# 생성형 AI 신뢰성 확보를 위한 데이터 검증

**김민우** 셀렉트스타 NLP팀 팀장
**이일구** 셀렉트스타 NLP팀 연구원

## 1. 머리말

생성형 AI란 텍스트, 이미지, 음성, 비디오 등 새로운 콘텐츠를 만들어 내는 AI 시스템을 말한다. 학술적으로 이야기하면, '데이터 셋의 분포를 학습하는 모델'을 의미한다. 데이터 셋의 분포를 학습함으로써 AI 시스템은 실제 데이터와 유사한 새로운 데이터를 생성할 수 있다. 이런 개념을 통해 여러 생성 모델들이 만들어졌고, 이를 다양한 데이터 형태에 적용해 큰 성공을 거두었다.

대표적인 예로, 사용자의 질문에 반응하여 텍스트를 생성하는 모델인 ChatGPT[1], 제미나이(Gemini)[2], 클로드3(Claude 3)[3]와 주어진 텍스트에 부합한 그림을 그리는 달리 3(DALL-E 3)[4], 스테이블 디퓨젼 3(Stable Diffusion 3)[5], 미드저니(Midjourney)[6] 등이 있다. 최근엔 텍스트로부터 비디오를 생성할 수 있는 모델인 소라(Sora)[7]가 등장해 큰 화제가 됐다. 이처럼 다양한 생성형 AI 모델이 나옴으로써 그 활용 범위가 매우 넓어지면서, 중요성 역시 강조되고 있다.

다양하게 출시된 생성형 AI 모델이 여러 분야에 활용되기 위해선, 신뢰성 확보가 매우 중요하다. ChatGPT가 출시된 지 얼마 되지 않았을 때 몇 가지 문제점들이 발견됐다. 그중 대표적인 것이 환각(Hallucination) 문제다. 우리나라에서 잘 알려진 환각의 예는 '세종대왕 맥북프로 던짐 사건'이다. 이는 "조선왕조실록에 기록된 세종대왕의 맥북프로 던짐사건에 대해 알려줘"라는 질문에, ChatGPT가 가공의 사건을 마치 있었던 것처럼 포장해 그럴듯하게 설명해 준 사건이다. 즉, 생성형 모델은 있지도 않은 사실을 그럴듯하게 이야기할 수 있는데, 사실관계가 분명해야 할 사건을 다룰 때 큰 문제가 된다.

또 다른 문제점으론 편향(Bias)이 있다. 최근 연구결과에 따르면, AI가 미국 표준 영어보다 아프리카계 미국 영어로 작성된 텍스트를 분석할 때 더 많은 인종차별적 경향을 보인다고 한다[8]. 이러한 사회적 편견은 생성형 모델을 사용하는 사람들에게 은연중에 고정관념을 강화시킬 수 있다. 이처럼 생성형 AI 모델이 급속도로 발전하는 만큼 발견되는 문제점을 해결하고 신뢰성을 높이는 것은 매우 중요하다. 따라서 이번 원고에선 신뢰성 확보를 위한 방법을 본격적으로 알아볼 계획이다.

그 전에, 자주 쓰이는 용어와 글의 범위를 제한하고자 한다. 이번 원고에서 '생성형 AI, 생성형 모델, 생성형 AI 모델, 생성 모델' 등의 용어는 전부 같은 의미로 사용된다. 그리고 생성형 AI 모델은 여러 데이터를 다룰 수 있지만, 여기서는 흔히 LLM(거대언어모델, Large Language Model)이라 불리는 텍스트 기반 모델들 위주로 설명한다. 본래 LLM의 신뢰성은 '답변의 정확성 및 일관성'을, 안전성은 'LLM의 답변이 유해하지 않고 윤리적인 내용을 지키는 것'을 의미한다. 하지만

이번 원고에선 신뢰성과 안전성 두 개념을 한데 묶어 신뢰성이라는 용어로 사용하고자 한다.

## 2. 생성형 AI 기술의 현황

### 2.1 자연어처리 분야에서 생성형 모델의 발전

자연어처리(Natural Language Processes) 분야에서 생성모델을 이야기하려면, 먼저 언어모델(Language Model)부터 설명해야 한다. 언어모델이란 주어진 텍스트의 다음 단어를 예측하는 모델이다. 예를 들면, "생성형 AI 모델의 발전이 ___"와 같은 문장의 빈칸에 들어갈 단어를 예측하는 것이다. 이때 "빠르다", "매우", "눈부시게" 등의 단어가 들어갈 확률이 "사과", "창밖"보다 높다는 것을 알 수 있다. 대량의 텍스트 데이터 셋을 학습해 모델을 만들면 문맥에 기반해 자연스럽고 의미 있는 다음 단어를 예측(생성)할 수 있다.

초창기 언어모델은 n-gram 모델이나 HMM(Hidden Markov Model)과 같은 통계적 모델을 써서 연구됐고, 이는 RNN, LSTM 같은 신경망(뉴럴네트워크) 기반 연구로 이어졌다. 최근 LLM 기반 생성모델은 Transformers[9] 모델로부터 시작됐다. Transformers는 self-attention이라는 새로운 알고리즘을 바탕으로 기존 모델들이 잘 처리하지 못했던 긴 문맥의 텍스트를 성공적으로 처리했고, 자연어처리 분야의 새로운 표준으로 자리잡았다.

Transformers는 Encoder와 Decoder로 이뤄져 있다. 이 중 Transformer의 Encoder만 갖고 만들어진 모델이 BERT[10], Decoder만으로 만들어진 모델이 GPT[11]다. 이후 각 모델을 바탕으로 Encoderonly 모델, Decoder-only 모델이 발전하게 된다. 생성모델은 둘 중 Decoder-only 모델을 중심으로 빠른 발전이 이뤄지고 있다. 우리가 잘 알고 있는 GPT-1 ~ GPT-4, 제미나이, 클로드 모두 Decoder-only 모델을 기반으로 한다. closed-source 모델이 아닌, 메타(Meta)에서 오픈소스(open-source)로 공개한 라마(LLaMA)[12]도 마찬가지다. Decoder-only 모델은 이전 텍스트 데이터에 기반해 바로 다음 단어를 예측하는 단방향 분석에 특화됐다. 덕분에 텍스트 생성에 뛰어난 역량을 보여준다.

대량의 텍스트 데이터를 통해 언어의 의미와 관계를 학습하는 것을 사전학습(Pre-training)이라고 한다. 사전학습을 하면 언어이해 능력과 일반화 능력이 향상된다. 사전학습 후 미세조정(Fine-tuning)을 하면 번역, 요약 등 하위 태스크의 성능을 높일 수 있다. 이러한 '사전학습 및 미세조정' 방법론을 쓰는 모델들을 PLM(Pre-trained Language Models)이라고 부른다. 대표적인 PLM으로 BERT, GPT가 있다.

PLM 이후 연구의 흐름은 L LM으로 이어졌다. LLM은 PLM에 비하면 매우 큰 사이즈를 갖고 있는데, 한 예로 GPT-3의 경우 175B개의 파라미터를 가지고 있다. 반면 BERT-Large는 0.3B, GPT는 0.1B 수준이다. LLM에선 이렇게 모델 사이즈와 함께, 학습 데이터 셋도 훨씬 커졌다. 이렇게 모델과 데이터 셋이 엄청나게 커지면서 emergent behavior 현상이 일어났다[13]. 이는 기존 PLM에서는 나타나지 않았던 현상인데, 문맥 내 학습(In-context learning), 지시사항 따르기(Instruction following), 단계적 추론(Step-by-step reasoning) 같은 능력을 뜻한다. 이런 능력으로 인해 번역이나 질의응답 같은 자연어처리에 대해, 특수한 태스크가 아닌 일반적인 상황에서도 답변이 가능해졌다. 이렇게 새로운 능력을 갖게 된 LLM은 다양한 분야에서 응용되고 있다.

2.2 다양한 생성형 AI 기술 소개

2.2.1 AI Assistant

우리가 잘 알고 있는 오픈AI의 ChatGPT, 구글의 제미나이, 앤트로픽(Anthropic)의 클로드 등은 생성형 모델의 좋은 예시다. 이들은 모두 대량의 데이터와 많은 수의 파라미터로 무장한 모델들이며, 어떠한 질문에도 막힘 없이 답변하는 비서와 같은 역할을 하고 있다. 2022년 11월 오픈AI에 의해 ChatGPT가 처음 출시된 이후, 2023년은 LLM의 한 해였다는걸 증명하듯이 여러 모델들이 서로 경쟁하듯 출시됐다. 그리고 바로 얼마 전인 2024년 5월, 오픈AI와 구글이 각각 ChatGPT-4o[14]와 제미나이 1.5 Pro[15]를 출시하며 LLM에서 LMM(Large Multimodal Models)으로 가는 신호탄을 쏘아 올렸다.

2.2.2 이미지 생성

최근의 이미지 생성은 주어진 텍스트를 잘 표현하는 그림을 그려주는 방향으로 발전해왔다. 대표 모델로는 오픈AI의 달리 3[4], 스태빌리티 AI(Stability AI)의 스테이블 디퓨전 3[5], 미드저니의 미드저니[6] 등이 있다. 이들 역시 서로 경쟁하듯 새로운 모델이 나오고 있고, 기존의 문제점을 빠르게 고쳐나가면서 발전하는 중이다.

2.2.3 비디오 생성

텍스트 기반 이미지 생성 연구가 진행되며, 당연하게도 텍스트가 주어질 때 비디오를 생성하는 모델도 나오리란 상상을 쉽게 할 수 있다. 하지만 비디오 생성은 생각보다 복잡한 작업이다. 간단하게 30fps라고 가정해도, 1초를 생성하는 데 이미지 30장을 만들어 내야 하는 것이다. 더불어 해당 영상과 어울리는 소리까지 같이 만들어야 한다. 결과적으로 비디오 생성은 이미지 한 장을 생성하는 태스크보다 훨씬 어렵다는 것을 쉽게 이해할 수 있다.

우리가 생각하는 수준의 모델은 2024년 2월 오픈 AI의 소라가 최초다[7]. 소라가 직접 생성한 영상을 보면, 인물의 자연스러운 연속성, 지하철 창문에 반사되는 인물의 모습 등 현실세계의 물리 현상들이 잘 구현돼 있는 것을 볼 수 있다. 앞으로 발전 가능성이 기대되는 분야다.

## 3. 데이터 검증의 중요성

LLM 발전에서 중요한 역할을 했던 요소가 바로 많은 양의 데이터다. 그런데 데이터 양만 많으면 모든 문제가 해결되는 것일까? 그렇지 않다. 데이터의 질은 AI 성능 및 신뢰성과 직접적으로 연결된다[16]. AI 모델은 주어진 데이터의 패턴을 분석하고 관계를 학습한다. 이때 정확하지 않은 데이터가 사용되면 관계 및 패턴 학습에 어려움을 겪게 되고, 부정확한 결과나 해로운 내용을 출력할 수 있다. 이는 AI 성능과 신뢰성을 크게 떨어뜨리고, 사용자에게 잘못되거나 위험한 정보를 제공해 해를 끼칠 수 있다.

잘못된 데이터를 사용했을 때 AI 모델이 어떤 위험과 문제점을 보이는지 실제 사례를 통해 알아보자. 구글에서 만든 제미나이는 텍스트에 맞는 이미지를 생성할 수 있다. 실제 제미나이에 "1929년 독일 군인을 그려줘"라는 요청을 한 결과, 1929년 독일 군인과는 관계없는 동양인이나 아메리카 원주민을 그려줬다. 이는 해당 모델이 가지고 있는 인종 편향을 보여준다. 이렇게 편향

이 나타나는 이유는 무엇일까. 백인 위주의 이미지 생성을 막기 위해 데이터 수집 및 정제에 관한 규칙을 심하게 적용했기 때문이다. 결과적으로 백인에 대한 역차별을 나타내는 모델이 만들어진 것이다.

이처럼 수집정제하는 과정에서 실제 사건과 일치하지 않는 데이터를 사용하면 잘못된 결과가 나온다. 따라서 AI의 신뢰성을 확보하기 위해서 다음 필수요소 세 가지를 고려해야 한다.

첫 번째는 정확한 데이터의 사용이다. 이는 사실관계가 명확한 데이터, 오류나 편향이 없는 데이터, 최신 데이터 등을 의미한다.

두 번째 방법은 데이터의 다양성과 포괄성을 담보하는 것이다. 특정 그룹이나 관점을 대변하지 않고, 여러 의견이나 관점, 상태들을 골고루 반영한 데이터를 사용해야 편향을 줄이고 신뢰성을 확보할 수 있다.

세 번째는 체계적인 피드백 시스템 구축이다. 이를 통해 모델의 성능을 지속적으로 모니터링하고 문제가 발생했을 때 신속하게 대응할 수 있도록 한다.

위의 요소들을 고려하여 신뢰성을 확보한 시스템을 갖춘 AI 모델은 현실세계를 좀 더 정확히 반영하며, 문제가 생겼을 때 빠르게 대응할 수 있게 된다. 이를 통해 사용자의 만족도를 더욱 높일 수 있다.

위에서 언급한 내용을 한 마디로 정리하면, 결국 AI 모델 신뢰성 확보를 위해서 정확하고 편향되지 않은 데이터를 사용해야 하기에 '데이터 품질'이 중요하다. 하지만 실제 데이터를 수집, 정제하다 보면 많은 오류와 편향이 발생할 수 있기에, 반드시 데이터 검증을 통해 이러한 과정에서 발생하는 오류를 줄여야 한다.

## 4. 데이터 검증 방법

### 4.1 데이터 수집 및 전처리 과정에서의 검증

먼저 데이터 수집할 때 출처가 확실한 곳에서 데이터를 얻는 것이 중요하다. 대표적인 예가 책, 연구 논문, 코드 데이터, 뉴스 기사 등이다. 아카이브(arXiv.org)나 네이처(Nature) 등 출처가 분명한 학술지, 깃허브(Github) 등 코드 관리 플랫폼에 있는 소스 코드와 관련문서, 발행기관이 명시된 뉴스나 책 등은 출처가 어디인지 확실히 알 수 있다. 위키피디아, 블로그, 커뮤니티 게시글, 댓글 등 인터넷 텍스트 역시 출처가 확실하게 명기됐다면 데이터에 포함할 수 있다. 출처가 확실한 곳에서 데이터를 얻는 것은 데이터의 신뢰도를 확보하는 일차적 요소이지만 이것으로 신뢰도와 정확성이 보장되지는 않는다. 확실한 출처라 해도 출처 자체의 신뢰도와 데이터의 정확성 등을 면밀히 고려해야 한다.

출처가 확실한 데이터만 모았더라도 크고 작은 문제들은 여전히 존재한다. 이 때문에 데이터 정제과정을 반드시 거쳐야 한다. 정제과정으로는 사실관계 확인을 통한 오류 잡기, 누락된 데이터 확인, 중복 데이터 제거 등이 있다. 추가로 데이터 형식을 일관되게 만들고 단어나 문장 구조 등을 표준화해 데이터의 일관성을 높여야 한다. 여기서 끝이 아니다. 생성형 AI의 답변에선 편견과 함께 윤리적인 문제도 나타날 수 있다. 이에 전처리 과정에서 윤리적이지 않거나 유해한 내용(폭력적인 내용, 성인 콘텐츠, 유해 콘텐츠 등)이 담긴 데이터를 잘 처리해야 한다.

최근엔 개인정보 유출 문제가 수면 위로 떠오르고 있다. 따라서 데이터 정제 과정에서 이름, 전

화번호 등 개인정보를 제거하고 철저하게 익명화시켜야 한다. 강화된 개인정보 보호 조치를 취해야 생성형 AI를 통한 개인정보 유출 문제도 막을 수 있고, 사용자의 안전을 지킬 수 있다.
정리하면, 먼저 데이터수집 과정에서 출처가 분명한 데이터를 모은다. 이후 정제 과정에선 사실관계 확인과 표준화 작업을 진행한다. 동시에 편견을 줄이고 유해한 콘텐츠를 생성하지 않기 위한 여러 가지 노력이 필요하다. 이러한 작업을 통해 AI 모델이 좀 더 신뢰성 있는 모습으로 발전할 수 있고, 사용자들에게 안전하고 유용한 정보를 제공할 수 있다.

### 4.2 모델 학습 과정에서의 검증

특정 주제나 표현 또는 편견이 담긴 데이터가 많이 포함될 경우, 모델이 편향되게 학습할 수 있다. 그러므로 여러 편견이 존재하는 데이터를 제거하거나, 편견이 존재하지 않게 다양한 케이스의 데이터를 추가 수집할 필요가 있다. 이렇게 특정 주제에 편중되지 않도록 데이터 분포를 잘 조절해야 편견 없는 모델이 만들어진다.
수많은 데이터에서 사람이 일일이 편향성이나 오류를 찾아내는 것은 굉장히 시간과 비용이 많이 드는 일이다. 편향성 및 오류를 탐지할 수 있는 머신러닝 모델을 사용하면 빠르고 효율적으로 데이터 검증을 할 수 있다.
모델 학습과정에서의 검증은 편향된 데이터를 검증하는 것이다. 다른 측면으로 본다면, 모델 자체의 신뢰도를 확보하는 것이라고도 볼 수 있다. 이를 위해 모델의 설명성을 높이고 검증 결과를 올바르게 해석해야 한다. 모델의 답변이 어떠한 근거를 통해 이뤄졌는지를 이해하고 설명할 수 있으면 모델 성능 향상에 큰 도움이 된다. 이처럼 모델이 어떻게 결론을 냈는지 명확하게 설명하고 해석할 수 있도록 하는 기술을 Explainable AI라고 한다.
최근 신경망 기반 AI 모델들은 어떤 내부 메커니즘을 통해 결론을 내는지 알려져 있지 않다. 이 때문에 블랙박스 모델이라고도 불린다. Explainable AI는 이러한 모델 내부 작동 방식을 이해하는 연구다. 이는 모델 자체의 신뢰성을 확보하고, 모델 개선과 디버깅을 도와주며, 결과적으로 사람이 AI 모델을 더 잘 이해할 수 있도록 한다.
최근 앤트로픽은 자체적으로 개발한 모델 클로드3 소넷(Sonnet)의 내부 작동을 이해하는 데 성공했다고 발표했다[17]. 이러한 연구들이 지속되면 AI 모델 내부를 더 자세히 파악할 수 있고, 모델의 예측 및 답변을 잘 설명할 수 있다. 이는 한층 더 신뢰성이 높은 AI 모델 제작으로 이어진다.

### 4.3 모델 평가 및 배포 과정에서의 검증

학습이 끝난 후 에는, 해당 모델이 제대로 된 성능을 발휘하는지 다양한 지표를 통해 알아봐야 한다. LLM 연구자들은 흔히 Open LLM Leaderboard[18]라는 것을 통해 LLM의 일반적인 능력을 평가하고 어느 모델이 더 좋은지 서로 비교한다. 여기서 LLM의 일반적인 능력이라 함은 답변 생성의 정확도, 유창성, 일관성, 사실 부합성이 얼마나 높은지, 여러 NLP 태스크(감정분석, 언어이해, 추론, 번역, 요약, 질의응답)들을 얼마나 잘하는지를 뜻한다. 이런 요소들을 평가하기 위해 각각의 벤치마크 데이터 셋이 있고, 이를 통해 각 LLM이 어떤 영역을 잘하고 못하는지를 알 수 있다.
이러한 여러 평가들을 통해 AI 모델의 신뢰성을 알아보고 부족한 부분을 더 보강할 수 있다. 한편 위에서 알아본 평가 지표들은 대부분 LLM의 능력을 평가하는 것이고, 안전성을 알아보는 평

가지표들도 있다. 유해성, 편견, 윤리성, 강건성, 데이터 거버넌스 등의 요소들은 모두 안전성을 평가하는 항목들이다. 이를 통해서 LLM이 폭력적인 표현, 욕설, 성적인 표현을 하는지 또는 인종, 성별 간의 편견, 고정관념 등을 가지고 있는지 판단할 수 있다.
이러한 검증을 마치고 나면, 실제 배포 환경에서 지속적으로 모니터링하고 관리하는 시스템이 필요하다. 문제가 생길 때 이를 바로 알아차리고 필요한 대응을 빠르게 할 수 있도록, 피드백 시스템을 갖추고 끊임없이 데이터를 검증해야 한다. 필요에 따라 데이터를 주기적으로 업데이트하며, 모델을 재학습하고 배포한다면 사용자들에게 신뢰도 있는 모델을 제공할 수 있다.

## 5. 맺음말

원고 초반에 언급했듯이, 생성모델을 학술적으로 이야기하면 데이터 셋의 분포를 학습하는 모델이라 할 수 있다. 즉 좋은 모델이라는 것은, 이상적으로 데이터 셋의 분포를 제대로 학습했다고 가정한다면, 결국 '좋은 데이터 셋'에서 나온다고 할 수 있다. 컴퓨터과학 분야에서 통용되는 'Garbage in, Garbage out'과 일맥상통하는 의미다. 그렇기에 체계적인 데이터 검증은 신뢰성 높은 AI 모델을 만들기 위한 필수요소다. 위에서 설명한 데이터 수집, 정제, 모델 학습, 평가, 배포 등 모든 과정에서 데이터 검증이 잘 녹아 들어간다면, 모델의 신뢰성과 안전성을 담보할 수 있다.
신뢰성 확보를 위한 향후 연구 방향을 짧게 이야기하며 글을 마치겠다.
첫 번째는 자동화된 데이터 검증 시스템 개발이다. 위에서 언급한 데이터 검증 프로세스에는 사람의 노력이 많이 들어간다. 특히 데이터 수집 이후 오류나 편향이 있는지 혹은 윤리적인 문제가 없는지 검토하는 부분에서 사람의 의견을 반영해야 하기에 아직은 쉽지 않다. 하지만 이런 자동화가 빠르게 이뤄져야 높은 데이터 품질을 유지하면서 AI의 신뢰성을 확보할 수 있다.
두 번째는 다양한 데이터의 출처와 문화적 배경을 반영한 데이터 셋 구성이다. 아직까지도 LLM에 사용되는 데이터는 대부분 영어로 구성됐으며, 서구 문화권의 내용을 많이 담고 있다. 한국적인 내용을 물어볼 때 부정확한 답변을 하는 것은 어쩌면 당연할 수도 있다. 향후에는 여러 문화권을 함께 고려할 수 있도록, 다양한 나라의 데이터를 포함해야 데이터 부족에서 오는 편향을 줄일 수 있을 것이다. 이와 관련, 최근 셀렉트스타에서 대국민 설문조사를 바탕으로 LLM 벤치마크를 구성한 연구가 있다. 한국의 문화적 배경을 고려한 상당히 의미 있는 연구다[19].
세 번째는 윤리적 AI 개발을 위한 세계적인 합의다. 작년 10월 30일, 미 바이든 정부에서 첫 AI 행정명령을 발표했다[20]. 주요 내용은 'AI 안전성 확보를 위한 표준화된 평가 마련'과 '개인정보 보호 강화'다. 이와 같이 AI 신뢰성과 안전성을 지키기 위한 노력들에 대해, 연구자, 개발자, 기업은 물론 각국 행정부가 동참해 세계적 합의를 이끌어 내야 한다. 이는 앞으로 AI 시대를 살아가는 데 있어 굉장히 중요할 것이다.

[참고문헌]

[1] https://chatgpt.com/
[2] https://gemini.google.com/

[3] https://claude.ai/chats
[4] https://openai.com/index/dall-e-3/
[5] https://stability.ai/news/stable-diffusion-3
[6] https://midjourney.co
[7] https://openai.com/index/sora/
[8] https://www.nature.com/articles/d41586-024-00779-1
[9] https://arxiv.org/abs/1706.03762
[10] https://arxiv.org/abs/1810.04805
[11] https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
[12] https://llama.meta.com/llama3/
[13] https://arxiv.org/abs/2206.07682
[14] https://openai.com/index/hello-gpt-4o/
[15] https://deepmind.google/technologies/gemini/pro/
[16] https://arxiv.org/abs/2203.15556
[17] https://www.anthropic.com/research/mapping-mind-language-model
[18] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[19] https://arxiv.org/abs/2402.13605
[20] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy development-and-use-of-artificial-intelligence

※ 출처: TTA 저널 제213호