# 생성형AI 기술 및 서비스 도입 방안

마창수*, 장석우**
*안양대학교 경영학과
**안양대학교 소프트웨어학과
e-mail: swjang@anyang.ac.kr

# Introduction of generative AI technologies and services

Chang-Su Ma*, Seok-Woo Jang**
*Dept. of Business Administration, Anyang University
**Dept. of Software Science, Anyang University

요 약

본 논문은 생성형AI의 산업 적용 관점에서 이해를 돕기 위해 서비스 기술의 생성과 발전, 기능적 특징에 대해 알아보고, 최근 생성형AI의 서비스를 제공하는 기업들과 서비스 방식에 대해 알아본다. 지금까지 생성형AI는 공공의 지식 제공자 역할을 수행하고 있다면 향후 기업 혹은 나아가 산업 도메인을 위한 특화된 영역에 쓰이기 위한 조정 모델의 기능 개발이 가속화할 것으로 예상하였다.

## 1. 서론

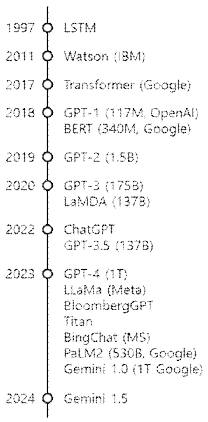
생성형AI는 2018년 OpenAI의 GPT[1] 발표로 인해 관심을 받기 시작하였고, 현재는 가장 큰 성장을 보이는 AI 영역으로 발돋움하였다. 초거대 규모의 데이터를 활용한 LLM(Large Language Model)을 훈련하기 위해서는 이전과 비교할 수 없는 큰 시스템 자원과 데이터를 필요로 하며, 모델은 세부적인 수정과 개선이 쉽지 않아 생각지 못한 윤리적 문제를 야기할 위험성도 존재한다. 현재는 보편적 AI에서 사업 특화된 Vertical AI로 확장하는 시기이며 기업에서는 생성형AI 도입에 박차를 가하고 있다. 대형 모델이기 때문에 기업의 환경에 적합한 적용 모델을 검토하고 활용 및 구축을 위한 방안을 정하되, 빠른 변화 속에 새롭게 출시되는 서비스와 오픈소스를 지속적으로 모니터링하는 전략이 필요하다.

## 2. 생성형AI의 역사

챗봇은 사람과 자연스러운 대화를 수행하는 기능은 제공하였고, 주로 문장의 의도(intent)와 단어에 해당하는 개체(Entity)와 규칙 기반의 대화(Dialog)로 구성되어 있고, 이 규칙은 사람에 의해 사례에 맞게 구성되어야 하므로 정해진 시나리오에 따르는 대화만이 가능한 한계를 갖는다.

이후 트랜스포머를 응용해 대규모의 지식 데이터와 컴퓨팅 자원을 이용해 방대한 지식에 대한 학습을 통해 Super AI를 지향하는 OpenAI의 ChatGPT가 2022년에 출현하게 되고, 이어서 Google의 Bard가 발표되었다. 2024년에 Google에서 발표한 Gemini에 이르기까지 여러 대형언어모델(LLM, Large Language Model)들이 공개되었다.

생성형AI는 단지 언어 모델만을 지칭하는 것이 아니라 텍스트, 코드, 이미지, 영상, 음악 등 다양한 형식의 데이터를 생성하는 것을 포함한다. 단, 언어 모델 이외의 생성은 주로 창의나 창작의 영역에 해당하는 경우가 많으므로 이 논문에서는 주로 언어모델에 대해 논한다. [그림1]은 대표적인 언어 모델이 LSTM부터 시작해 최근 구글이 발표한 Gemini 1.5까지의 발전 단계를 보여주고 있다. 대표적인 글로벌 기업 간에 파라미터 개수를 크게 늘리며 모델을 고도화하고 있는 것을 볼 수 있다.




[그림 1] 대표적인 언어 모델
(괄호 안의 숫자는 모델의 파라미터 수)

언어 모델은 딥러닝을 적용한 LSTM 모델에서부터 큰 발전이 시작되어 기존 통계 방식의 문장분류, 구문 분석, 형태소 분석, 통계적 기계번역 등의 적용 단계에서, 대규모 지식을 기반으로 자연어를 이용한 대화, 문서 요약, 문서 작성 등 고급 답변을 제공해 주는 수준으로 발전하였고, 여러 질문에 대해 의도와 맥락을 유추하여 가장 적합한 답변을 제공하는 수준에까지 이르렀다.

## 3. 최근 대형 언어 모델의 서비스 동향

가트너의 2023년 AI 보고서[2]에 따르면 생성형AI는 가트너 Hype Cycle상에 과장된 기대의 정점(Peak of inflated expectation) 단계를 지나고 있다. 이후 큰 환멸(Trough of Disillusionment) 단계에 빠져 헤어 나오지 못할지, 빠르게 계몽 단계(Slope of Enlightenment)와 생산성 안정 단계(Plateau of Productivity)로 나아갈지 알 수 없다. 하지만 추운 겨울이 지나고 딥러닝으로 인해 큰 성장을 이룬 AI 산업 전반으로 보면 이미 환멸 단계를 벗어나고 있다고 여겨진다.

생성형AI는 OpenAI의 GPT 연구의 높은 성과에 따라 Meta, Google, MS, Naver 등과 같은 대형 IT 기업들이 잇따라 생성형AI 모델 개발 및 발표가 잇따르고 있다.

ChatGPT는 가장 대표적인 LLM 모델로 비영리 단체인 OpenAI 주도로 개발되었다. GPT 엔진을 기반으로 사람과 자연스러운 대화를 통해 질문 답변, 산술 연산, 언어 이해, 번역, 상식 추론, 코드 완성, 논리적 추론, 패턴 인식 및 독해 등의 기능을 제공한다. 이전의 챗봇 서비스와 차별화되는 점은 답변 형식 문장을 사람이 직접 지정하지 않아도 적합한 문장을 자체 생성한다는 것이다. 때문에 답변이 좀 더 자연스럽고, 때에 따라 원하는 형식을 요청할 수 있다. 기존에도 지식 기반에서 온톨로지 등을 이용해 적합한 답을 추론할 수 있는 시스템이 있었지만 답변을 다양한 형식으로 생성하여 사람과 보다 자연스러운 대화를 통해 의사소통 한다는 것이 생성형 AI의 특징이라고 할 수 있다.

구글은 24년 2월에 Gemini 1.5를 발표하였다. Gemini 1.5는 유사한 파라미터 크기이지만 더 적은 컴퓨팅 자원으로 더 긴 문맥을 이해할 수 있게 발전되었다고 발표하였다. 약 100만개의 토큰을 처리할 수 있으며 PDF파일, 보다 긴 코드 저장소, 비디오 및 오디오 처리 등 보다 큰 크기의 데이터 처리가 필요한 비즈니스 업무에 적용할 수 있다는 장점이 있다.

## 4. 생성형AI 도입 시 고려사항

생성형AI의 제공 방식은 웹 UI, Restful API, 클라우드 서비스 등을 통해 제공한다. 기업에서 생성형AI를 비즈니스에 적용하려면 Restful API 방식이나 Cloud Service를 통한 Application 개발 방식이 적합하다.

웹 방식의 경우 특정 사이트에 접근하여 자연어 대화, 사용자 프롬프트, 파일 업로드 등을 통해 질의하고 답변을 얻을 수 있다. 대화 방식의 경우는 다중 질의 및 답변을 통해 답변의 정확도와 품질을 높여가는 방식이 사용된다. 또한 생성형AI에게 역할을 부여하여 원하는 답변의 형식을 기대할 수 있다. 웹 방식의 경우 웹에 접속하여 진행하므로 산업에서 특정 도메인이나 Use Case에 적용해 사용하는 데는 한계가 있다.

Restful API 방식은 애플리케이션에서 접근 허용 토큰을 이용해 인증 후 통신을 통해 생성형AI 기능을 사용할 수 있다. Restful API 방식이 기업의 비즈니스 도입에 편한 이유는 개발 언어 선택에 제한을 받지 않기 때문에 애플리케이션 개발에 좀 더 자유롭기 때문이다. API를 통해 생성형AI 모델 프로토타입을 생성하고, 프롬프트 엔지니어링이나 모델 튜닝을 수행할 수 있고, 질의 및 응답을 확인할 수 있다.

Cloud Service 방식의 경우 AWS, GCP, Azure와 같이 Cloud Provider에서 제공하는 AI Platform을 이용해 개발하는 방식을 의미한다. LLM이라는 용어에서도 의미하듯이 생성형AI는 초거대 데이터를 활용한 대형 모델이다. 일반 기업이나 개인이 모델을 직접 구성하거나 훈련시키는 것은 거의 불가능하다. 따라서 기업에서 생성형AI를 자신의 비즈니스에 잘 녹이려면 기업의 환경 측면에서 사용하기에 적합한 생성형 모델을 선정하기 위한 검토가 선행되어야 한다.

생성형AI를 선정할 때 기존 Cloud 기반 시스템을 활용하게 될 경우 퍼블릭 클라우드 사용시 발생 가능한 보안 위협[3]에 대해 고려해야 한다. 특히 산업기술의 유출방지 및 보호에 관한 법률에 의거해 국가에서 보호하는 산업의 경우 퍼블릭 클라우드를 사용하는데 보다 세심한 주의가 필요하다. 물론 2022년 개정된 '전자금융감독규정' 개정과 '금융 분야 클라우드컴퓨팅서비스 이용 가이드'에 따라 금융 분야에서 클라우드 활용이 다소 완화되었으나, 그만큼 보안에 대한 안전조치 확보가 더 중요하게 되었다.

## 5. 생성형AI의 진화 전망

생성형AI는 그동안 보편적AI에 해당하는 수평적 AI에 집중해 왔으나, 향후에는 특정 영역에서 역할을 담당할 수 있도록 맞춤 설계되고 특정 영역에 전문화된 지식을 갖는 수직적 AI로 발전[4]해 갈 것이다. 수직적AI는 기업에 직접적으로 필요하며 수익에 직결되어 있다. 보편적 지식으로 훈련된 수평적AI에 특정 산업의 지식과 페르소나를 접목해 업무에 투입할 수 있는 생성형AI로 진화해 나갈 것이고, 이를 기업에서 보다 쉽게 사용할 수 있도록 모델 조정, 파인 튜닝, 퓨샷 러닝

등의 방법이 발전되어야 한다.

생성형AI는 멀티 모델로 발전되고 있다. 멀티 모달은 기계 학습 분야에서 사용되는 용어로, 데이터가 여러 가지 형태나 '모드'를 가질 수 있다는 개념을 나타낸다[5]. 예를 들어, 이미지와 텍스트를 동시에 처리하는 시스템은 '다중모드' 시스템이라고 할 수 있고, 이러한 시스템은 각각의 모드에서 정보를 추출하고, 이 정보를 결합하여 더 복잡한 작업을 수행할 수 있다. 멀티 모달은 자연어 처리(NLP), 컴퓨터 비전, 음성 인식 등 다양한 기계 학습 분야에서 중요한 연구 주제이다.

## 6. 생성형AI의 산업 도입 현황

Dataiku와 Databricks가 2023년 6월 조사한 '제조 및 에너지 분야 AI 현황'[6]에 의하면 제조 산업이 전체 산업에 비해 AI 내재화 및 확장성이 더 앞서는 것으로 나타났다. 반면 생성형AI 혹은 대규모 언어모델 기술사용 계획에 대해서는 오히려 낮은 비중을 나타내고 있다. 이는 생성형AI에 대한 이해가 부족하거나 생성형AI를 어떻게 도입해야 할지에 대한 의사 결정이 내려지지 않은 이유로 해석할 수 있다. 하지만, 또 다른 이유로 대규모 언어 모델을 이용한 생성형AI가 산업 특화된 지식 보다는 대중을 상대로 한 공공의 지식을 기반으로 구축되어져 있고, 산업 혹은 특정 기업의 업무와 지식을 고려한 특성화된 모델을 확보할 수 있는 프롬프팅 기술의 발전과, 산업 특화 모델을 위한 모델 조정화(Instruction tuned)기술의 고도화가 필요하다.

### 참고문헌

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Open AI, "Improving Language Understanding by Generative Pre-Training", 2018.06.11

[2] Gartner, "Hype Cycle for Artificial Intelligence 2023"

[3] 박형근, "퍼블릭 클라우드 컴퓨팅 사용 기업이 고려해야 할 보안 위험과 대응 방안에 대한 小考", 정보보호학회지 제22권 제7호 2012.11 46 - 53

[4] 서보배, 삼성SDS, "글로벌 AI 리더들이 전하는 생성형 AI 기술의 전망", 2023.11.10

[5] Letitia Parcalabescu, Nils Trost, Anette Frank, "What is Multimodality?", arXiv:2103.06304v3 [cs.AI] 10 Jun 2021

[6] data iku & databricks, "설문조사 보고서, 제조 및 에너지 분야의 AI 현황", 2023.06