니 GPT-3.5나 GPT-4는 그 이상의 비용이 들었을 것이다.

이런 상황이 되니 대학에서는 새로운 AI 모델을 개발하는 데 발생하는 비용을 감당할 수가 없는 상황이다. 실제로 GPT-3가 출시된 2020년 이후 발표된 새로운 AI 모델을 살펴보면 모두 대학이 아닌 오픈AI, 구글, 메타 등 세계적인 기업의 연구소에서 개발한 것이다. 이제 새로운 AI 모델은 굳이 유명 AI 학회를 거치지 않고도 논문 형태가 아닌 기술 문서(Technical Report) 형식으로 발표되고 있다.

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

[그림 4] 오픈AI의 Technical Report

오픈AI는 자신들이 개발한 소스 코드를 공개하지 않다. 따라서 2023년 3월 14일에 발표한 GPT-4의 기술 문서에는 중요한 기술적 내용이 빠져 있다. 일각에서 이제 오픈AI는 더 이상 ‘오픈(Open)’이 아니라고 이야기하는 이유다.

구글도 이에 오픈AI에 대항하기 위해 2023년 5월 10일 차세대 언어 모델 PaLM2를 발표했다. 이를 지메일, 구글 독스 등 자사 25개 제품에 적용하고 개발자들을 위해 PaLM API도 제공하고 있다. 그리고 구글 역시 PaLM2의 파라미터 개수나 구체적인 기술 정보를 공개하지 않았다. 오픈AI와 구글 모두 소스 코드를 공개하지 않는 폐쇄형 AI 개발을 추진하고 있는 것이다.

이들 기업과 달리 메타에서는 ChatGPT와 유사한 기능을 갖는 LLaMA2를 개발한 후 이를 근간으로 누구나 새로운 상용화 제품을 만들 수 있도록 했다. 실제로 메타를 중심으로 한 많은 개발업체들이 LLaMA2를 파인 튜닝해서 서비스화하고 있다. 즉 오픈 소스 운동이 LLM에서 끊어진 것은 아니다.

그러나 AI가 본격적으로 유료화되면 소스 코드는 더 이상 공개되지 않고 AI의 주도권이 대학에서 전문기업으로 더 확실히 넘어갈 것이다. 또한 AI의 새로운 이론보다는 상용화할 수 있는 서비스를

확장하거나 기업을 대상으로 서비스 제공 범위를 다각화하는 방향으로 변화할 것이다. 이렇게 되면 전 세계를 대상으로 AI 발전을 주도하는 몇 개의 기업만이 살아남고, 나머지 기업들은 사용자나 거대 테크기업의 협력사로 남을 가능성이 높다. 이것은 우리나라의 영세한 AI 기업의 미래와도 연결되는 문제다. 대한민국 AI의 미래를 위해 하루라도 빨리 대응책을 마련하는 것이 중요하다.

D AI 연구자에서 AI 사용자로

지금까지 'AI 연구자'라는 영역은 있어도 'AI 사용자'라는 영역은 없었다. 대학에서는 얼마 전만 해도 이전에 없던 이론과 AI 모델을 만들기 위해 AI 연구자들의 역할이 중요했고 기업에서도 이들은 AI 도입을 주도하는 데 꼭 필요한 인력들이었다. 그런데 AI의 헤게모니가 초거대 AI를 만드는 기업 연구소로 옮겨지며 고학력 AI 연구자보다는 초거대 AI를 제대로 구현하는 인력이 더 중요해졌다. 또한 ChatGPT가 코딩도 할 수 있으니 이를 잘 사용할 줄 아는 AI 사용자를 더 선호하게 되었다.

앞으로는 이렇게 ChatGPT와 같은 도구를 잘 활용할 줄 아는 프롬프트 엔지니어 및 컨설턴트의 수요가 더 늘어날 것이다. 이러한 상황에서 대학들은 AI 연구자보다는 AI를 잘 활용해 기업이 원하는 결과물을 만들어 낼 수 있는 AI 사용자 배출에 힘써야 한다.

예를 들어 건설기업에서 앞으로 가장 필요한 AI 관련 인력은 건설업을 이해하고 그에 필요한 AI를 만들어 활용할 수 있는 사람이다. 반면 AI는 아는데 건설업을 모른다면 앞으로 건설 업계에서는 일하기 힘들어질 것이다. 이러한 현상은 곧 모든 산업에 적용될 것이다. 따라서 AI에 직접 명령하는 프롬프트 엔지니어와 각 분야의 기업 컨설턴트는 해당 산업을 먼저 이해한 다음 그 지식을 바탕으로 AI를 활용할 수 있어야 한다.

이제 AI는 무조건 고부가 가치 산업이라는 생각을 버려야 한다. AI 깃발만 내걸면 투자가 쏟아지던 시대는 지났다. 구직자들 또한 AI만 알면 어디든 취직할 수 있다는 생각을 버려야 한다.

앞으로는 기업에서 AI 투자 비용에 대한 손익 계산서를 뽑을 것이다. 단지 AI를 도입하는 것만이 능사가 아니라 이로 인해 기업의 수익을 얼마나 창출했는지, 고객은 만족하는지, 비용은 얼마나 절감했는지 등 실질적인 혜택을 모두 따져 봐야 하는 시대가 되었다.

E 보편적 서비스에서 산업별 서비스로

현재 오픈AI가 LLM 산업을 주도하고 있다고 해도 세상의 모든 대화형 AI 서비스를 점유할 수는 없을 것이다. 아무리 많은 정보를 학습했다 해도 모든 분야의 문서를 전부 습득할 수는 없기 때문이다. 이런 측면에서 앞으로는 오픈AI와 같은 글로벌 기업이 지향하는 보편적 서비스 외에 각 산업별 특성에 따라 세분화된 도메인별 서비스가 등장할 것으로 예상된다.

보편적 서비스가 글로벌 대기업의 몫이라면, 도메인별 서비스는 각 산업별 노하우를 집약한 서비스라고 볼 수 있다. 이미 여러 업체에서 산업별로 이러한 서비스를 제공하고 있다.

다음은 ChatGPT에 현존하는 산업별로 정보를 서비스하는 기업이 어떤 것들이 있는지 물어본 데 대한 답변이다.



산업별 정보 서비스 회사를 산업의 매출 규모 순서로 10개 정도 나열해줘. 산업별 정보 서비스 회사란 예를 들면 금융 산업의 FactSet 같은 회사를 말해.



산업별 정보 서비스 회사는 특정 산업에 대한 데이터와 분석, 연구, 소프트웨어 등을 제공하는 업체입니다. 아래는 다양한 산업에서 활동하는 대표적인 정보 서비스 회사를 해당 산업의 매출 규모를 기준으로 나열한 것입니다. 매출 규모는 계속 변동하는 수치이므로 아래 목록은 대략적인 지표로 봐주시면 좋겠습니다.

1. 기술 산업: Gartner
 - 세계적으로 IT 분야의 리서치와 컨설팅을 제공하며, 기업들이 기술 투자를 최적화할 수 있도록 지원합니다.
2. 금융 산업: Bloomberg
 - 금융 데이터, 분석, 뉴스 등을 제공하며, Bloomberg Terminal은 세계 금융 시장에서 표준적으로 사용되고 있습니다.
3. 건강 산업: IQVIA
 - 제약, 바이오테크, 의료 기기 산업에 대한 데이터와 분석 서비스를 제공합니다.
4. 에너지 산업: Platts
 - 에너지와 금속 분야의 정보, 가격, 분석 서비스를 제공합니다.
5. 자동차 산업: IHS Markit
 - 자동차 산업에 대한 시장 연구, 데이터, 분석 등을 제공합니다.
6. 소비자 산업: Nielsen
 - 소비자 산업에 대한 시장 연구와 분석을 제공하며, 특히 미디어와 소비 패턴에 대한 데이터가 유명합니다.
7. 부동산 산업: CoStar Group
 - 상업용 부동산에 대한 정보와 분석 서비스를 제공합니다.
8. 교육 산업: Eduventures
 - 교육 산업에 대한 시장 연구와 컨설팅 서비스를 제공합니다.
9. 미디어 및 엔터테인먼트: Comscore
 - 미디어 산업에 대한 소비자 행동 분석과 측정 서비스를 제공합니다.
10. 광고 및 마케팅: Kantar
 - 광고, 브랜드, 마케팅 등에 대한 연구와 분석 서비스를 제공합니다.

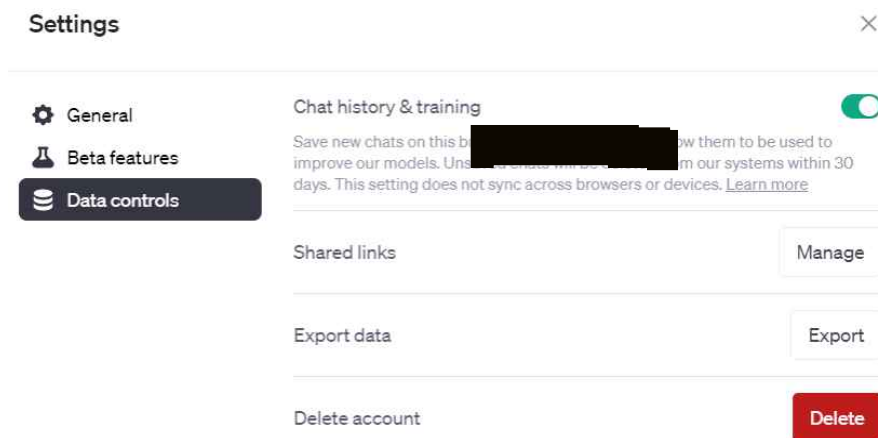
각 회사가 제공하는 서비스의 세부 내용, 특징점, 시장 점유율 등은 다양하므로, 특정 산업에 대한 더 깊은 이해를 위해 추가적인 연구가 필요할 수 있습니다.

[그림 5] GPT-4의 답변

여기에 언급된 기업들은 앞으로 산업별로 특화된 ChatGPT를 만들 가능성이 높다. 이미 수십 년에 걸쳐 축적된 데이터를 정제해 유료로 정보를 제공하고 있기 때문이다. 앞서 한 질문은 가장 큰 규모의 사업 위주로 답하도록 유도한 것이지만, 여기에 언급되지 않은 거의 모든 분야에서 이러한 정보 서비스를 운영하고 있을 것이다.

F 일반적인 용도에서 기업에 특화된 LLM으로

ChatGPT에 입력한 질문은 고스란히 오픈AI의 데이터로 쌓이는데 이에 대해 많은 기업에서 우려를 표하고 있다. 기업의 기밀 등 중요한 내용이 입력되면 이것이 다시 ChatGPT를 학습하는 데 사용될 수 있고, 이를 학습한 ChatGPT의 답변을 통해 그 내용이 만천하에 공개될 위험이 있기 때문이다.



[그림 6] ChatGPT에서 “Chat history & training”을 해제하면 입력된 데이터를 학습에 사용하지 않음

물론 입력한 프롬프트 내용을 ChatGPT 학습에 활용하고 싶지 않다면 설정에서 ‘Chat History & Training’ 옵션을 해제하면 된다. 이 경우 그동안 입력된 대화 히스토리가 보이지 않게 되고 입력된 문장은 학습에 사용되지 않는다. 하지만 히스토리가 안 보이게 되면 사용하는 데 매우 불편하고, 이 설정을 잘 모르는 사람들도 아직 많다는 것이 문제다.

물론 ChatGPT 활용의 긍정적인 측면도 많다. 예를 들어 각 기업에는 오랜 시간에 걸쳐 쌓은 기록이나 노하우를 담은 문서들이 있고 이는 주로 기업의 지식 관리 시스템(Knowledge Management System)에 들어 있다. 그런데 이 시스템에서 원하는 정보를 찾으려면 검색을 수없이 반복해야 한다. 만약 그 문서들을 학습시켜 기업용 ChatGPT를 만들면 직원들이 훨씬 편리해질 것이다.

이렇게 각 기업에 특화된 ChatGPT를 만들면 그 자체로 기업의 중요한 자산이 된다. 신입 사원도 자신의 업무에 관한 정제된 답변을 받아 빠르게 업무에 적응할 수 있는 것이다. 이러한 장점 때문에 앞으로는 기업용 ChatGPT를 만들어 주는 새로운 비즈니스의 시대가 열릴 것으로 전망된다.

G 주니어 여러 명 대신 ChatGPT를 잘 쓰는 시니어 한 명

ChatGPT는 기업의 일하는 방식을 근본적으로 바꿀 것이다. 특히 전문 지식을 갖춘 ChatGPT를 잘 활용하면 늘 하던 업무가 매우 수월해지고 업무량도 상당히 줄어들 것이다. 자신이 하는 일에 대한 전반적인 이해를 바탕으로 그 일을 세분화한 다음 특정 부분을 ChatGPT에 시키면 되기 때문이다.

예전에는 많은 기업들이 야근도 불사하며 열심히 일하는 사람에게 높은 평가를 주곤 했다. 하지만 ‘머리가 나쁘면 몸이 고생한다’라는 말처럼 우리나라 기업 문화의 근간이던 ‘농경 시대의 근면성’은 이제 옛말이다. 상황에 따라 맡은 일의 일부는 ChatGPT에 맡기고 사람은 ChatGPT가 일한 결과를 보고 빠진 것이 있는지, 오류는 없는지 등의 문제만 해결하면 된다.

앞으로 ChatGPT를 잘 활용하는 사람은 큰 규모의 일도 쉽고 빠르게 처리할 수 있는 반면, ChatGPT를 활용하지 않거나 사용할 줄 모르는 ChatGPT에 의해 쉽게 대체될 것이다. 중국에는 ChatGPT를 잘 활용하는 소수의 엘리트 직원 한 명이 ChatGPT를 활용하지 못하는 열 명을 대신해 전체 생산성을 주도할 수도 있다.

사람은 보통 특정 분야만 잘 안다는 한계가 있지만 기업은 모든 문제를 한번에 해결할 수 있는 슈퍼 능력자를 원한다. 그러나 그런 사람은 찾기 힘들다. 그러나 이제는 달라질지 모른다. 지금까지 전혀 몰랐던 분야라도 ChatGPT에게 물어보면 단번에 답변을 해주기에 약간의 관련 상식과 하고자 하는 열정만 있으면 누구나 능력자가 될 수 있다.

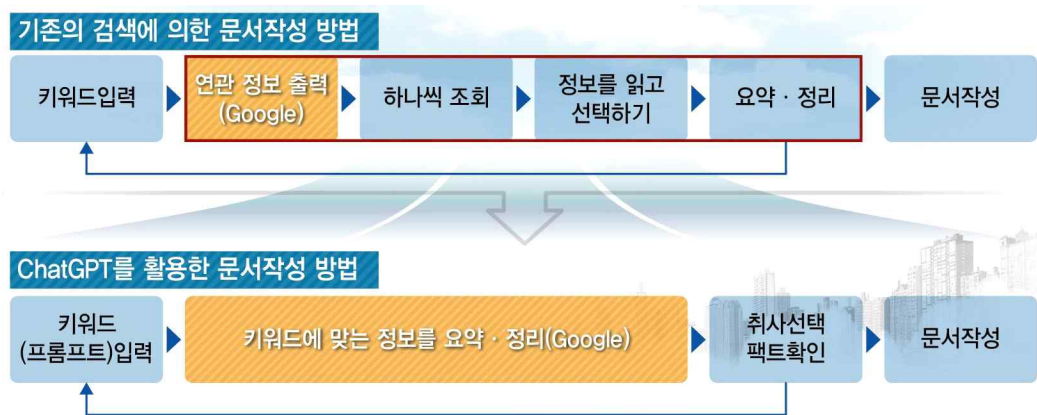
따라서 기업에서는 처음부터 하나하나 일을 가르쳐야 하는 주니어보다 축적된 경험을 바탕으로 ChatGPT를 잘 사용해 빠른 성과를 내는 시니어를 더 찾게 될 것이다.

3

검색 시장의 변화를 비롯한
ChatGPT 파급 효과 분석

A 일하는 방식의 변화

ChatGPT가 세상에 나오자 사람들의 일하는 방식이 많이 달라졌다. 기존에는 먼저 일의 계획을 세우고, 문서를 작성하고, 보고 및 회의하는 것이 순서였다. 그리고 이를 위해 관련 자료를 검색하는 데 많은 시간을 소비했다. 그런데 ChatGPT가 등장하면서 전 세계적으로 검색량이 눈에 띄게 줄어든 것으로 나타났다.



[그림 7] 기존의 검색에 의한 문서 작성 방법과 ChatGPT를 활용한 문서 작성 비교

보고서 작성을 예로 들어 보자. 기존에는 어떤 문서를 작성하려면 포털 사이트에 키워드를 입력한 후 수많은 검색 결과를 하나씩 클릭하면서 관련 정보를 읽고 그 내용을 요약 정리해 문서를 완성하곤 했다. 하지만 이제 ChatGPT는 알아서 그 수많은 정보를 읽은 다음 필요한 정보 몇 가지만 정리해서 보여준다. 물론 그 정보가 사실인지 아닌지 확인하는 과정에서 또 다시 포털 검색을 해야 하지만 이전보다는 훨씬 시간과 노력을 덜게 되었다.

세상의 변화는 이처럼 일하는 방식의 변화부터 시작된다. 이러한 일하는 방식의 변화는 기존 검색 시장에서 정확도와 신속성을 무기로 압도적 우위에 있던 구글에도 치명적인 타격이 될 것이다.

앞서 구글에서 출시한 PaLM2 LLM 기반의 챗봇 Bard는 실시간 검색 기능을 강점으로 내세웠다.

하지만 GPT-4에도 실시간 검색 기능이 추가됨에 따라 큰 차이점이 없어졌다. 심지어 추론 능력이나 답변의 품질은 GPT-4가 더 월등하다.

또한 현재 구글은 문장뿐 아니라 음성과 사진까지 이해하는 완전히 새로운 멀티모달 LLM인 Gemini를 준비 중이지만, 오픈AI는 이미 음성과 사진을 이해하는 GPT-4V를 출시했고 텍스트를 사진으로 만들어 주는 DALL-E-3도 사용자 화면에 구현 중이다. 결국 구글은 오픈AI에 대응하기 위해 ChatGPT의 최대 단점인 정보 신뢰도 문제를 넘어설 수 있도록 검색 결과 신뢰도를 기존 보다 높이는 전략이 필요할 것이다.

B 생성형 AI를 활용한 사용자 인터페이스의 전환

사용자 인터페이스, 즉 UI(User Interface)는 사용자들이 웹사이트나 애플리케이션을 이용할 때 만나는 화면을 말한다. 이는 사용자가 쉽고 편리하게 사용하는 데 즐거운 경험을 제공하는 화면의 모든 구성 요소를 가리킨다. 여기에서 더 나아가 사용자가 최고의 경험을 할 수 있도록 한 차원 더 높은 것을 사용자 경험, 즉 UX(User Experience)라고 지칭한다.

UI와 UX는 모두 사용자의 경험을 업그레이드하는 것이 목표다. 따라서 사용자들이 어떤 화면을 열면 UI/UX를 파악하는 과정이 선행된다. 아무리 UI/UX가 발전해도 사용자들이 이를 이해하려는 노력을 먼저 해야 하는 것이다.

하지만 세상에 존재하는 모든 웹사이트나 애플리케이션의 사용법을 누구나 바로 알 수는 없다. 때문에 사용자들은 포털 사이트에서 원하는 키워드로 간편하게 검색하는 방식을 주로 이용한다. 여기서 문제는 원하는 결과를 얻을 때까지 검색을 반복하는 과정은 다소 번거롭다는 것이다.

이런 측면에서 생성형 AI를 애플리케이션에 도입하는 사례가 주목받고 있다. 대표적으로 클라우드 SaaS(Software-as-a-Service)로 유명한 세일즈포스(Salesforce)의 EinsteinGPT라는 LLM이 있다. 세일즈포스는 자사의 CRM 패키지에 이를 도입해서 전체 애플리케이션을 쉽게 사용하는 데 활용하고 있다.

한편 공공기관에서는 대국민 서비스를 위해서 많은 포털을 운영하고 있다. 이러한 포털들에는 각기 어마어마하게 많은 정보와 사용 기능이 있어서 원하는 국민들이 서비스를 받으려면 많은 단계를 거쳐야 한다. 또한 원하는 정보가 있는지 검색을 해야 하고 그 검색의 결과는 너무도 많다.

앞으로 이러한 정보의 복잡성은 더욱 증가할 것이다. 따라서 공공 포털을 운영하고 있는 관리자의 입장에서는 국민들이 쉽고 빠르게 서비스를 활용하는 방안을 찾는 것이 매우 중요하다. 예를 들어 공공 포털의 첫 번째 페이지를 중요한 이벤트가 있을 때마다 실제 사용자 맞춤형 메뉴나 안내로 바뀌어서 서비스를 하는 것이다. 연말이 되면 모든 국민이 해야 하는 세금 보고 때나 3월 말 기업의 결산 보고 등에 이를 적용할 수 있을 것이다.

이렇게 대국민 서비스를 위한 공공 포털에 LLM 챗봇을 도입해서 국민들이 원하는 서비스를 바로 제공해 줄 수 있다면 대국민 서비스에 대한 만족도를 획기적으로 높일 수 있을 것이다.

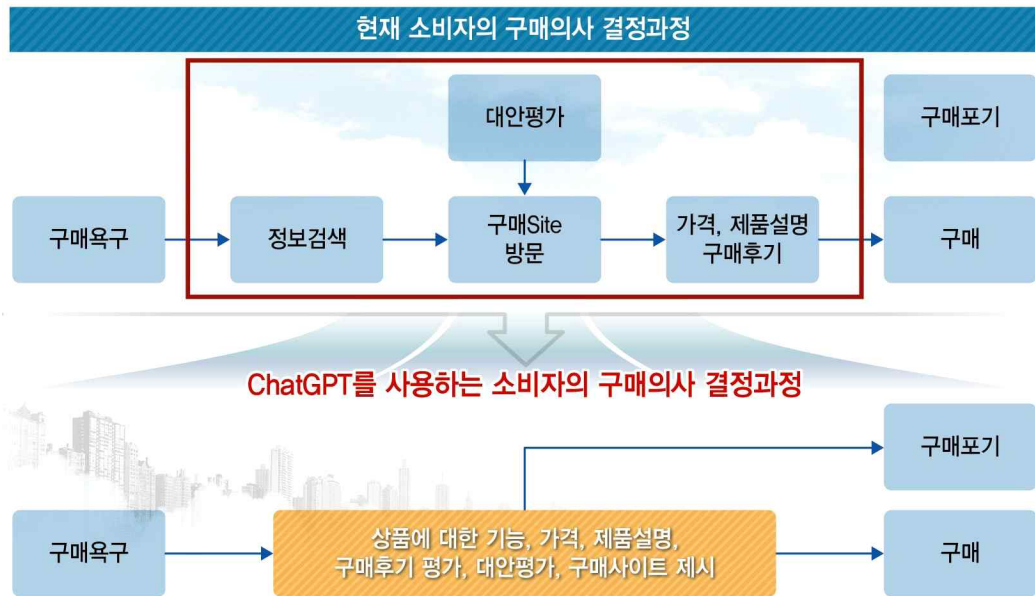
C 소비자의 구매 의사 결정 과정 중 ChatGPT의 역할

현대 소비 사회에서는 소비자가 어떤 과정으로 상품을 구매하는지가 중요하다. 이것은 유통 산업뿐 아니라 제조 산업 전반을 좌우하기에 더욱 중요하다.

인터넷이 없던 시절, 사람들이 원하는 물건을 사려면 발품을 많이 팔아야 했다. 기업 입장에서는 여디가 저렴하다라는 입소문과 소비자가 방문하기 용이한 교통이 좋은 곳에 입점하는 것이 중요했다. 그런데 인터넷이 등장하면서 온라인으로 가격 비교가 가능해지자 소비자들은 더 이상 발품을 팔 필요가 없어졌다. 이것은 소비자의 구매 의사결정 과정이 근본적으로 변했음을 의미한다.

유통업체는 유통 매장이 필요 없게 되었지만 대신 가격 비교에서 조금이라도 우세를 점하기 위해 제조업체에 가격 인하를 요구한다. 제조업체는 좀 더 싼 가격으로 제품을 만들기 위해서 다양한 방식으로 원가를 낮추어야 한다. 이처럼 소비자의 구매 패턴의 변화는 유통업체뿐 아니라 제조업체들에게 커다란 영향을 주고 있다.

이런 가운데 ChatGPT와 같은 다양한 LLM이 출현하면서 소비자의 구매 의사 결정 과정을 다시 한번 흔들고 있다.



[그림 8] 구매의사결정과정의 변화

예를 들어 일주일에 한두 번 러닝을 하는 사람이 러닝화를 구매하려면 포털에서 러닝화 정보를 검색하고 수많은 결과를 보면서 고민해야 한다. 동시에 다른 대안을 떠올리기도 한다. ‘꼭 러닝화를 사야 하나?’, ‘지금 가지고 있는 신발을 신고 뛰어도 괜찮지 않을까?’라는 생각 등이다. 그러다 어떤 러닝화가 마음에 들어 상품 페이지를 방문한다고 해도 곧바로 구매를 결정하지는 않는다. 가격뿐 아니라 기능이나 소재, 용도 등 여러 가지 옵션을 좀 더 검토한 후에야 결정을 내린다. 이처럼 소비자의 구매 의사 결정 과정은 생각보다 간단하지 않다.

그런데 ChatGPT를 활용하면 러닝화 정보를 검색하는 것부터 구매까지의 수고를 상당히 덜어 준다. 러닝화를 사고 싶다는 생각이 들면 ChatGPT가 관련 정보를 알아서 탐색하도록 하면 되기 때문이다. 즉 평소에 자신의 행동 패턴을 바탕으로 ChatGPT에 다음과 같이 물어보면 된다.

“일주일에 1~2회 5km 정도씩 달리는데, 어떤 러닝화를 신는 것이 좋은가?”



일주일에 1~2회 5km 정도씩 달리는데, 어떤 러닝화를 신는 것이 좋은가?



Finished browsing



여러 출처에서 얻은 정보를 바탕으로 일주일에 1~2회 5km 정도 달리는 활동에 적합한 러닝화에 대한 권장사항을 제공할 것입니다.

1. 일반 러닝화 추천:

- **adidas 4D FWD 2 Shoes:** 이 모델은 2023년의 최고 러닝화로 평가받았으며, 남녀 공용으로 제공됩니다¹.
- **Nike ZoomX Invincible Run Flyknit:** 이 러닝화도 남녀 공용으로 제공되며, Nike의 러닝화 중에서 최고로 평가받았습니다¹.

2. 5K 러닝에 특화된 러닝화:

- **Nike ZoomX Vaporfly Next% 3:** 이 러닝화는 경쟁을 즐기는 러너들을 위해 제작되었으며, 더 가볍고 안정적이며 더욱 빠른 반응성을 제공한다고 합니다².
- **adidas Adizero Takumi Sen 9:** 이 러닝화는 경주용으로 제작되었으며, 매우 가벼운 LIGHTSTRIKE PRO 폼과 탄소가 섞인 ENERGYRODS가 더 빠른 발바닥을 도와줍니다².
- **ASICS Metaspeed Sky Plus:** 이 러닝화는 빠르게 시작하고 더 빠르게 끝내려는 러너들을 위해 설계되었으며, 각 스텝마다 에너지를 절약하면서 더 긴 보폭을 유지할 수 있습니다².

3. 개인적인 요인 고려:

- 러닝화를 선택할 때 여러 개인적인 요인을 고려해야 합니다. 예를 들어, 당신의 고유한 생체 역학, 몸무게, 달리는 표면, 발의 형태 및 선호하는 느낌 등이 있습니다³.

이러한 권장사항과 함께 당신의 러닝 습관과 개인적인 편안함을 고려하여 러닝화를 선택하시면 좋을 것입니다. 또한 가능하다면 전문 러닝화 상점에서 발의 측정과 러닝화의 시착을 해보는 것도 좋은 방법입니다.

[그림 9] 러닝화에 대한 GPT-4의 답변

이때 LLM에 따라서는 실시간 가격 검색까지 가능하다. 실제로 네이버가 개발한 하이퍼클로버의 ‘Cue:’는 가격을 제시해 주고 네이버 쇼핑몰로 바로 연결도 해 준다. 사실 네이버가 하이퍼클로버를 개발한 궁극적인 목적도 네이버 쇼핑몰을 활성화하는 것인지 모른다.

Cue:

답변 과정 ^



러닝화는 개인의 발 모양, 러닝 스타일, 운동 목적 등에 따라 다르기 때문에, 정확한 추천을 위해서는 개인의 발 모양과 러닝 스타일에 대한 정보가 필요합니다.

Q '러닝화 추천' 검색 중

Q '러닝화 리뷰' 검색 중

Q '러닝화 사이즈' 검색 중

≡ 검색 결과 분석 중

✓ 답변 준비 완료

일주일에 1~2회 5km 정도씩 달리시는 분들을 위한 러닝화 제품을 추천해드리겠습니다.

1. 국내매장판 나이키 에어맥스 90 97 SC LTR 에센셜 남자 운동화 러닝화 러닝화 신발

- 나이키 브랜드의 에어맥스 90, 97 SC LTR 에센셜 제품입니다.
- 경량, 충격흡수, 통풍 기능이 있으며 고무, 스웨이드, 폴리에스테르 등의 소재로 제작되었습니다.
- 로우탑 발목높이와 3cm대 굽높이로 편안한 착화감을 제공합니다.

2. 국내매장판 나이키 리액트 X 인피니티 런 플라이니트 4 3 2 남성 여성 러닝화 런닝화 운동화

- 나이키 브랜드의 리액트 X 인피니티 런 플라이니트 제품입니다.
- 고무, 폴리에스테르, 폴리우레탄 등의 소재로 제작되었으며 경량, 충격흡수, 통풍 기능이 있습니다.
- 로우탑 발목높이와 고무 솔로 제작되어 편안한 착화감을 제공합니다.

3. 아디다스 운동화 발편한 발볼넓은 런닝화 조강화 헬스화 남성 운동화

- 아디다스 브랜드의 운동화 제품입니다.
- 메시, 폴리에스테르 등의 소재로 제작되었으며 경량, 충격흡수, 통풍 기능이 있습니다.
- 로우탑 발목높이와 3cm대 굽높이로 제작되어 편안한 착화감을 제공합니다.

위 제품들은 모두 경량, 충격흡수, 통풍 기능이 있으며 발볼이 넓은 분들도 착용하기 좋은 제품입니다. 나이키와 아디다스 브랜드의 제품들이 인기가 많으니 참고하시면 좋을 것 같습니다.

네이버 쇼핑 추천 ①



국내매장판 나이키 에어맥스 90 97 SC LTR 에센셜 남자 운동화 런...

69,000원 오늘출발

브랜드몰s Npay+



국내매장판 나이키 리액트 X 인피니티 런 플라이니트 4 3 2 남성 ...

119,000원 오늘출발

와일드덕 Npay+



아디다스 운동화 발편한 발볼넓은 런닝화 조강화 헬스화 남성 운동...

37,800원 오늘출발

스니커오프 Npay+

[그림 10] 러닝화에 대한 네이버 Cue:의 답변

구글도 SGE(Search Generative Experience)라는 유사한 서비스가 있다. 그런데 여기서 네이버나 구글이 자신들의 검색 광고 시장을 그대로 두고 이와 같은 서비스를 하게 되면 검색 광고 시장이 위축될 수 있다는 문제가 발생한다. 바로 이것이 네이버와 구글의 고민일 것이다.

네이버의 Cue와 구글의 SGE가 활성화돼서 트래픽이 유입되고 매출이 올라가면 기존의 검색 광고를 하던 기업들이 Cue나 SGE로 이동하는 결과를 초래할 수 있다. 그리고 그렇게 되면 검색이 힘을 발휘하지 못할 가능성이 높다. 구글 검색 엔진은 너무나 많은 정보를 주어 오히려 소비자 들이 구매를 위한 조언을 얻기 불편하기 때문이다. 결국 검색 시장의 93%를 차지하고 있어 세계 최대의 검색 엔진으로 불리는 구글이지만 ChatGPT로 인해 흔들릴지도 모른다.

구글은 Bard에게 질문한 결과를 만드는 과정에서 맞춤 정보 제공과 고객이 필요한 키워드로 검색 하는 것은 서로 다른 성격이라고 이야기하지만, 소비자가 구매 행동을 바꾸게 되면 유통 시장과 제조 산업은 또 한번 혁명을 맞게 될 것이다. 고객은 구매를 위해 검색은 구글에서, 자신에게 맞는 정보는 ChatGPT에서 찾는 등 불편하게 두 군데를 방문하고 싶지 않아 할 것이기 때문이다.

현재는 ChatGPT가 고객의 쇼핑과 관련해 완벽한 서비스를 하고 있지 않지만, 이는 오픈AI가 고객이 문의한 상품 설명에 해당 상품을 판매하는 쇼핑몰 링크와 사진을 추가하기만 하면 쉽게 해결이 가능한 사항이다. 즉 오픈AI는 매우 쉽게 쇼핑몰 광고 비즈니스를 할 수 있는 것이다.

오픈AI가 쇼핑몰 광고 비즈니스에 참여할지는 미지수이지만, 만약 참여한다면 현재의 쇼핑몰 비즈니스는 엄청난 변화를 겪게 될 것이다. 제품을 판매하는 쇼핑몰뿐 아니라 서비스를 제공하는 여행사, 부동산, 금융 거래 등 수많은 부문에서 새로운 비즈니스가 창출될 것이다.

그리고 ChatGPT로 인해 비즈니스 구조가 변화되면 소비자를 상대로 하는 기존의 B2C 비즈니스도 다른 방식으로 바뀌게 될 것이다. 즉 제품 판매 업체를 모아 놓은 현재의 아마존 같은 플랫폼 비즈니스가 아니라 경쟁력 있는 단위 사업체 형태로 각 제품 카테고리별 1, 2위의 전문 제품을 판매하는 방식으로 변화될 가능성이 높다.

4

공공기관의 LLM 활용을 위한 제언

현재 AI 업계의 이슈 중 하나는 AI로 어떻게 돈을 벌 수 있는가다. ChatGPT 열풍이 어느정도 지나가고 이제 사용자들이 AI가 자신에게 얼마나 도움이 되는지를 따지는 시기에 도달하자, AI 기업들은 어떻게 하면 AI로 수익을 올릴 것인가에 대한 근원적인 질문을 하고 있다. 즉 어디에 AI 사업을 포커싱할 것인지가 초미의 관심사라고 할 수 있다.

이런 상황에서 공공기관들은 어떻게 AI 기업들의 LLM 모델을 적절히 활용할 수 있을지에 대해 구체적인 논의가 필요하다. 우선 ChatGPT의 활용은 업무에 실제로 적용해서 생산성을 높이는 데 중점을 두어야 한다. 이를 위해서는 공공기관마다 PE(Prompt Engineering) TF팀을 운영해서 하나하나 업무에 적용하는 사례를 만들 필요가 있다. 그리고 PE 사례집, 업무별 ChatGPT 적용 매뉴얼 같은 것들도 갖추어야 한다.

보안 문제를 피하는 길을 찾는 것도 매우 중요하다. ChatGPT 프롬프트에 쓴 내용들이 오픈AI로 흘러들어 학습 데이터로 사용되지 않도록 오픈AI의 API와 랑체인을 통해 ChatGPT 인터페이스를 따로 만들 필요가 있다. 이 경우 사용자마다 월 정액제로 결제하는 대신 사용 토큰당 결제가 가능해 비용 절감도 가능하다.

IT 부서의 역할에도 변화가 요구된다. 지금까지 AI에 대해 IT 부서에서는 ‘AI팀이나 CDO(Chief Digital Officer)가 알아서 할 것이므로 우리는 신경 안 써도 된다’라는 생각이었다. 그러나 앞서 말한 보안 문제를 피할 수 있는 LLM 시스템을 사내에 구현하려면 다른 아닌 IT 부서가 중요한 역할을 해야 한다. AI 전문가의 경우 AI 모델을 만들고 학습시키는 것은 잘하지만 API를 연결하고 통합하는 등 시스템 구축의 경험이나 지식은 없는 경우가 많기 때문이다. 즉 IT 전문가들이 LLM 시스템을 구축할 때 적극적으로 지원해야 하며, 이에 따라 공공기관에서는 업무 조정이 필요할 것이다.

또 하나 중요한 것은 공공기관의 데이터를 활용할 수 있는 사내 챗봇을 만드는 것이다. LLM 시스템을 구축하다 보면 기왕 API를 쓰는 건데 공공기관 데이터로 어떤 데이터베이스를 만들어서 활용하면 어떨까 하는 생각이 든다. 그래서 나온 것이 벡터DB(VectorDB) 다. 이 기술을 사용해서 특정 공공기관의 문서와 매뉴얼 등 다양한 자료들을 벡터화해 저장하면 그 기관만의 LLM 시스템을

활용할 수 있다. 즉 해당 기관에 맞는 특화된 지식 체계가 생기는 것이다. 이것 역시 IT 부서에서 개발을 해야 한다.

나아가 해당 공공기관만의 독자적 AI 챗봇을 만들려면 파인 튜닝을 해야 한다. 그런데 여기에는 문제가 있다. 일단 비용이 많이 든다. GPU 등 하드웨어를 구입하는 데도 돈이 들고 데이터를 모으고 전처리하는 데도 돈이 든다.

그런데 비용보다 더 중요한 문제는 명확한 용도를 찾지 못하는 것이다. 거의 모든 분야에 사용할 수 있는 ChatGPT와 달리 파인 튜닝을 통해 AI 챗봇을 만들게 되면 특정한 목적에만 사용할 수 있다. 따라서 챗봇을 개발하기 전에 사용자 요구사항을 먼저 정립해야 한다. 앞서 언급한 PE TF 팀에서 기관의 업무나 산업 특성에 맞는 요구사항을 발굴해서 제대로 정리해야 한다.

사실 많은 공공기관들이 임원 의사결정 시스템, CRM 시스템과 같은 것들을 만들어 놓고 제대로 활용 못한 사례가 많다. LLM 시스템 역시 마찬가지로 되어선 안 된다.

무엇보다 AI는 제품화된 기술이 아니기 때문에 AI를 제대로 활용하려면 교육이 필요하다. 지금까지 나온 거의 모든 IT는 제품이 있다. ERP, CRM, 빅데이터, 클라우드처럼 각 기술을 다루는 전문 업체가 있고 제품이 있다. 그것을 사서 설치하고 사용법만 교육하면 된다. 그러나 AI의 경우 특정 제품을 도입한다고 되는 것이 아니다. AI는 제품이 아니라 기술이기 때문이다.

따라서 공공기관에서는 실제로 AI 모델을 사용할 공무원들에게 먼저 AI 교육을 해야 한다. 이미 AI 모델을 도입한 공공기관에서 투자에 따른 수익(ROI)이 생기지 않는 것은 그것을 만든 사람과 사용하는 사람 간의 미스매치 때문이다. 실제 활용하는 부서의 공무원들이 AI를 모르니 요구사항을 제시할 수 없었을 것이고, AI 전문가들은 업무를 모르니 활용도를 높일 수 없었을 것이다. 그러니 공공기관에서 AI 모델을 통해 성과를 만들기 위해서는 전사적으로, 특히 의사결정자들에 대해 먼저 AI를 교육해야 할 것이다.

5

공공기관의 LLM 도입을 위한 5개의 아키텍처

공공기관에서 ChatGPT를 비롯한 LLM 모델을 사용하기 위해서는 다음의 5가지 이슈를 중요하게 고려해야 한다.

가장 큰 이슈는 보안이다. 프롬프트에 입력하는 공공기관 정보가 AI 모델 개발사에 흘러든다면 최악의 경우 자사의 기밀 정보가 차기 GPT 버전에서 답변으로 튀어나올 수도 있다. 오픈AI는 API를 통한 프롬프트 정보에 대해서는 절대 차기 버전 학습에 사용하지 않는다고 발표했지만, 이에 대해 많은 이들은 여전히 의문을 품고 있다.

두번째 이슈는 어떻게 학습 데이터를 만들고 학습시킬 것인가라는 것이다. 모든 공공기관이 자신들의 데이터를 학습시켜서 해당 기관에 대한 구체적인 내용을 답변하게 하고 만들고 싶어한다.

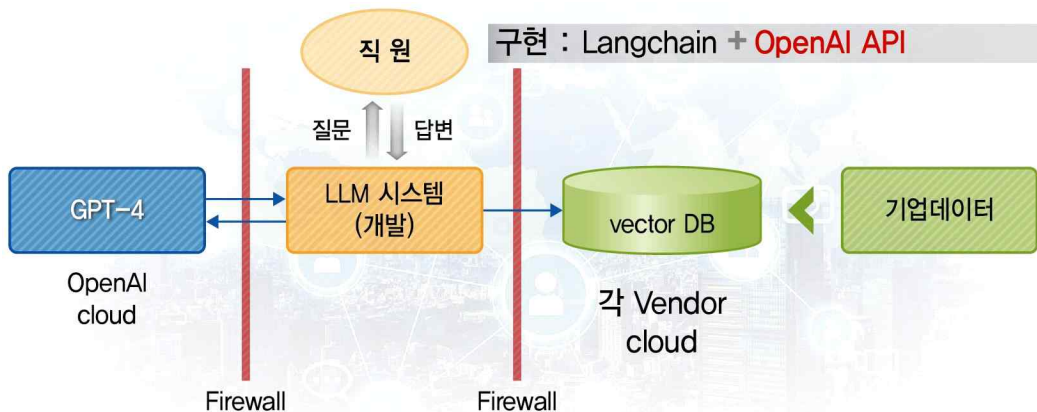
세번째 이슈는 실제로 필요한 기능의 정의와 구축 비용 그리고 시간이다. 이 부분은 LLM을 구축할 때 공공기관 공통적인 부분과 해당 공공기관의 특수한 부분으로 구분해 효율적으로 구현해야 할 것이다.

네번째는 공공 클라우드를 활용하거나 내부 망(Firewall 내부)에서 운영하는 것이다.

다섯째는 LLM을 파인 튜닝하고 벡터DB 운영 기술을 확보하기 위해 전문가를 영입하는 것이다. 이를 위해서는 기존의 전자정부 프레임워크와 공공 클라우드 운영 경험을 살릴 필요가 있다.

다음에서는 이러한 5가지 이슈를 중심으로 현재 구현 가능한 5가지 LLM 아키텍처를 소개한다.

A 타입A: GPT-4 액세스를 API를 통해서 구현



[그림 11] 타입A. GPT-4 직접 액세스를 오픈AI API로 구현

A 타입은 LLM 시스템을 통해서 GPT-4에 액세스하는 방법을 개발하는 것이다. 이는 랭체인이라는 파이썬 패키지와 오픈AI의 API를 활용해서 구현할 수 있다. 다만 이 경우 각 기관의 데이터를 GPT-4에 직접 학습시키기는 어렵고 벡터DB를 거쳐야 하는데, 현재 수십 종의 무료 또는 유료 벡터DB가 나와 있다.

이렇게 각 벤더에서 제공하는 클라우드에 데이터를 먼저 올릴 때는 오픈AI의 API를 사용해 벡터화한 후에 입력한다. 그리고 LLM 시스템에 직접 사용자 입력을 받을 수 있는 UI를 구축한 후 랭체인을 활용해서 직원이 프롬프트를 작성하면 그 프롬프트 내용을 벡터DB로 조회한다. 그 결과를 오픈AI API를 통해 GPT-4에 보내고 답변을 받은 후 사용할 수 있다.

이 방식은 오픈AI에 입력한 프롬프트 정보를 API를 통해서 보내는 것이 특징이다. 한편 앞서 말했듯 오픈AI는 이처럼 API를 통해 수집한 데이터는 학습 용도로 사용하지 않는다고 발표했다.

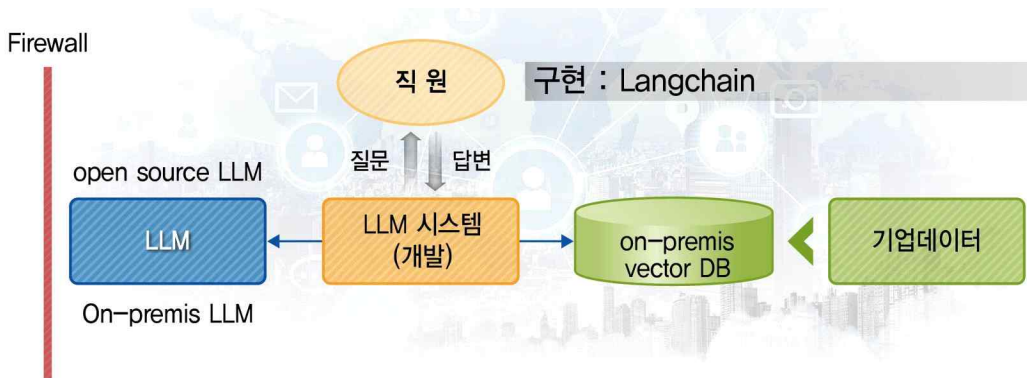
(물론 이 약속을 믿을 것인가는 다른 문제다.)

한편 이 방식의 경우 오픈AI는 종량제로 과금한다. 즉 사용하는 프롬프트의 단어 수와 답변의 단어 수를 합해서 과금하는 방식이다.

이렇게 LLM 시스템을 개발하는 방식은 전산팀이 있는 공공기관에서는 그리 어렵지 않게 구현할 수 있고, 전산팀이 없다면 외주를 줄 수도 있다. 그러나 오픈AI의 GPT-4와 벡터DB가 내부

망 밖에 있기 때문에 LLM 시스템을 외부에 구축해야 한다는 큰 단점이 있다. LLM 시스템 자체가 공공기관의 중요한 정보라는 측면에서 보안 우려가 발생할 수밖에 없는 대목이다.

B 타입B: 오픈 소스 LLM을 사용



[그림 12] 타입B: 오픈 소스 LLM 사용

B 타입은 오픈AI를 신뢰하지 않을 경우 선택 가능한 대안으로, 보통 오픈 소스 진영에서 만든 LLM들을 사용한다. 허깅페이스(Hugging Face)라는 벤처업체가 운영하는 웹사이트에 가면 지금까지 만들어진 수백개의 LLM들이 있는데, 이 중 적절한 것을 다운로드해서 사용하거나 사용 벤더의 LLM을 사용할 수 있다. 랭체인을 통해 거의 모든 LLM들을 사용할 수 있으므로 구현하는 방법은 A 타입과 동일한다.

이 방식의 경우 오픈AI에 비용을 지불하지 않아도 되지만, 다운받은 LLM을 구동할 GPU가 달린 서버를 사용하는 비용은 필요하다. 예를 들어 LLM을 제공하는 공공기관의 클라우드를 사용하면 과금이 된다. 이때 사용하는 LLM이 GPT-4보다 성능이 떨어지거나 한글을 제대로 지원하지 못할 수도 있으니 미리 확인해야 한다.

내부 망에 구축하는 경우에도 내부용(On-premise) LLM과 내부용 벡터DB를 사용하기 위해 시스템 설치, 유지, 보수, 업그레이드, 운영에 드는 비용과 인력이 필요하므로 이러한 작업을 대행해주는 기관을 활용할 필요가 있다. 또한 각 기관별로 내부용 LLM과 내부용 벡터DB의 유지 보수를 위해서 공공 클라우드를 활용할 필요가 있다.

C 타입C: 파인 튜닝 방식



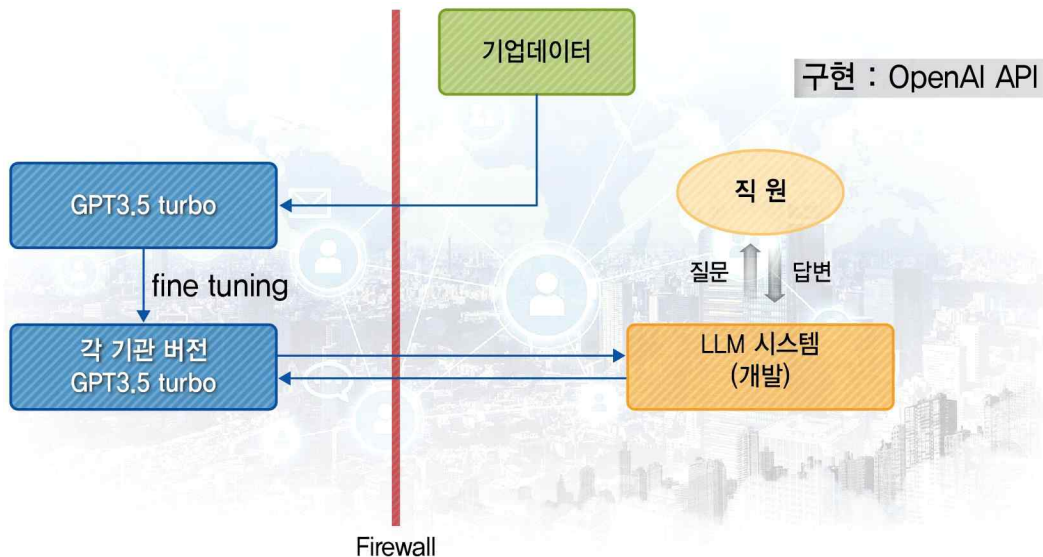
[그림 13] 타입C: 파인 튜닝 방식

C 타입은 A, B 타입에서 사용하는 벡터DB를 완전히 각 공공기관의 데이터로 교체해서 학습하는 방식이다. 아직은 벡터DB의 성능이 완전하지 않아 많은 양의 공공기관 데이터를 저장하기 힘들고 검색의 정확도가 떨어지므로 이때는 파인 튜닝 방식을 사용한다. 공공기관의 데이터를 전처리해서 메타의 LLaMA2와 같은 오픈 소스 LLM 기초 모델에 파인 튜닝을 하는 것이다. 그 결과물은 모델 파일(키 포인트 파일)로 나오며, 이를 공공기관 내 GPU 서버에 올려서 활용하는 방식이다.

이 경우 공공기관 내 데이터를 전처리하는 비용과 파인 튜닝하는 비용 그리고 GPU 서버 비용이 들게 된다. 따라서 공공기관을 대표한 기초 모델을 만들고 이를 기반으로 파인 튜닝을 하는 것이 좋다.

이 방식의 단점은 파인 튜닝을 하게 되면 기초 모델의 일부분만 사용하게 되기 때문에 GPT-4에 비해 성능이 떨어진다는 것이다. 반면 가장 안전하고 무엇보다 공공기관의 특별한 목적에만 맞는 일을 할 수 있다는 것은 장점이다.

D 타입D: 오픈AI의 파인 튜닝 방식



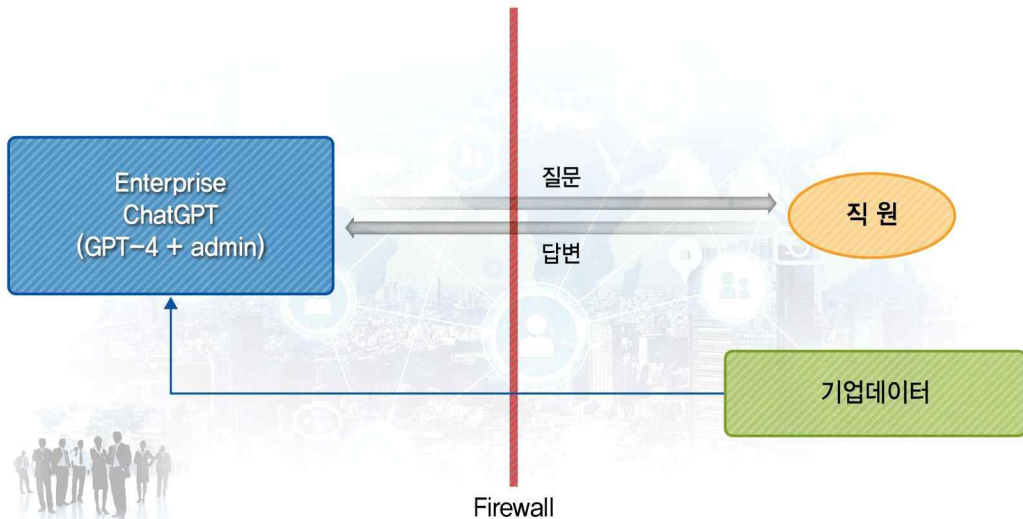
[그림 14] 타입D: 오픈AI의 파인 튜닝

D 타입은 오픈AI의 파인 튜닝을 활용하는 것이다. 오픈AI는 2023년 8월 22일 자사의 ChatGPT (GPT-3.5 turbo)를 파인 튜닝하는 방식을 출시했다. 이를 활용해 공공기관 데이터를 전처리해서 오픈AI의 클라우드 서버에 올린 후 오픈AI의 API를 사용해서 파인 튜닝할 수 있다.

이때 오픈AI의 GPU를 사용하게 되므로 공공기관 버전의 ChatGPT가 생기게 된다. 이것은 오픈 AI에 보관하고, 각 공공기관의 LLM 시스템을 개발해서 오픈AI를 사용해 해당 기관 버전의 ChatGPT를 사용할 수 있다.

이 방식의 장점은 파인 튜닝과 GPU 사용 비용을 종량제로 지불하면 되기 때문에 C 타입보다 훨씬 저렴하다는 것이다. 반면 A, B 타입과 마찬가지로 중요 정보를 외부 망에 구축해야 한다는 것 그리고 오픈AI에 데이터를 보관해야 한다는 점은 여전히 보안상 문제가 될 수 있다.

E 타입E: 오픈AI의 Enterprise ChatGPT



[그림 15] 타입E: Enterprise ChatGPT 형태

E 타입은 오픈AI에서 2023년 8월 28일 출시한 Enterprise ChatGPT를 활용하는 방식이다. 한마디로 오픈AI가 공공기관용으로 ChatGPT를 만들어 주는 것이라 할 수 있다. 공공기관 데이터를 오픈AI에 보내면 해당 공공기관 버전의 GPT-4를 만들어 그 기관만 사용할 수 있도록 하는 것이다. 한편 오픈AI는 공공기관이 보내온 데이터를 SOC2(Service Organization Controls 2) 기준에 맞추어 운영하는데, SOC는 미국회계사협회(AICPA)가 만든 내부 통제 기준이다.

이 방식의 장점은 한꺼번에 많은 질문을 처리할 수 있다는 것이다. 프롬프트에서 3만 2000개의 토큰을 사용할 수 있으며, 이는 A4 용지 약 30페이지 분량에 해당한다. 속도 역시 2배 빠르고 무제한으로 사용할 수 있는 것도 장점이다.

아직 이 방식을 사용하는 비용에 대해서는 구체적인 산정이 되지 않았다. 그리고 외부 망에 구축하는 것이기에 보안 문제는 여전히 존재한다.

F 공공기관에 적합한 모델

일반적인 기업들은 앞서 설명한 5가지 방식 중 A 타입을 많이 활용하고 있다. 가장 비용이 적게 들고 보안 문제를 해결할 수 있기 때문이다. 하지만 최근에는 다른 옵션으로 마이그레이션 계획을

가진 기업들도 적지 않다. 사용을 이어가면서 발생하는 다양한 요구사항에 대응하기 위해 계속 자체 업그레이드를 하다 보니 한계를 느끼기 때문이다.

보안이 중요한 대기업은 B, C 타입으로 전환하고 있다. 추후 오픈AI 한국 지사가 생긴 후에는 E 타입을 사용하는 기업들도 늘어날 가능성이 있다.



[그림 16] 공공기관의 LLM 모델

일반 기업과 달리 공공기관은 데이터의 보안 때문에 B 또는 C 타입을 고려할 수밖에 없다. 가장 좋은 것은 그림과 같이 그 두 방식을 통합한 것이다.

다만 공공기관의 업무는 상당히 많은 법령과 규제들에 의해서 운영되므로 공공기관의 LLM 역시 많은 법률, 조례, 규칙, 규제 등의 내용을 기본적으로 담아야 한다. 따라서 일반 기업에서 사용하는 오픈 소스 LLM 기초 모델과 더불어 우리나라 전체 공공기관과 공기업에서 사용할 수 있는 ‘공공기관 공통 기초 모델(Common Public Foundation Model)’도 만들 필요가 있다. 각 공공기관은 이 기초 모델을 파인 튜닝해서 기관별 sLLM 모델을 만들어 사용하는 것도 고려할 수 있다.

또한 환각 현상을 최대한 줄이기 위해 방화벽 내부에 있는 내부용 벡터DB에 신규로 생성된 데이터나 중요한 개별 공공기관 데이터는 벡터화해 저장할 필요가 있다. 이렇게 구현하는 방식을 RAG(Retrieval Augmented Generation)라고 한다.

보안, 환각 외에도 공공기관 LLM은 구축 기술, 비용, 시간, 인력, 운영, 유지 보수 등 많은 이슈들이 있다. 따라서 이러한 이슈들을 미리 걸러내 범정부적으로 LLM 도입 전략을 수립해야 할 것이다.

NIA

Digital Insight 2023

ChatGPT의 파급 효과와 기관의 LLM 도입 전략

■ 발 행 : 2023.12.14.

■ 발행인 : 황종성

■ 발행처 : 한국지능정보사회진흥원(NIA)정책본부AI·미래전략센터

■ 기획 및 문의 : 김소미 선임(somikim@nia.or.kr)

■ 작 성: KAIST 김재철SI대학원 장동인 교수(donchang0725@gmail.com)

- NIA 「Digital Insight 2023」는 디지털 트랜스포메이션(Digital Transformation) 시대를 맞아해 다가오는 미래를 준비하고, 미래 지능화 시대를 선제적으로 대응하기 위해 한국지능정보사회진흥원(NIA)에서 발간하는 보고서입니다.
- 본보고서는 방송통신발전기금으로 수행한 정보통신·방송 연구개발 사업의 결과물이므로, 보고서의 내용을 발표할 때는 반드시 과학기술정보통신부 정보통신·방송 연구개발 사업의 연구 결과임을 밝혀야 합니다.
- NIA의 승인 없이 본 보고서의 무단전재나 복제를 금하며, 인용하실 때는 반드시 NIA 「Digital Insight 2023」라고 밝혀주시기 바랍니다. 보고서 내용에 대한 문의나 제안은 위의 연락처로 해주시기 바랍니다.
- 본 보고서의 내용은 한국지능정보사회진흥원(NIA)의 공식 견해와 다를 수 있습니다.