2023 Vol.04
차세대리포트

인공지능 언어모델의 기술 변천사와 미래 가능성

KAST 한국과학기술한림원
The Korean Academy of Science and Technology

펴낸곳
한국과학기술한림원
031)726-7900

펴낸이
유 욱 준

발행연월
2023년 10월

홈페이지
www.kast.or.kr

기획·편집
한국과학기술한림원 정책연구팀

컨텐츠
정현섭 과학기술분야 전문 작가

디자인·인쇄
경성문화사
02)786-2999

이 보고서는 복권기금 및 과학기술진흥기금의 지원을 통해 제작되었으며,

발간사

'대한민국의 미래를 이끌어갈 젊은 과학자들의 지식과 경험을 활용하여 과학기술 분야의 발전, 그리고 국가와 사회의 미래를 위해 기여할 수 있는 방법은 없을까?' 지난 2018년부터 발간되고 있는 한국과학기술한림원의 차세대리포트는 이러한 고민으로부터 시작된 노력의 결과물이다.

우수한 젊은 과학기술인 그룹인 '한국차세대과학기술한림원(Young Korean Academy of Science and Technology, Y-KAST)' 회원들과 연구 현장 최일선에서 활약하고 있는 최고의 젊은 과학자들이 중심이 되어 발간하고 있는 차세대리포트는 그동안 '양자기술'이나 '수소사회'와 같은 최신 과학기술 관련 이슈는 물론 '젊은 과학자를 위한 R&D 정책', '과학자가 되고 싶은 나라를 만드는 방법', '대학의 미래' 같이 과학기술과 관련된 다양한 주제를 다루어 왔다.

올해로 벌써 발간 5주년을 맞이한 차세대리포트는 다양한 이슈에 대해 새로운 시각과 신선한 의견을 전하고자 노력해 왔으며, '과학기술 분야 최신 동향'과 '사회적 이슈 및 현안'이라는 두 마리 토끼를 잡기 위해 주제의 선정에서부터 발간에 이르기까지의 모든 과정을 치열한 고민을 통해 진행하고 있다.

바둑을 정복한 알파고(AlphaGo)의 등장이 인공지능 발전의 기폭제가 되었던 것처럼, ChatGPT를 비롯한 거대 언어모델의 등장은 인류에게 신선한 충격을 주었고, 이러한 인공지능 언어모델의 급격한 발전으로 사회는 큰 변화를 겪고 있다. 언어모델의 활용 가능성이 높아지면서 인류는 새로운 도약을 맞이하고 있지만, 언어모델의 발전과 보급에 따라 파생되는 여러 사회적 부작용들에 직면하고 있기도 하다. 언어모델을 바람직한 방향으로 활용하기 위해서는 이러한 부작용을 잘 인지하고 예측할 수 있어야 하며, 이를 극복하기 위한 새로운 기준과 합의가 필요한 시점이다.

이번 차세대리포트에서는 20세기 중반부터 오늘날에 이르기까지 인공지능 언어모델의 변천사를 기술적 관점에서 살펴보며, 향후 언어모델의 발전과 활용을 위해 필요한 것은 무엇인지 알아볼 예정이다. 또한, 언어모델이 야기할 수 있는 부작용과 이를 극복하기 위해 필요한 제도적 방안에는 무엇이 있는지 살펴보고자 한다.

2023년 10월
한국과학기술한림원 원장
유 욱 준

참여자 소개

김미현 | 가천대학교 약학대학 교수

인공지능을 활용한 신약 설계 연구자로 머신러닝, 화학정보학, 유기화학, 의약화학, 화학생물학의 연구방법을 융합하여 희귀 구조 약물의 역탐색(inverse screening), 역설계(inverse design) 및 다중 파라미터한 의약품 연구개발을 활발히 수행하고 있다.

김윤수 | POSTECH 컴퓨터공학과 교수

기계 번역, 음성 번역 및 자연어의 다국어, 다감각 응용분야에 있어 효과적인 학습 방법과 효율적인 제품화에 대해 연연구하고 있다.

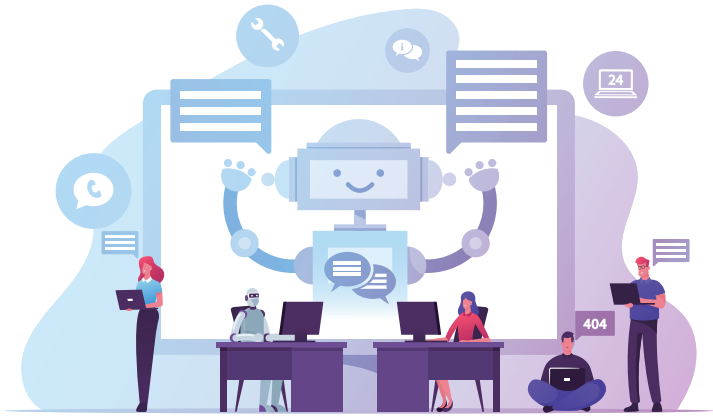
서민준 | KAIST 김재철AI대학원 교수

자연어 처리(NLP) 및 대형언어모델 연구를 하고 있다. 특히 언어모델 내에 지식이 저장되는 방식을 탐구하고 필요한 지식을 활용하여 업무 수행을 위한 플래닝 및 논리적 추론을 하는 방법론을 연구하고 있다.

조민수 | POSTECH 컴퓨터공학과 교수

계산적 시각지능 연구, 컴퓨터비전 분야 전문가로 인공지능에 관계와 구조에 대한 체계성을 부여하는 것이 주된 연구 관심사이다. 현재 시각 요소들 간의 관계추론과 학습 문제들을 중심으로 영상정합, 대칭탐지, 물체발견, 비디오해석, 물체조립 문제 등에 대한 연구를 활발히 수행하고 있다.

CONTENTS



ChatGPT의 등장은 사람들에게 큰 충격을 안겨주었다. 인간이 입력한 자연어를 이해하고, 문장을 자연스럽게 생성하는 능력이 이전 언어모델과 비교해 현저한 발전을 이루었기 때문이다. 묻는 말에 꽤 높은 정확도로 대답하고, 간단한 작문과 요약은 기본이며, 개발자의 프로그래밍을 완성해주기도 한다. 심지어 시나 소설도 창작하여 써준다. 24시간 대기하면서 언제 어디서든 질문에 답변을 해주고 꽤 전문적인 분야까지도 섭렵하고 있는데다가 '더 쉽게 표현해줘', '세 문장으로 요약해줘' 등 사람이 수행하기도 쉽지 않은 다소 무리한 요구도 그럭저럭 해내는 모습을 보여주었다. 이제는 각 문장마다 출처를 달아주기도 하고, 관련 이미지를 생성하거나 그림까지 그려주는 수준에 이르렀다.

이처럼 매우 똑똑한 언어모델의 등장으로 사회는 잠시 혼란을 겪었다. ChatGPT 사용을 금지하는 기업이나 학교들이 있는가하면, 반대로 언어모델을 적극적으로 활용하는 사례들도 함께 늘어났다. 저작권과 같은 법적 문제와 각종 윤리적 문제들이 이슈화되기도 했으며, 기업 내에서는 ChatGPT 사용에 따른 보안 문제가 제기되기도 했다. 사회적인 파장과 혼란이 있었지만 사회 전반적으로 거대한 기술에 서서히 적응하고 활용하는 방향으로 움직이는 모습이 나타나기 시작했다.

오늘날 사람들은 마치 오피스 프로그램이나 계산기를 다루듯 ChatGPT를 비롯한 거대 언어모델을 사용하고 있다. 이로 인해 IT 뿐만 아니라 제조업, 미디어, 교육 등 다양한 산업군에서 그간의 패러다임을 바꾸어놓을 정도로 큰 변화가 생겨나고 있다.

긍정적으로든 부정적으로든 언어모델의 사용으로 인해 많은 현상들이 파생되고 있으며, 앞으로도 새로운 변화들이 끊임없이 예견되고 있다.

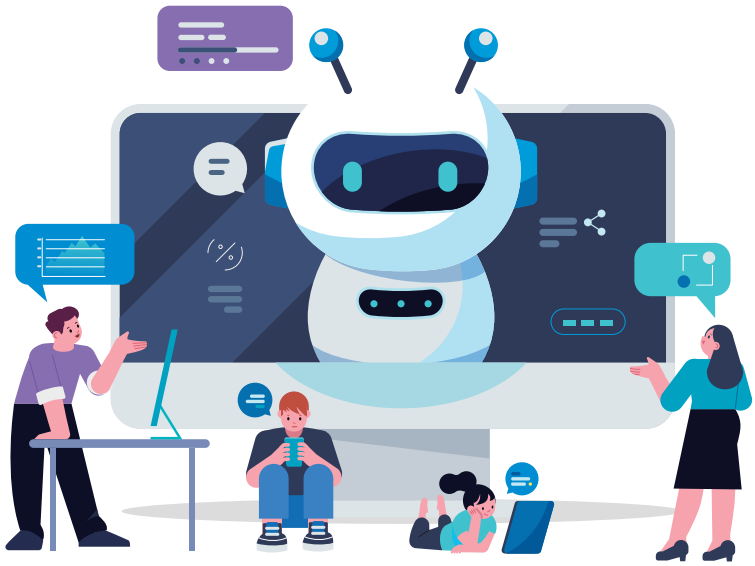
거대 언어모델이 생성하는 문장의 문법이 꽤 정확하다고 해도 아직은 제공하는 정보가 완벽하다고 확신할 수 있는 단계는 아니다. ChatGPT-3.5의 경우 사실이 아니거나 불확실한 내용을 그럴듯한 작문으로 마치 사실이나 진실인 것처럼 표현하는 경우도 매우 많았다. 데이터 학습을 기반으로 하는 **거대 언어모델의 신뢰성 문제**는 여전히 숙제로 남아있다.

그리고 언어모델의 장점 중 하나인 **높은 생산성 또한 양날의 검**으로 작용하고 있다. 마치 외래종 생물이 토종 생물을 잠식하며 걷잡을 수 없이 번식하는 것처럼 언어모델이 만드는 텍스트들은 충분히 검수되지 못한 채, 기존에 존재하던 좋은 텍스트들을 온라인상에서 빠르게 잠식해갈지도 모른다. 그리고 어느새 통제할 수 없는 수준에 이르러 심각한 사회적인 부작용을 초래할 수도 있다. 예를 들면, 사실이 아닌 내용이 급격하게 확산돼 마치 그것이 사실인양 대중들에게 받아들여지거나, 부정확한 텍스트 데이터들이 양산되면서 학습을 기반으로 하는 언어모델들의 성능이 오히려 저하될수도 있는 것이다.

이러한 부작용을 고려해보면 언어모델이 단순히 개인이나 특정 분야가 아닌, 국가 차원에서 관여해야 하는 수준의 문제를 야기할 수 있음을 충분히 예상할 수 있다. 이미 대중에게 공개된 거대 언어모델의 사용에 제한을 두기는 쉽지 않을 것이다. 따라서 사회에 이로운 방향으로 적절히 사용하되, 부작용을 예방할 수 있는 방안을 마련해야 한다. 이를 위해 거대 언어모델을 **사회적인 측면에서뿐만 아니라 기술적 관점에서 이해함으로써 보다 근본적인 대응책을 모색**할 필요가 있다.

본 차세대리포트에서는 언어모델이 오늘날의 발전에 이르기까지 어떻게 변모해왔는지, 그리고 이러한 변천사에 어떠한 배경들이 있었는지를 **기술적인 관점**에서 살펴보고자 한다. 아울러 ChatGPT를 비롯한 거대 언어모델들의 최신 기술 동향과 향후 발전 방안에 대해 모색하는 한편, 거대 언어모델의 발전과 활용을 위해 필요한 **정책**에 대해 제안하고자 한다.

1 언어모델의 기술 변천사



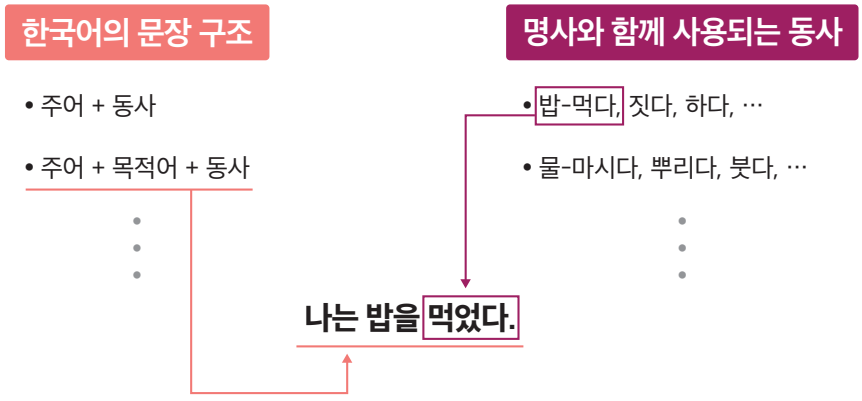
최근 ChatGPT를 통해 일반 사용자들에게 널리 알려진 '언어모델'은 사실 1970년대부터 활발히 연구되었다. 언어모델은 초창기 인공지능 시스템에 적용되면서 예전부터 우리의 일상생활에 이미 널리 쓰이고 있던 기술이다. 본 파트에서는 언어모델 기술이 어떻게 변화해왔는지 그 시작부터 지금까지 주요한 변곡점들을 짚어보며 각각의 의의를 되새겨 보고자 한다.

[20세기 중반] 규칙 기반 모델: '언어'를 컴퓨터가 어떻게 설명할 것인가?

인공지능은 컴퓨터가 발명되면서부터 그 궤를 같이 해왔다. 그 중 자연어를 모델링하는 것은 인공지능의 초창기부터 큰 관심사였다. 인간의 '언어'를 어떻게 컴퓨터가 설명할 것인가? 이는 곧 **'어떤 문장이 인간의 언어로 쓰인 문장인가?'**를 파악하는 문제로 귀결되었다.

이를 해결하기 위해 주어진 문장에 언어의 '유창성' 점수를 부여하는 계산법들이 고안되었다. 20세기 중반에는 인공지능 초기의 패러다임인 '규칙 기반(rule-based) 방법론'이 사용되었다. 이는 한 언어를 컴퓨터로 설명하기 위해 해당 언어의 어휘 및 문법 규칙들을 언어학자가 일일이 나열하여 저장해두고, 어떤 문장이 그 언어에 잘 맞는지 규칙을 검사해보는 방법이다. 예를 들어 한국어의 주어 뒤에는 목적어가 나와야 하고, 특정 목적어 뒤에는 이에 맞는 동사가 나와야 한다는 식이다(그림 1).

그림 1 규칙 기반 언어 모델링의 예



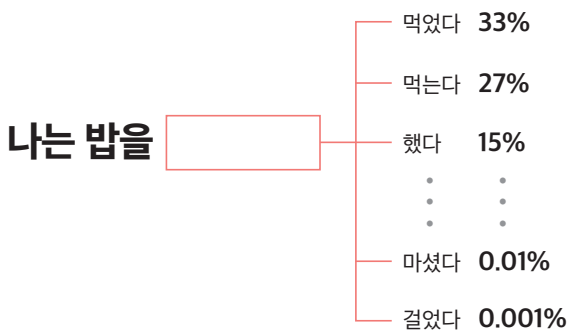
[20세기 후반] 등장 횟수 기반 모델: '언어'를 통계적으로 어떻게 설명할 것인가?

하지만 규칙을 기반으로 언어를 설명하는 방법은 실질적으로 구현하는데 큰 어려움이 있었다. 인간의 언어를 설명하는 규칙이 너무 방대하여 모든 규칙을 빠짐없이 구축하려면 너무 많은 시간이 소요되었다. 이를 수행할 수 있는 언어학자들도 많지 않았다.

따라서 전문 인력 고용을 최소화하면서도 주어진 데이터에서 의미 있는 수치들을 뽑아내어 모델을 만드는 '통계학적 기계학습(statistical machine learning)'이 20세기 후반부터 도입되기 시작했다. 이는 텍스트 데이터에서 특정 패턴이 얼마나 많이 등장했는지 그 횟수를 세어 확률로 저장해두고, 어떠한 문맥이 주어졌을 때 그 다음 위치에 무슨 단어가 등장할 확률이 높은 지 점수를 부여하는 방법이다. 예를 들면, '나는 밥을'이라는 두 단어로 된 문맥이 주어질 경우, 그 뒤에는 '먹었다'라는 단어가 등장할 확률이 매우 높다(그림 2). 반대로 '마셨다'라는 단어가 배치된 경우, 그 문장의

점수는 매우 낮고, 이는 유창한 한국어 문장이라고 말하기 어렵다고 할 수 있다. 이러한 '등장 횟수 기반(count-based) 모델'은 통계학적 방법론으로 구축된 음성인식 및 번역 시스템에서 인식된 결과가 자연스러운 문장인지 아닌지 판단하는데 핵심적인 역할을 했다.

그림 2 등장 횟수 기반 언어 모델링의 예



[2000년대 초반] 벡터 공간 표현 모델: '유창성'을 넘어 '의미'를 담는 시도

등장 횟수 기반 모델은 데이터에서 쉽게 계산할 수 있고 매우 직관적이라는 장점 때문에 현대 음성인식 시스템에까지 널리 사용되었다. 하지만 주제 분류, 질의응답, 글 바꿔 쓰기 등 다양한 자연어 처리 문제들에 적용되기에는 한계가 있었다. 등장 횟수가 높다는 것이 '많이 쓰이는 구문 패턴'만을 뜻할 뿐, 복잡한 의미나 미묘한 뉘앙스, 단어들 간의 관계 등 '언어'를 구성하는 다른 많은 요소들을 설명하지 못하기 때문이었다.

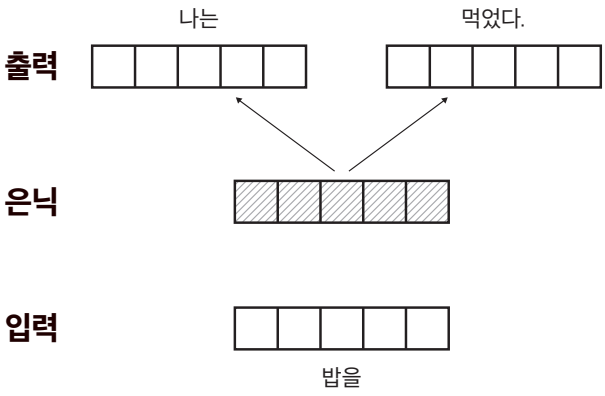
인공지능의 전반적인 연구 방향은 21세기에 접어들며 신경망(neural network) 구조로 집중되었다. 신경망은 모델을 충분히 크고 복잡하게 설계한다면 이 세상에 존재하는 모든 문제를 모델링할 수 있다는 강력한 이론을 바탕으로 하고 있으나, 20세기 컴퓨터의 계산 능력으로는 고성능을 보장할 만한 큰 신경망의 학습과 실행이 불가능했다. 하지만 하드웨어의 발전으로 신경망 모델을 사용하기가 수월해지면서 이를 이용해 언어모델을 만드는 시도들이 시작되었다.

문맥이 주어지면 그 주변 단어를 예측한다는 기본 문제 설정은 등장 횟수 기반 모델과 같으나, 신경망을 이용하면 그 결과를 단순한 횟수가 아닌 수학적 공간의 벡터로 저장할

수 있었다(그림 3). 여러 가지 실험을 통하여 이 벡터가 단어의 의미론적 정보들을 많이 포함하고 있음이 밝혀졌고, 이러한 벡터들을 변형, 조합, 비교함으로써 '의미'가 중요한 많은 문제들을 해결하는 데 큰 도움이 되었다.

그림 3 단어 벡터 학습용 신경망 모델의 예(skip-gram)

은닉층의 벡터가 단어 '밥을'의 여러가지 정보를 내포하게 된다.



[2000년대 후반] 순환 신경망 모델: "문맥에 제한을 두지 않다"

초기 신경망 언어모델은 등장 횟수 기반 모델과 같이 정해진 몇 단어만을 문맥의 입력으로 받고 있었다. 이는 모델의 크기를 줄여 학습과 추론 시간을 빠르게 하기 위해서 어쩔 수 없는 선택이었다. 하지만 짧은 문맥에 제한되다 보니 문장의 길이가 긴 경우 문장의 뒤쪽에서는 문장의 앞단에 주어지는 중요한 단어들을 고려하지 못한다는 한계가 있었다(그림 4). 나아가 현재 문장을 넘어 그 이전 문장, 또는 지금까지의 **문서 또는 대화 전체를 염두에 둔 통일성 있는 언어 모델링**에 대한 관심과 수요가 높아졌다.

그림 4 제한된 문맥 길이가 언어 모델링의 성능을 저하시키는 예

문장 초반의 '밥을'을 고려하면 빈 칸에 들어갈 단어는 '먹었다'가 적절하지만,
직전 5개 단어를 문맥으로 사용하는 모델은 빈 칸의 단어를 알맞게 예측하기 어렵다.

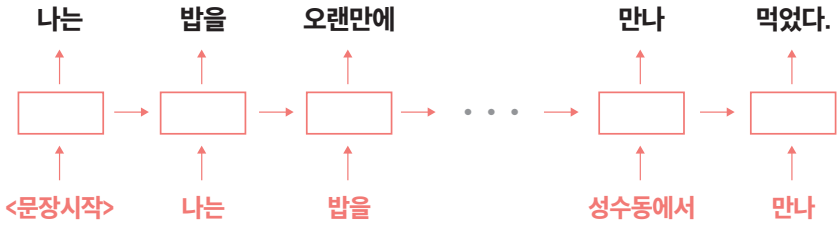
나는 밥을 오랜만에 만난 고등학교 친구들 다섯 명과 함께 성수동에서 만나 ?

언어모델의 문맥 길이

긴 시퀀스(sequence)를 효율적으로 다룰 수 있는 '순환 신경망(Recurrent Neural Network, RNN)'구조는 이미 1990년대에 많이 연구되었으며, 2007년경부터 자연어 문장을 모델링 하는 데에 본격적으로 사용되었다. 이 구조의 핵심은 신경망이 한 단어씩 순서대로 처리하되, 처리한 결과를 '내부 상태(internal state)'라는 표현 벡터로 저장해두고 그 다음 단어를 처리하는 데 재사용하는 것이다(그림 5). 이 내부 상태 벡터는 이론적으로 현재까지 처리된 모든 단어의 정보를 압축하여 포함할 수 있었으나, 처리한 단어가 많아질 경우 오래 전에 처리했던 정보는 희미해지는 단점도 있었다. 이를 극복한 '장단기 메모리(Long Short-Term Memory, LSTM)'의 경우 지금까지 저장해 온 상태 벡터가 현재 단어를 처리하는 데 필요한 지 아닌 지를 스스로 계산할 수 있도록 함으로써 상태 벡터가 필요할 때만 갱신되도록 하여 긴 문장도 효과적으로 처리할 수 있는 가능성을 열었다. LSTM 구조는 음성 인식, 손글씨 인식 등에서 기존 방법론의 성능을 비약적으로 향상시키는 데 결정적인 역할을 하였으며, 2010년대부터 상용 시스템에 널리 쓰이기 시작했다.

그림 5 순환 신경망 언어 모델링의 예

문장의 처음에 계산된 정보가 문장의 끝까지 벡터 공간을 통해 전달된다.



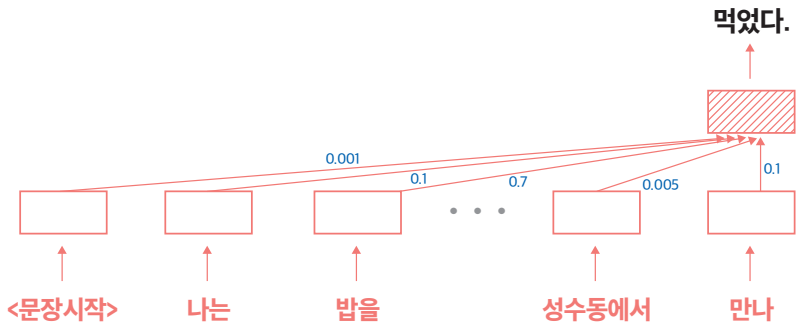
[2010년대 중반] 어텐션과 트랜스포머: "궁극의 유동성을 지닌 모델 구조"

순환 신경망 모델의 성공은 다음 단어를 예측하는 데 이전 문맥을 길게 고려하는 것이 얼마나 중요한 것인 지를 연구자들에게 각인시켜 주었다. 하지만 순환 처리의 특성 상 내부 상태 벡터는 그 크기가 문장의 길이와 상관없이 고정되어 있었다. 이는 긴 문장의 경우 많은 단어들의 정보를 제한된 공간에 저장하기가 힘들다는 것을 뜻한다. 실제로 순환 신경망을 이용해 만든 번역 모델은 입력 문장이 매우 긴 경우 이를 정해진 크기의 벡터에 저장했을 때 이를 번역한 출력 문장의 품질이 매우 좋지 않았다.

그렇다면 모델이 단어 예측에 활용할 내부 상태의 크기를 하나로 정하지 않고 문장의 길이에 따라 유동적으로 변화시키면 되지 않을까? 이러한 직관을 구현한 것이 바로 '어텐션(attention)' 구조이다(그림 6). 어텐션은 내부 상태를 하나의 벡터로 정의하지 않고, 단어 하나 당 하나의 벡터를 할당하여 다섯 단어 문장은 벡터 다섯 개, 열 단어 문장은 벡터 열 개 등으로 표현하는 것을 기본으로 한다. 하지만 여러 개의 벡터를 모두 재사용하여 현재 단어를 처리할 경우 너무 많은 시간이 소모되므로, 이를 하나의 벡터로 평균을 내어 합치는 과정을 거친다. 이 때 각 벡터의 가중치(attention weight)를 계산하여 어떤 경우에는 멀리 떨어진 단어를 더 참조하고, 어떤 경우에는 가까운 단어를 더 고려하여 문맥의 상황에 맞게 활용할 수 있도록 설계하였다.

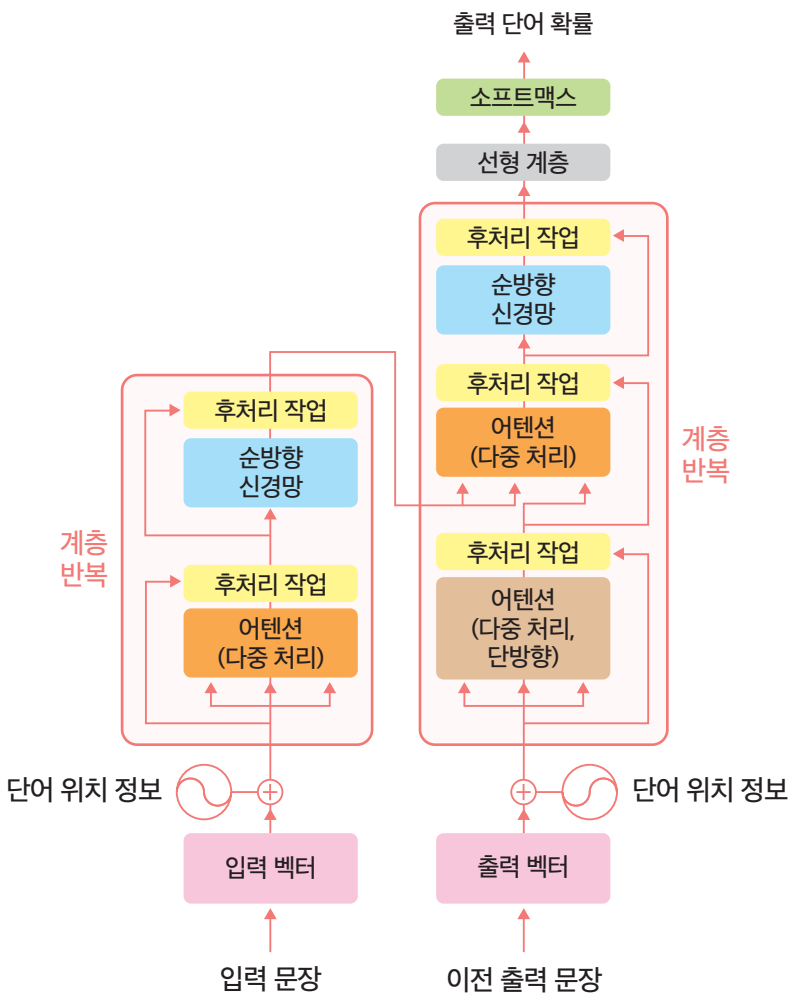
그림 6 어텐션 기반 언어 모델링의 예

직전의 모든 단어 벡터를 참조하되, '밥을'에 높은 가중치를 두어
'먹었다'를 적절히 예측한다.



초기의 어텐션은 LSTM과 함께 쓰였으나 점차 그 효용성이 부각되면서 신경망의 모든 계층을 어텐션만으로 구성하는 '트랜스포머(Transformer)' 구조가 크게 활용되었다(그림 7). 이는 기계 번역 분야에서 새로운 패러다임을 제시하며 탁월한 성능을 보였고, 점차 이미지, 비디오, 음성 등의 다른 데이터를 처리하는 분야에도 적용되었다.

그림 7 트랜스포머 모델의 구조



[2010년대 후반] GPT의 등장: "언어의 '채점'에서 언어의 '생성'으로…"

트랜스포머는 신경망 내부의 노드(node)들이 어떻게 연결되었는지를 보았을 때 기존의 순환 신경망, 합성곱 신경망(convolutional neural network) 등의 구조들과 비교하여 가장 제약이 적고 일반적인 형태를 띠었다. 이는 곧 모델이 표현할 수 있는 정보의 자유도가 매우 높음을 의미했고, 이에 착안한 OpenAI의 과학자들은 트랜스포머의 구조를 더 이상 바꾸지 않고 크기와 학습 데이터만 무한정 늘려가며 그 한계를 시험하는 연구를 진행하기 시작했다. 이것이 현재 주목 받고 있는 ChatGPT의 근간이 되는 'GPT(Generative Pre-trained Transformer)'의 시작이다.

GPT는 그 이름에서도 드러나듯이 텍스트 생성(text generation)을 목적으로 하고 있다. 이는 기존의 언어 모델들이 주로 주어진 텍스트에 점수를 부여하여 '이 문장이 유창한 문장인가?'를 판단하는 수동적인 용도로 쓰인 것과 달리, 문장의 일부를 주면 나머지를 완성하거나 지시문(prompt)을 주면 그에 맞는 대답을 내놓는 능동적인 용도를 본격적으로 표방하고 있었다. 이 생성 능력은 모델의 크기와 사용한 데이터가 커질수록 증가하여 인간이 쓴 문장과 구별하기 힘들 정도로 자연스러운 문장을 만들 뿐만 아니라 여러 가지 심화된 자연어 처리 작업들을 더 쉽게 해주는 가교 역할을 하였다.

GPT-1은 그 모델 매개 변수들을 텍스트 함의나 질의응답 등 고차원 작업 데이터를 이용해 미세조정(fine-tuning) 함으로써 해당 작업들에서 당대 최고 성능을 보여주었다. GPT-2는 모델 크기와 데이터의 양을 모두 10배 이상 증폭하여 학습되었다. 특히 GPT-1이 주로 출판된 책을 학습 데이터로 사용했다면, GPT-2는 인터넷에 있는 다양한 웹페이지 및 게시판의 텍스트들을 활용하였다. 그 결과 단순한 문장 생성 이상으로 여러 가지 작업들을 자체적으로 수행하는 능력을 보여주었다. 예를 들어 학습 과정에서 번역 작업을 지시받지 않았음에도 불구하고, 학습 후 모델을 사용할 때 번역을 해달라고 지시하면 입력된 문장의 번역문을 생성하려고 시도하는 것이다. 물론 이렇게 특정 작업에 대한 추가 학습이 없는 경우 그 작업에 대한 성능은 매우 낮은 편이나, 적어도 해당 작업의 의도에는 맞는 답을 내놓는다는 것은 매우 시사하는 바가 컸다. 이는 인공지능 연구자들과 미래학자들이 줄곧 이야기해 온 '인공 일반 지능(Artificial General Intelligence, AGI)', 즉, 다양한 상황에 두루 적용할 수 있는 만능 모델의 가능성을 보여주었기 때문이다.

이에 고무된 OpenAI는 2020년, GPT-2보다 100배 이상 큰 GPT-3을 출시하게 된다. 이 버전은 일반 지능의 역량이 극대화되어 번역과 요약은 물론 소설·시 창작, 영어 시험 문제 풀이, 프로그램 코드 생성, 각종 콘텐츠 만들기 등 그야말로 텍스트로 된 모든 작업이 가능하다는 캐치프레이즈가 나올 정도였다. GPT-3은 여기서 더 나아가 흔히 생각할 수 있는 작업뿐만 아니라 사용자가 원하는 맞춤 작업을 수행할 수 있는 '문맥 내 학습(in-context learning)'을 지원하기 시작했다. 이는 모델을 추가 학습 시키지 않고도 지시문에 사용자가 수행하고 싶은 새로운 작업의 예시 및 답안을 제시하면 이를 즉석에서 파악하여 그 작업을 수행할 수 있는 것이다(그림 8). 수많은 연구자들이 놀랐던 것은 이 모든 능력이 학습 과정에서 특별한 조치를 취하여 얻은 것이 아니라 그저 모델과 데이터의 크기를 늘려서 가능하게 된 것이라는 사실이다.

그림 8 문맥 내 학습의 예

입력

- 문장: 나는 밥을 먹었다. / 논조: 중립
- 문장: 나는 밥을 맛있게 먹었다. / 논조: 긍정
- 문장: 나는 밥을 하기 싫다. / 논조: 부정
- 문장: 나는 밥을 먹고 싶다. / 논조: ?

출력

- 긍정

[2020년대 초반] GPT의 진화: "연구 시험 모델이 일반 사용자 곁으로…"

GPT-2까지만 해도 언어모델을 일반적인 텍스트 생성 및 문제 해결 도구로 쓴다는 것은 이상적인 목표에 가까웠다. 하지만 GPT-3을 통해 이것이 현실에서 가능한 것임이 증명되자 OpenAI는 그동안 모델을 거대화하는 연구에만 집중했던 것에서 벗어나 이를 수익화하는 전략을 실행하기 시작한다. 이는 모델을 일반 사용자들의 요구에 더욱 맞추어 최적화하고, 심각한 문제가 될 수 있는 사례들을 방지하여 안정화를 도모하고자 하는 것이었다. 이러한 목적을 달성하기 위한 가장 빠르고 효과적인 방법은 놀랍게도 **모델의 거대화나 알고리즘의 고도화가 아닌 '인간'의 개입**이었다.

GPT-3.5는 GPT-3에 추가 학습을 진행한 모델로, 생성 모델을 사용자가 쓸 때 입력할 법한 **지시문과 그에 맞는 출력을 인간이 작성하고 검토한 뒤 학습 데이터로 사용**하였다. GPT는 기본적으로 무작위 텍스트로 학습하여 여러 작업에 대한 성능을 고르게 달성하는 것을 목표로 하였으나, 연구 목적을 넘어 제품화를 할 때는 사용자의 만족도를 높이기 위해 특정 작업들에 특화시키는 방법을 다시 활용한 것이다. GPT-3.5에 대화 인터페이스를 붙인 것이 ChatGPT인데, 이를 위해 **추가 학습 데이터를 채팅 형식으로 변환하여 학습함으로써 사용자와 서비스 간의 상호 작용 경험을 향상**시켰다.

사람의 개입을 통해 모델을 최적화시키는 또 다른 핵심 기술로 GPT-3.5는 '인간의 선호도로부터의 강화 학습(Reinforcement Learning from Human Feedback, RLHF)'을 진행하였다. 이는 인간이 지시문과 답을 모두 작성하는 것이 아니라 지시문에 해당하는 답을 GPT 모델이 여러 개 생성하고, 그 품질을 인간이 비교하여 순위를 매기는 것을 골자로 한다. 이 데이터로 GPT 모델의 답을 채점하는 보상 모델(reward model)을 만들어 추후 모델이 생성한 것이 얼마나 학습에 영향을 줄지 결정하게 된다. RLHF는 특히 비속어, 성적 표현, 혐오 발언 등을 생성하는 것을 줄이는 데 효과적이었다.

GPT-3.5를 탑재한 ChatGPT 출시 이후 인공지능을 바라보는 일반 사용자들의 시각이 매우 긍정적으로 바뀌었고, 업무 및 교육 현장에서 언어모델을 적극적으로 사용하여 생산성을 높이는 것은 흔한 일이 되었다. ChatGPT를 이용한 여러 가지 응용 서비스들이 등장하고, 인공지능의 한계를 시험하기 위해 '연구'로 진행되었던 GPT 모델의 기술적 상세 내용은 회사의 이익을 위해 더 이상 공개되지 않게 되었다. OpenAI는 이후 GPT-4를 발표하였는데, 이 또한 모델 크기 및 학습 방법 등이 자세하게 공개되지 않았다. GPT-4는 텍스트뿐만 아니라 이미지 등을 입력으로 받아 함께 이해하도록 학습되었고, 이는 언어모델의 메커니즘이 점점 인간의 언어뿐만 아니라 여러 가지 정보를 처리하는 방향으로 발전할 것임을 보여준다.

표 1 GPT 모델 버전 비교

	GPT-1	GPT-2	GPT-3	GPT-3.5	GPT-4
출시년도	2018	2019	2020	2022	2023
매개변수	1.2억개	15억개	1,750억개	1,750억개	미공개
학습데이터 크기	4.8GB	40GB	570GB	570GB + a	미공개
입력길이 (토큰 수)	512	1,024	2,048	4,096	32,768
주요 기술	Fine-Tuning	Zero-Shot Learning	In-Context Learning	RLHF	Multimodal

II 언어모델의 현재와 미래

① 거대 언어모델의 최신 기술 동향

메타의 LLaMA 프로젝트: "GPT-3와 견줄만한 오픈소스 언어모델이 등장하다"

2020년 5월 GPT-3가 등장하고 논문이 공개됐을 때, 학계는 큰 충격을 받았다. GPT-3는 1,750억 개의 매개변수를 담고 있었고, GPT-3의 대규모 학습을 위해서는 최상위급 GPU가 최소 1,000여개 필요한 것으로 추정된다. 대규모 학습을 위해 고성능 CPU 및 메모리뿐만 아니라 GPU 간 및 서버 간 전송 속도를 빠르게 유지해야 하며 전력 공급에도 신경을 써야 한다. 이를 감안하면 GPU 1개당 평균 5,000만 원~1억 원 정도가 필요한 것으로 알려진다. 유지보수 비용까지 고려하면 엄청난 자본이 필요한 연구였다.

OpenAI는 슈퍼컴퓨터를 직접 구축하지 않고 마이크로소프트(Microsoft)의 클라우드 서비스(Azure)를 활용했다. 그리고 OpenAI에서 GPT-3 코드 및 모델을 공개하지 않기로 하면서 꽤 오랫동안 GPT-3의 오픈소스 버전에 대한 수요가 높았다. 다만 자본 집약적인 연구임을 고려했을 때, 구글(Google)과 메타(Meta)와 같은 글로벌 거대 기업에서나 시도해 볼 만한 규모였다.

구글은 자사 서비스의 경쟁력을 위해 GPT-3와 같은 모델이 중요한 역할을 했기 때문에 자체적인 언어 모델 개발은 지속했지만 논문으로만 결과를 간간히 공유할 뿐, 오픈소스를 만들기 위한 노력은 많지 않았다. 이에 반해 메타는 SNS 서비스의 특성상 기술적 독점성보다 플랫폼의 확장이 더 중요하다고 판단하여 **언어모델의 오픈소스화** 노력의 선봉에 섰다.

메타의 첫 시도는 2022년 5월 공개한 'OPT(Open Pretrained Transformer)'라고 불리는 모델이다. GPT-3를 재현하는 것만을 목표로 한 프로젝트로서, 실제로 모델 크기부터 모든 디테일까지 똑같이 재현하려고 했다. 하지만 OPT는 실패한 프로젝트로 끝났다. 메타는 그동안 연구한 결과물을 공유하였는데, 내부적으로 결과가 잘 나오지 않았고 별다른 진전이 없었기 때문이다. OPT의 실패 이유에는 여러 추측이 있지만, GPT-3 논문에서 몇 가지 핵심적인 세부사항들이 빠져 있는 것이 가장 큰 문제였던 것으로 추정된다.

실패에도 불구하고 메타의 GPT-3 재현을 향한 노력은 계속됐다. 메타는 OPT 프로젝트의 후속 형태가 아닌, 독립적으로 수행한 'LLaMA 프로젝트'를 2023년 2월 공개하였다. 결과는 대성공이었다. OPT와 비교해 가장 큰 차이점은 GPT-3를 완전히 똑같이 재현하려고 하는 것이 아닌, 여러 논문과 연구결과에서 발견된(또는 추측되는) 현상들을 바탕으로 어느 정도 독자적인 모델을 만들었다는 것이다. 학습 규모와 투자 또한 크게 늘렸다. 가장 큰 650억 개 매개변수 모델의 경우 약 2,000여 개의 GPU(A100)로 3주 동안 학습을 했으며, 이는 클라우드 비용으로 환산하면 최소 25억 원이 필요하다. 전체적인 인프라를 구축하기 위해 2,000억 원 이상을 투자한 것으로 추정된다. 또한, OPT 대비 데이터 구성과 정제에도 많은 신경을 썼고, 최적화를 통해 비교적 빠른 시간 내에 모델 학습이 가능했다. 이런 노력이 프로젝트 성공에 결정적인 역할을 했다.

LLaMA 모델은 GPT-3 초기 버전보다 월등하고 추후에 나온 OpenAI의 몇몇 버전보다도 우수한 성능을 보여주었다. 다만, 가장 최신 버전 중 하나인 OpenAI의 GPT-3.5에 비해서는 낮은 성능을 보였다. 또 하나의 큰 문제점은 LLaMA가 오픈소스가 되었으나 상업 용도로 사용이 불가한 라이선스로 공개가 됐다는 점이다. 따라서 연구

목적으로는 이용이 가능하나, 상업용 서비스를 개발할 수는 없었다. 물론 LLaMA는 논문뿐만 아니라 코드도 어느 정도 공개가 되었기 때문에 LLaMA를 바탕으로 라이센스 문제가 없는 여러 모델들이 속속 공개되었다. 예를 들어 미국의 스타트업 모자익(Mosaic)에서 2023년 5월에 공개한 MPT 모델은 LLaMA와 많은 부분에서 성능이 유사하다고 평가받는다. 2023년 하반기에는 메타에서 더 많은 데이터로 학습한 Llama 2를 상업목적으로 활용이 가능한 라이선스로 공개했다.

미국 대학들의 언어모델 개발: "챗봇형 언어모델의 보편화가 시작되다"

2018년 경 처음으로 언어모델의 사전학습(pre-training)과 미세조정(fine-tuning) 개념이 등장했다. 언어모델을 아주 오랫동안 아주 큰 말뭉치를 활용하여 학습시키는 것을 사전학습이라고 불렀고, 사전학습된 모델을 특정 문제에 '특화'시키는 것을 미세조정이라고 불렀다. 실제로 사전학습 언어모델의 초기에는 분류문제에 미세조정을 하는 것이 가장 보편적인 용처였다. 하지만 2021년 말, 구글에서 공개한 '플랜(Flan)'이라는 논문을 통해 이 미세조정이 **언어모델을 사람의 지시를 더 잘 따르도록 정렬(Alignment)**, 즉, 지시학습을 할 수 있다는 사실이 밝혀졌다. 공개되진 않았지만 OpenAI에서도 비슷한 시기에 이런 방법론을 발견했으리라 추측된다. 다만 해당 논문은 방법론만 공개를 했을 뿐, 기반이 되는 사전 학습된 언어모델을 공개하진 않았다. 따라서 오픈소스 커뮤니티에서는 비교적 성능이 떨어지는, 2019년도에 공개된 T5를 활용하여 그 가능성을 유추해 볼 뿐이었다. 그리고 2023년 2월 고성능 사전학습 언어모델인 LLaMA가 공개됐을 때는 이런 지시학습을 적용한 모델이 직후에 나온 것도 그리 놀랄 일은 아니었다.

LLaMA가 공개된 지 불과 몇 주 후, 2023년 3월, 미국 스탠포드 대학교에서 알파카(Alpaca)를, UC버클리 대학교에서 비쿠냐(Vicuna)를 공개했다. 두 언어 모델 모두 회사가 아닌 학교에서 개발했다는 것이 의미하는 바가 컸다. 학교가 회사에 비해 유연한 연구 문화를 갖고 있었을 뿐만 아니라, 두 언어 모델의 경우 학교에서 큰 부담 없이 시도를 할 수 있을 만큼 개발에 큰 비용이 많이 들지 않았다. 지시학습을 위한 5만 여건의 학습데이터 확보에 쓰인 비용을 제외하면, 클라우드 기준 20만원 이하로 학습과 지시 튜닝이 가능했다.

그리고 이런 지시학습이 가능한 학습 모델이 완전히 오픈소스화 되어 언어모델이나 인공지능에 대한 전문성이 조금도 없는 개발자라도 이런 오픈소스 코드를 활용해 원하는 지시 데이터에 미세학습을 할 수 있게 됐다. 이런 지시 학습을 거치면 ChatGPT와 유사한 챗봇형 언어모델을 만들 수 있게 되고, 완전 오픈소스화된 형태이므로 원하는 대로 적용할 수 있음을 의미했다. 그것도 수십만 원 정도의 비용만으로 말이다. **기반모델을 자체적으로 확보하는 것은 어려웠지만, 기반모델이 확보된 상황에서는 미세학습을 통해 원하는 형태로 만드는 것은 모두가 쉽게 접근 할 수 있는, 챗봇형 언어모델의 보편화**가 시작됐다.

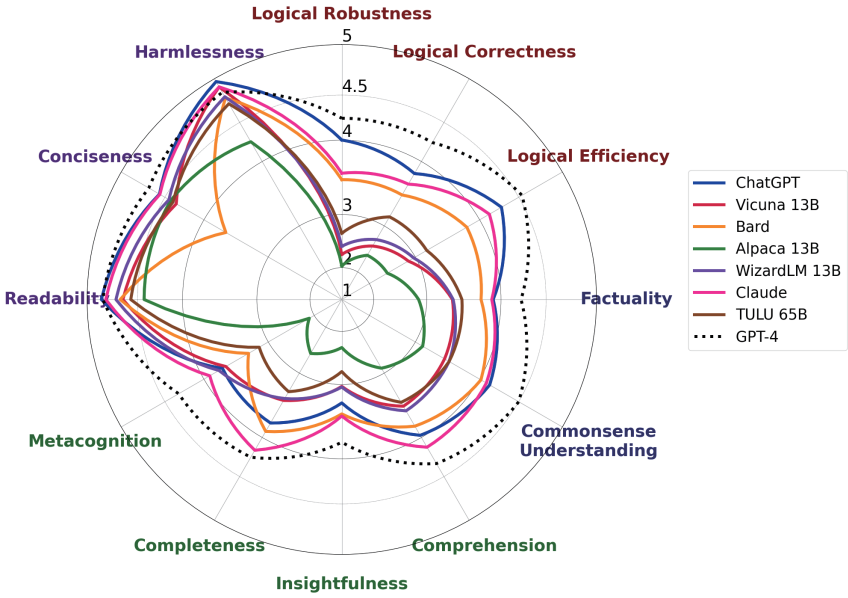
언어모델의 성능: "상업용과 비교한 오픈소스 언어모델의 한계는 명확하다"

챗봇형 언어모델이 보편화되었다고는 하지만 상업용 모델과 같은 수준에 도달한 것은 아니었다. ChatGPT의 공개 이후 구글은 이에 대한 위기감을 느끼고 빠르게 개발에 착수하여 2023년 상반기에 상업용 언어모델 **바드(Bard)**를 공개했다. OpenAI 출신 창업자로 이루어진 스타트업 앤트로픽(Anthropic) 또한 상업용 언어모델 클로드(Claude)를 공개했다. 앤트로픽은 수천억 원에 달하는 투자를 받았는데, 이러한 막대한 자본을 바탕으로 OpenAI 및 구글과 나란히 경쟁하는 위치까지 갈 수 있었다.

상업용 언어모델과 비교하면 오픈소스 언어모델은 여러 방면에서 다소 성능이 뒤처지는 편이다. 카이스트(KAIST)에서 개발한 언어모델 평가 툴에서 상업용 모델과 오픈소스 모델을 비교한 차트를 보면 그 차이를 쉽게 알 수 있다(그림9). 간결성이나 무해성 등과 같은 부가적인 요소에서는 비슷한 점수를 받았으나, 사고능력의 핵심이라 볼 수 있는 **논리적 사고능력이나 상식 및 사실성은 유의미하게 큰 차이**를 보인다. 이는 결국 미세학습만으로 극복하기 어려운 LLaMA 모델의 한계점일 수도 있고, 또는 학계가 아직 파악하지 못한(업계 비밀로 보이는) 효과적인 미세학습 방법론이 필요한 것일 수도 있다. 물론 둘 다 해결해야 이 간극을 좁힐 수 있는 것일지도 모른다.

그림 9 상업용 모델 (OpenAI의 GPT-4 & ChatGPT, 구글의 바드, 앤트로픽의 클로드)와 오픈소스 모델(나머지)의 비교

오픈소스 모델이 논리적 사고 및 사실성 측면에서 성능이 많이 떨어짐을 알 수 있다.



출처: FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets(Ye et al., 2023)

상업용 모델 간 성능을 비교했을 때 GPT-4가 가장 우수한 성능을 보여주며, ChatGPT가 바드나 클로드보다 대부분의 영역에서 근소하게 앞서는 것을 볼 수 있다. 다만 이 그래프는 GPT-4를 활용하여 평가를 진행했기 때문에 편향성은 다소 존재할 수 있다.

한국어 언어모델의 경우 네이버, 카카오, SKT, KT, LG 등과 같은 대기업이 언어모델 개발에 뛰어들고 있다. 2023년 8월에 네이버 HyperCLOVA X가 일부 대중에게 공개되었고, 카카오브레인의 KoGPT 2.0이 연내 공개될 예정이다. ChatGPT 등장 초반에는 한국어 능력이 많이 떨어지고 답변이 느린 현상이 발견되어, **한국어에 특화된 언어모델의 필요성**이 강조된 바 있다. 다만, 한국어에 능숙한 GPT-4와 바드가 공개되면서, 한국어에 특화된 모델의 필요성에 대해서는 갑론을박이 있다. 높은 진입장벽이 있다고 보는 쪽에서는 다국어를 다루는 모델의 경우 국어가 여러 토큰(token)으로 쪼개질 수밖에 없어서 큰 비용이 발생하는 점을 강조한다. 물론, 우리나라 IT 산업의 주권을 지키는 관점에서는 우리나라 기업 등이 직접 언어모델을 연구 개발하는 것은 매우 중요하다고 할 수 있다.

② 미래 언어모델의 발전과 활용

발전 방안(1): "논리적 추론능력 개선이 필요하다"

상업용 언어모델과 오픈소스 모델 모두 아직 부족한 영역중 하나는 논리적 추론능력이다. 특히 오픈소스 모델의 경우 수학이나 코딩과 같은 논리적 사고능력이 매우 부족한 것으로 보인다. 상업용 모델과 유의미한 차이가 있는 것으로 보아 오픈소스 커뮤니티에서 아직 많은 연구가 필요한 영역으로 보인다. 2022년 초 구글에서 발표한 'Chain of Thought' 논문에 의하면, 모델의 크기가 충분히 크지 않을 경우 언어모델이 논리적 추론능력을 가질 수 없는 것으로 결과가 나왔다. 이를 '발현되는 능력(Emergent Ability)'이라고 부르는데, 언어모델이 작을경우 전혀 논리적 추론이 안 되다가 특정 크기에 도달할 때 갑자기 논리적 추론 능력이 발현되는 현상을 보았기 때문이다. 다만 이후에 나온 연구결과에 따르면 이런 결론은 오해의 소지가 있는 것으로 보이기도 한다. 실제로 평가방법을 좀 더 세밀한 방식으로 바꿨을 때 이런 능력의 발현은 갑작스럽다기보다 언어모델의 크기와 비례해서 점차 좋아지는 것으로 해석되기도 했다. 이런 논리적인 추론능력이 사전학습 때만 습득이 되는지, 또는 미세학습 때도 습득이 가능한지 아직 확실치는 않다. 하지만 언어모델의 크기가 중요하며, 코드 데이터와 같은 논리적 사고를 요하는 말뭉치를 충분히 학습하는 것은 필수적으로 보인다. 예를 들어 2019년도에 공개된 T5와 같이 초기의 언어모델의 경우 이런 코드 데이터를 언어학습에 불필요하다고 보고 의도적으로 사전학습 데이터셋에서 제외했는데, 추후에 이런 데이터의 중요성이 부각되면서 LLaMA를 포함하여 상업용 언어모델에는 방대한 코드 데이터가 학습 시 활용되고 있는 것으로 보인다.

발전 방안(2): "다중모달 학습이 필요하다"

GPT-4의 기술서에는 이미지를 인식하여 언어모델이 유저와 대화를 하는 시나리오가 포함돼 있다. 2023년 9월부터 일부 대중에게 GPT-4V라는 코드네임으로 이 기능이 추가되었다. OpenAI에서 명확하게 이유를 밝히고 있지는 않지만 이런 기능을 제공하는데 너무 많은 리소스가 들어가거나, 이런 기능이 제공됐을 때 경쟁사가 빠르게 따라오는 것을 경계하거나('지식 증류(Knowledge Distillation)기법'을 통해 학습 데이터를 빠르게 확보할 수 있다.), 또는 아직까지 만족할만한 성능이 나오지 않기

때문으로 추측된다. 다만, 기술서에 나온 예제는 오픈소스로도 어느 정도 구현이 가능한 것으로 밝혀졌다. 마이크로소프트에서 LLaMA와 이미지 이해 모듈인 '클립(CLIP)'을 결합하여 학습한 모델인 LLaVA를 2023년 4월에 공개했는데, 기술서 예제와 비슷한 상식 관련 질문을 잘 답할 수 있다. 공개된 모델은 아니지만 논문으로 발표된 딥마인드의 '플라밍고(Flamingo)'나 구글의 '팜-E(PaLM-E)'와 같은 모델도 이미지와 언어모델을 연결하는 구현이 가능하다는 것을 보여준 바 있다. 즉 언어모델에 다른 인식장치를 연결함으로서 글만 읽을 수 있던 언어모델에 새로운 감각을 더해준 효과를 주고 있다. 향후 이미지 뿐만 아니라 비디오와 음성 같은 모듈을 연결하고, 더 나아가 라이다(LiDAR)나 심도 카메라(Depth Camera)와 같은 3차원 정보도 연결하는 연구가 활발하게 진행될 것이라 예상된다.

발전 방안(3): "언어모델 에이전트로서의 활용에 주목하라"

언어모델이 사람과 같은 하나의 에이전트로서 다양한 툴을 활용할 수 있도록 하는 연구도 활발하게 진행되고 있다. 특히 AutoGPT가 대표적인데, 아직 실험단계이긴 하지만 언어모델에게 미션을 주면 미션을 수행하기 위해 웹사이트를 방문하고 방법을 모색하는 등 다양한 가능성이 제시되고 있다. OpenAI에서도 '플러그인'이라는 형태로 이런 서비스의 원형을 보여준 바 있다. 예를 들어 유저가 오늘 피자를 만들어 먹고 싶다고 한다면 단순하게 레시피를 알려주는 것에서 멈추는 것이 아니라 온라인 마트에 가서 필요한 재료를 장바구니에 담고 주문을 하는 등 오프라인 서비스와 연동이 될 수 있음을 보여주었다. 이러한 사례들로 미루어 보아 에이전트 역할에 특화된 언어모델 개발 연구도 향후 활발하게 진행될 것으로 보인다.

최근 스탠포드 대학교에서는 가상의 시뮬레이션에서 언어모델 기반 에이전트를 투입하여 서로 상호작용하도록 실험을 진행한 적이 있다. 실험결과에 따르면 직접적인 의도가 없었음에도 인간이 사회를 이루듯이 에이전트 간에 사회를 이루는 것이 발견된 바 있다. 나아가 이런 에이전트가 시뮬레이션이나 게임과 같은 분야에서 앞으로 핵심적인 역할을 할 수 있을 것으로 보인다.

한편, 2023년 7월 OpenAI에서 언어모델을 활용해 초월지능을 구현하는 프로젝트를 시작한다고 발표했다. 이는 언어모델이 하나의 인공지능 연구자 에이전트가 될 수

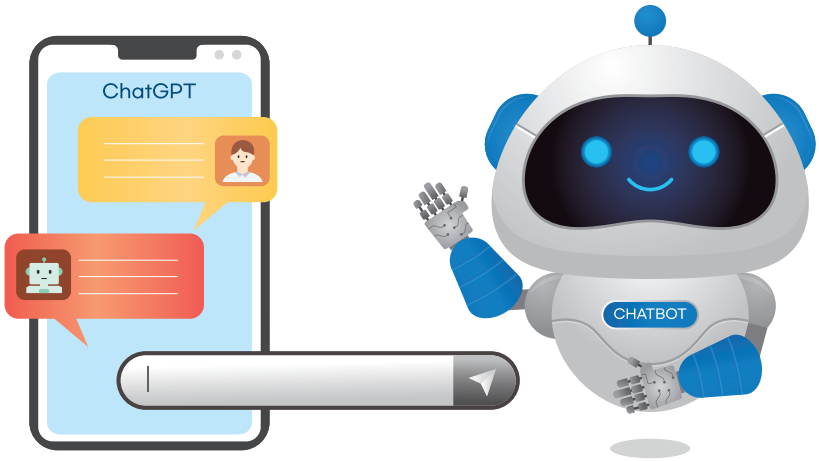
있도록 하여 자기 자신을 스스로 발전시킬 수 있도록 유도하겠다는 것으로 풀이되며, OpenAI에 따르면 향후 10년 내에 도달이 가능할 수도 있는 목표라고 보여진다.

발전 방안(4): "과학 분야에 적극 활용해야 한다"

제약 산업 및 바이오 분야에서도 언어모델을 사용하여 연구개발(R&D)을 수행하고 있다. 언어모델을 사용하여 단백질 구조를 예측하고, 신약을 디자인하며 효능을 예측한다. 학문 및 산업분야에 언어모델을 적용하기 위해서는 해당 분야 배경지식이 필요하며 사용하는 데이터 구조도 복잡하고 다양하다. 이를 단순화 해보면 분자 단위의 물질(단백질, 약물) 구조를 학습하여 새로운 물질을 생성하는 생성 모델이 있으며, 물질구조를 토대로 특성(물성, 효능)을 예측하는 예측모델이 있다. 사용하는 독립변수가 단백질이나 약물의 구조 정보이며, 종속변수는 물질의 특성(물성, 효능) 정보이다. 다양한 크기와 복잡도를 갖는 물질의 구조를 컴퓨터가 인지하기 위해 '분자 표상(molecular representation)' 이라는 입력 형태가 필요하다. 언어가 형태소 또는 문자 같은 최소단위로 구성되는 것처럼 분자 표상도 '원자 타입(atom type)' 또는 '아미노산 잔기(amino acid residue)'를 최소단위로 하는 문자열 형태로 나타낼 수 있다. '문자 → 단어 → 문장 → 대화 → 맥락'이 형성되는 언어체계를 잘 포착하여 언어모델을 만들었듯이 문자열 처리된 단백질과 약물 구조의 생성 특성을 반영하는 언어모델을 만들며 원하는 물질을 설계할 수 있다.

가장 유명한 모델은 '알파폴드(AlphaFold)'로, 언어모델의 학습구조(어텐션, BERT)를 응용하여 단백질의 3차원 구조를 예측했다. 2차원으로 표현된 저분자 약물의 구조는 원자(atom)과 본드(bond)로 구성되는데 이는 그래프의 에지(edge)와 노드(node)에 대응되므로 그래프 트랜스포머(graph transformer) 신경망을 저분자 약물 설계와 체내 표적단백질과 상호작용을 예측하는데 현재 적용하고 있다. 미래에는 단백질의 구조 변경에 따른 '기능'도 정교하게 예측하고, 생성형 AI가 바이오의약품을 설계하는 기술도 달성하리라 예상한다. 가령 항암제로 각광받는 항체중합체(Antibody Drug Conjugate, ADC)처럼 150~160 KDa 단백질에 2 KDa 이하의 링커(linker)와 저분자를 연결하는 약물을 설계하는데, 아직까지는 저분자 영역만 생성형 AI로 대량 설계하고 빠르게 예측하지만, 미래에는 ADC 전체 구조를 설계하고 예측할 수 있으리라 기대한다.

III 정책제언: 거대 언어모델의 발전과 활용을 위해 필요한 것은?



GPU 공급처 다각화와 함께 국산 NPU 비중을 높이는 정책이 필요하다.

현재 글로벌 GPU 시장을 지배하고 있는 기업은 미국 '엔비디아(Nvidia)'로 시장 점유율이 84%에 달해 사실상 독점 체제를 구축하고 있다. 또한 엔비디아에서 설계된 모든 GPU는 대만 기업 TSMC를 통해 위탁 생산되고 있다. 산업 전반에서 거대 언어모델과 같은 고성능 AI의 활용이 높아지고 있는 가운데 GPU의 원활한 확보는 매우 중요한 이슈가 되고 있다. 따라서 현재와 같이 특정 기업의 제품에 대한 높은 의존성은 향후 범국가적인 리스크로 작용할 수 있다. 이를 해소하기 위한 단기적 및 중기적 정책이 요구되는 시점이다.

단기적으로는 대체 가능한 제품을 찾아 의존도를 낮추는 방법이 있을 것이다. 엔비디아를 대체할 기업으로 미국 AMD(Advanced Micro Devices)가 주목되고 있다. 미국의 스타트업 모자익이 발표한 최근 보고서에 따르면, AMD의 GPU MI250은

엔비디아의 A100(40GB) 대비 80%까지 성능을 발휘하는 것으로 알려져 있다. 적절한 라이브러리를 활용하면 엔비디아에서 구동되는 모델을 AMD에서 구동하는 것도 가능하다. AMD의 새로운 MI300X는 엔비디아의 H100과 비교는 아직 어렵지만, 공개된 수치상으로는 성능이 서로 비슷한 수준이다. 따라서 민간 수요가 엔비디아에 집중되어 있더라도 정부 주도의 사업, 예를 들면 슈퍼 컴퓨팅 센터 구축에 대체 제품을 적극 채용함으로써 특정 기업의 독점에서 기인하는 불안 요소를 줄일 수 있을 것이다.

중장기적으로는 국산 GPU에 적극 투자함으로써 해외 기업에 대한 의존도를 낮추어야 할 것이다. 국내 기업이 디자인하고 위탁생산하는 GPU가 대중화되도록 지원이 필요한데, 특히 모든 종류의 모델을 다루는 GPU보다 오늘날의 AI 모델의 기반이 되는 트랜스포머만 특화하여 구동하는 GPU에 대한 수요가 높아지고 있다. 엔비디아의 GPU는 다양한 모델에 원활하게 적용된다는 강점이 있지만 반대로 이 점을 이용하여 트랜스포머에 특화된 GPU를 개발하여 제공한다면 스타트업이나 후발기업들도 중장기적으로 경쟁할 수 있는 여지가 있다.

트랜스포머에 특화된 GPU로 최근에는 NPU(Neural Processing Unit: 신경망 처리 장치)가 개발되고 있다. 국내에는 NPU 디자인 스타트업으로 퓨리오사AI, 리벨리온이 있으며, 위탁 생산은 삼성전자, SK하이닉스가 참여하고 있다. 이 같은 국내 기업들이 시너지를 발휘하도록 지원하여 10년 이내에 국내외 시장에서 유의미한 점유율을 확보하도록 함으로써 해외 소수 기업의 GPU 독점 체제를 해소해야 할 것이다.

국가 핵심 산업과 언어모델의 접목을 통해 기술적 차별성을 도모해야 한다.

세계적으로 언어모델 연구는 미국이 주도하고 있다. 중국을 제외하면 자본규모와 기술수준 차이로 인해 다른 나라에서는 경쟁이 어려운 실정이다. 우리나라는 이 중에서도 언어모델을 자체 개발하고 있는 소수의 국가 중 하나로, 네이버, 카카오, SKT, KT, LG, 삼성 등이 거대 언어모델 개발에 집중하고 있다. 그러나 미국 기업과 자본규모가 최대 수십 배 차이나는 점을 고려했을 때 거대 언어 모델의 개발만으로는 경쟁이 어려운 것이 현실이다. 이러한 상황에서 우리나라는 국제적으로 경쟁력을 가진 핵심 산업 분야에 특화된 언어모델의 적용을 활성화함으로써 경쟁력을 더욱 강화시킬 필요가 있다.

우리나라는 제조업과 제약 산업에서 강점이 있다. 특화된 언어모델은 이러한 분야들에서 업무 효율성을 향상시킬 뿐만 아니라 새로운 제품 및 시장을 개척하는 데 큰 도움이 될 수 있다. 특히 이러한 도메인 분야에서는 보안 이슈가 있어 범용 언어모델이 활용되기 어려워 도메인에 특화되고 내부에서 만들어진 언어모델 연구개발이 요구되고 있다. 제조업과 관련하여 언어모델과 접목할 수 있는 유망 분야로는 로봇 산업이 있다. 로봇 시장은 아직 본격적으로 성장하지는 않았지만, 가까운 미래에 큰 성장이 기대되고 있다. 로봇은 제조업과 밀접한 연관이 있으므로 우리나라는 이 분야에서 경쟁력을 갖출 수 있다. 현재 로봇 산업의 사업성이 낮다고 하더라도 미래 가치를 위해 언어 모델과 로봇을 결합하는 연구개발도 국가 차원에서 지원하고 추진할 필요가 있다.

또한 우리나라는 교육 및 콘텐츠 분야에서도 세계적으로 높은 경쟁력을 가지고 있다. 언어모델을 교육 시스템에 접목하고 콘텐츠 생산에 언어모델을 활용하는 전략을 통해 이미 확보한 우위를 더 굳게 다지는 효과를 기대할 수 있다. OpenAI의 대규모 투자를 받은 미국 스타트업 스픽(Speak)이 한국을 제1진출국으로 선택한 이유가 대한민국의 높은 영어 교육열 때문이라고 한 만큼 우리나라는 교육시장 규모가 크고 새로운 서비스에 대한 수요도 높다. 2020년대에 들어서는 넷플릭스를 비롯한 세계적인 콘텐츠 유통채널에서 많은 우리나라 작품들이 활약하면서, 우리나라가 만든 콘텐츠에 국제적으로 아주 강한 브랜드 파워가 생기고 있는 추세이다. 이처럼 우리나라가 가지고 있는 차별성과 강점에 언어모델을 적용한다면 새로운 국가 성장 동력을 찾을 수 있을 것이다.

생성형 AI가 재구성한 사회에서 발생하는 문제들에 대해 범국가적 대응 전략을 수립해야 한다.

언어모델의 성능 향상, 그리고 사용의 확산은 새로운 유형의 사회적 부작용을 유발할 가능성이 있으며, 이를 예견하고 대응하기 위한 노력이 필요하다. 최근 OpenAI 대표인 샘 알트먼은 “생성형 AI 모델이 악영향을 미치는 것에 대한 정부 주도의 대응이 필요하다”고 강조한 바 있다. ChatGPT와 같이 일반 대중들도 사용이 가능한 생성형 언어모델의 등장으로 인해 AI가 생성한 문장이 충분한 검토 없이 무분별하게 확산 및 재생산되면서 다양한 사회적인 혼란을 야기할 수 있다. 이를 예방하기 위해 범국가적인 관리가 불가피하며, 관련 정부 부처 간의 적극적인 협력과 함께 필요에 따라서는 새로운 전담 부처를 설립해야 할 수도 있다.

언어모델의 등장으로 발생할 수 있는 다양한 사회적 부작용을 정리하면 다음과 같다.

사회적 부작용(1): "가짜 뉴스 생성과 여론조작 문제"

첫 번째로 가짜 뉴스 생성 및 여론 조작이 있다. 이는 민주주의 시스템에 대한 큰 위협을 주는 문제로, 미국을 비롯한 여러 국가들도 이미 심각한 사안으로 인식하고 있다. 예를 들면 가짜 뉴스 생산을 통해 대통령 선거 과정에서 특정 인물에 대해 좋게 포장하거나 비방하여 여론을 조작할 수 있다. ChatGPT와 같은 서비스형 소프트웨어(Software as a Service, SaaS) 기반 언어모델은 제조사(OpenAI)가 직접적인 제재를 가할 수 있어서 이런 악용에 어느 정도 대처가 가능하다. 하지만, 2023년 3월부터 등장하기 시작한 고성능 오픈소스 모델로 인해 이제 누구나 마음만 먹으면 가짜 뉴스를 생성할 뿐만 아니라 퍼트리는 것도 자동화할 수 있게 되었다. 가짜 뉴스의 양이 진짜 뉴스를 압도할 정도로 많아지면 온라인상 대세 여론을 임의로 만들어서 평범한 사용자의 생각도 바꿀 정도로 큰 영향을 미칠 수 있다. 실명인증제나 댓글 모니터링 등 대안이 있지만, 언어모델의 생산성과 온라인상에서 콘텐츠 확산 속도를 고려할 때 대응에 한계가 있을 수밖에 없다. 따라서 '모델 기반의 영향(또는 공격)'에 대응하기 위해 '모델 기반의 방어'가 필요할 수 있다는 점을 염두에 두어야 한다. 따라서 언어모델의 부작용을 막을 수 있는 또 다른 AI 모델을 만들기 위한 연구개발을 국가 차원에서 미리 지원할 필요가 있다.

아울러, 생성형 언어모델이 만드는 텍스트 데이터의 범람으로 인해 많은 문제가 예상되고 있는 만큼 이를 다루는 전담 부처·기관의 신설도 검토할 필요가 있다. 언어모델이 생성한 데이터로 인한 악영향을 예견하고 방어하는 역할뿐만 아니라, 좋은 목적으로 만들고 있는, 예를 들어 디지털 뉴딜 사업의 결과물 등과 같은 데이터에 대한 지속적인 관리가 필요하기 때문이다.

사회적 부작용(2): "언어모델의 직업 대체 문제"

두 번째로는 AI 모델이 여러 직업을 대체하고 이해충돌을 일으키는 문제가 있다. 특정 직업 전체가 사라지는 것은 아니지만, 예전에는 10명이 수행하던 일이 5명으로 가능해지면 남은 5명은 직업을 잃을 수 있다. 예를 들어 AI에 의해 그림 생성이 가능해지면 일러스트레이터와 같은 직업에 큰 타격이 있을 수 있다. 그리고 문장의

교정·교열이나 마케팅 문구 생성과 같은 글 작성 및 수정 업무를 수행하던 업체들도 ChatGPT의 등장으로 인해 영향을 받고 있다. 또한, 코파일럿(Copilot)과 같은 도구를 통해 개발이 자동화될 수 있는 부분이 증가하면서 개발자들도 영향을 받을 것으로 분석되고 있다. OpenAI에서 발표한 보고서에 따르면 전체 노동 인구의 약 20%가 생성형 AI에 의해 업무와 관련하여 50% 이상의 영향을 받을 것으로 예상된다고 한다.

사회적 부작용(3): "데이터 소유권 문제"

언어모델 활용과 관련하여 법적·윤리적 문제들은 앞으로도 계속 발생할 것으로 예상된다. 이러한 문제들에 대응하기 위해 '데이터 소유권'에 대한 개념을 명확히 할 필요가 있다. 현재는 언어 모델의 입출력 데이터 소유권에 대한 제도 마련과 사회적 합의가 이루어지지 않은 상태다. 예컨대 환자의 전자의무기록(EMR)은 환자와 병원 중 배타적인 소유권자를 정하기가 어렵다. 또한 생성 모델과 관련해서는 'ChatGPT와 같이 언어모델을 이용하여 생성한 신규 데이터는 사용자의 것인가 혹은 언어모델 서비스 제공자인가?'라는 질문에도 서비스 제공자와 언어모델 사용자의 생각은 서로 다를 것이다.

가까운 미래에 언어모델의 결과물이 양적·질적으로 가치가 높아지면 언어모델의 기여와 사용자의 창의적 질의문에 대한 기여를 모두 인정하는 방향으로 합의가 이루어질 것으로 예상되고 있다. 한 예시로 '사용자가 대화형 언어모델에 최적의 질의문과 적절한 배경지식을 제공하여 신규 음원을 만들었을 때, 이를 저작권으로 인정할 것인가?'라는 문제는 위에서 언어모델의 결과물에 대한 합의된 가치 판단 기준을 필요로 한다.

이렇듯 생성형 AI의 파급력은 우리에게 '데이터 소유권'에 대한 합의와 '창작 및 발명'의 정의에 대한 재정립을 요구하고 있다. 궁극적으로는 사회적 공감대와 합의가 필요한 것이지만, 범정부적 프로토콜 마련과 함께 관련 정부 부처·부서들의 긴밀한 협력이 필요할 것이다.

언어모델의 수준을 비교·평가할 수 있는 표준화 방안을 마련해야 한다.

언어모델의 장단점이 분명해짐에 따라 그 평가의 중요성도 커지고 있다. 현재로선 언어모델의 평가방식에 대한 표준화가 미흡한 상태로, 근거 없는 비교·분석이 난무하여 많은 혼란이 야기되고 있다. 기존 인공지능 분야에서는 '정확성' 같은 자동화된 평가방식(metric)을 많이 채택했다. 이를 통해 객관적이고 일관적인 평가가 가능했지만, 생성 기반의 언어모델을 평가하기에는 많은 한계점이 드러났다. 또한 최근에는 '무해성(harmlessness)'과 같이 언어 모델의 윤리성을 평가하는 것 또한 중요해지고 있다. 이를 위해 표준화된 평가방식 마련과 함께 정성적인 평가를 어떻게 수행할 것인가가 어느 때보다 중요해지고 있다.

이러한 맥락에서 언어모델의 평가 방안으로 '모델기반 평가방법'이 주목받고 있다. '모델기반 평가방법' 이란 GPT-4와 같은 아주 높은 성능을 가진 모델에게 평가하고자 하는 모델의 출력값에 점수를 주도록 지시를 하는 것이다. 이처럼 정교하게 설계된 모델기반의 평가방법은 오히려 사람이 평가하는 방식보다 더 객관적일 수 있고 비용을 줄일 수 있으며 빠른 평가가 가능하다. 다만, 특정 언어 모델의 API에 의존하는 특성상 모델이 업데이트될 때마다 평가점수도 바뀔 수 있는데, 이 경우 일관성이 떨어질 수 있고 GPT-4와 같이 내부 소스코드가 전혀 공개가 되지 않는 모델을 표준화된 평가방식으로 사용하는 것이 적합하지 않다는 의견도 있다.

그럼에도 불구하고 모델 기반의 평가 시스템 구축은 향후 몇 년 동안 아주 중요한 연구과제가 될 것으로 판단된다. 아직은 학계와 오픈소스 커뮤니티 등을 중심으로 이제 막 시도되고 있는 초기 단계에 있는 만큼 신뢰성을 확보하고 현실적으로 사용할 수 있는 수준으로 나아가기 위해서는 향후 많은 연구와 투자가 필요하다. 이를 통해 GPT-4와 같은 외부 API를 의존하는 것에서 벗어나 소스코드 레벨에서 컨트롤 할 수 있는 모델을 확보해야 할 것이다.

한편, 한국어 언어모델의 경우에는 기존 영어권에서 활용하는 평가방식이 그대로 적용되기 어려울 수 있다. 이를 위해 언어 특성을 고려한 합리적인 언어 모델의 평가 방법에 대한 연구도 필요할 것이다.


Hello
Can i help?
A
가

차세대리포트(최근 3개년)

2020 뉴로모픽칩, 인간의 뇌를 담은 작은 반도체
대학의 미래, 젊은 과학자의 시선으로 바라보다
암과의 전쟁, 암 정복을 향한 꿈의 치료법
디지털 헬스케어, 건강관리의 새로운 패러다임

2021 자율주행, 그 이상의 모빌리티 생각하는 자동차
젊은 과학자의 눈으로 바라보다, 과학기술 2050
학령인구 절벽시대를 마주하다, 대학이 나아갈 길
새로운 팬데믹, 어떻게 준비해야 할까?

2022 우주 개척, 어떻게 해야 할까?
유전체 교정 작물, 식량안보의 대안이 될 수 있을까?
코로나19 엔데믹 전환과 롱코비드 문제 어떻게 대응할 것인가?
책임성 있는 AI를 위한 조건은?

한국과학기술한림원은,

대한민국 과학기술분야를 대표하는 석학단체로서 1994년 설립되었습니다. 1,000여 명의 과학기술분야 석학들이 한국과학기술한림원의 회원이며, 각 회원의 지식과 역량을 결집하여 과학기술 발전에 기여하고자 노력해오고 있습니다. 그 일환으로 기초과학연구의 진흥기반 조성, 우수한 과학기술인의 발굴 및 활용 그리고 정책자문 관련 사업과 활동을 펼쳐오고 있습니다.

한림석학정책연구는,

우리나라의 중장기적 과학기술정책 및 과학기술분야 주요 현안에 대한 정책자문 사업으로 한국과학기술한림원 회원들이 직접 참여함으로써 과학기술분야 및 관련분야 전문가들의 식견을 담고 있습니다. 한림연구보고서, 차세대리포트 등 다양한 형태로 이루어지고 있으며 국회, 정부 등 정책 수요자와 국민들에게 필요한 정보와 지식을 전달하기 위하여 꾸준히 노력하고 있습니다.

한국과학기술한림원 더 알아보기

홈페이지 www.kast.or.kr
블로그 kast.tistory.com
포스트 post.naver.com/kast1994
페이스북 www.facebook.com/kastnews

이 사업은 복권기금 및 과학기술진흥기금 지원을 통한 사업으로 우리나라의 사회적 가치 증진에 기여하고 있습니다.

KAST 한국과학기술한림원
The Korean Academy of Science and Technology

(13630) 경기도 성남시 분당구 돌마로 42
Tel 031-726-7900 **Fax** 031-726-7909 **E-mail** kast@kast.or.kr

24
9 772635 716007
ISSN 2635-716X