초거대 언어모델과 수학추론 연구 동향

Research Trends in Large Language Models and Mathematical Reasoning

권오욱 (O.W. Kwon, ohwoog@etri.re.kr) 언어지능연구실 책임연구원/실장
신종훈 (J.H. Shin, jhshin82@etri.re.kr) 언어지능연구실 선임연구원
서영애 (Y.A. Seo, yaseo@etri.re.kr) 언어지능연구실 책임연구원
임수종 (S.J. Lim, isj@etri.re.kr) 언어지능연구실 책임연구원
허정 (J. Heo, jeonghur@etri.re.kr) 언어지능연구실 책임연구원
이기영 (K.Y. Lee, leeky@etri.re.kr) 언어지능연구실 책임연구원

ABSTRACT

Large language models seem promising for handling reasoning problems, but their underlying solving mechanisms remain unclear. Large language models will establish a new paradigm in artificial intelligence and the society as a whole. However, a major challenge of large language models is the massive resources required for training and operation. To address this issue, researchers are actively exploring compact large language models that retain the capabilities of large language models while notably reducing the model size. These research efforts are mainly focused on improving pretraining, instruction tuning, and alignment. On the other hand, chain-of-thought prompting is a technique aimed at enhancing the reasoning ability of large language models. It provides an answer through a series of intermediate reasoning steps when given a problem. By guiding the model through a multi-step problem-solving process, chain-of-thought prompting may improve the model reasoning skills. Mathematical reasoning, which is a fundamental aspect of human intelligence, has played a crucial role in advancing large language models toward human-level performance. As a result, mathematical reasoning is being widely explored in the context of large language models. This type of research extends to various domains such as geometry problem solving, tabular mathematical reasoning, visual question answering, and other areas.

KEYWORDS 소형 초거대 언어모델, 수학 추론, 초거대 언어모델

Ⅰ. 서론

일반 대중이 OpenAI의 ChatGPT를 대화 형식으로 활용하면서 기존의 인공지능과 달리 인간처럼 의사소통할 수 있는 방식의 인공지능에 대해 인지하게 되었고, 이에 따라 초거대 언어모델의 능력과

* DOI: https://doi.org/10.22648/ETRI.2023.J.380601
* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행한 연구임[RS-2023-00216011, 사람처럼 개념적으로 이해/추론이 가능한 복합인공지능 원천기술 연구].

유용성이 널리 알려졌다. 초거대 언어모델의 능력과 일반화에 대한 유용성은 GPT-3에서부터 언급되었지만, 학습된 정보로부터 새로운 정보를 이끌어낼 수 있는 추론 능력에 있어서는 일반화의 가능성을 보여줄 정도로 뛰어나지 않았다. 특히, 서술형 수학추론에서 성능이 낮았는데, 이후 ChatGPT와 GPT-4로 발전하면서 서술형 수학추론 능력 또한 비약적으로 발전하였다. 본고에서는 초거대 언어모델과 이를 개선하기 위한 연구 동향, 그리고 인공지능의 추론 능력 검증에 적합한 수학추론 연구 동향에 대해 살펴보고자 한다.

II. 초거대 언어모델 최근 동향

1. 초거대 언어모델 개발 동향

언어모델이란 '특정 문장이 등장할 확률을 계산해 주는 모델'로 전산언어학에서 시작된 개념이지만, 딥러닝 전이학습(Transfer Learning)과 결합하여 BERT, GPT-3 공개 이후 인공지능 기술의 중요한 축으로 자리 잡았고, 언어모델의 매개변수(Parameter) 크기를 늘리는 형태의 초거대 언어모델(LLM: Large Language Models) 구축 경쟁이 촉발되었다.

LLM은 언어모델의 한 종류인 언어 '생성' 모델을 방대한 데이터, 수백억 개 매개변수, 이를 학습하기에 충분한 컴퓨팅 자원을 기반으로 계산량을 크게 키워서 다음 단어 예측 정확도를 획기적으로 향상한 모델을 말한다. 이러한 규모의 학습을 통해 구축된 초거대 언어모델은 사전학습(Pre-training) 후에 미세조정(Fine-tuning)이 필요한 기존 언어모델과 다르게 10여 개의 소량 학습데이터로 응용 태스크에 적용이 가능한 few-shot 학습, 추가 학습이 필요없이 단지 태스크에 대한 설명 문구를 지정해 주는 것만으로도 응용 태스크에 적용이 가능한 zero-shot 학습을 가능하게 했다. Few-shot 학습과 zero-shot 학습처럼 추가학습 없이 입력 컨텍스트로만 응용 태스크에 적용 가능한 학습을 ICL(In-Context Learning)이라고 한다.

LLM으로 사람 수준의 문장 생성이 가능하고 소량의 학습 데이터로도 응용 태스크 적용이 가능하여 범용 인공지능(Artificial General Intelligence)에 한걸음 전진했지만, 모델 학습과 구축에 너무 많은 컴퓨팅 비용이 소요되고 또한 실서비스에서도 비용 문제가 걸림돌이 된다.

본격적인 LLM인 OpenAI의 GPT-3 언어모델의 성능은 모델의 구조보다는 매개변수, 데이터, 컴퓨팅 파워에 좌우된다는 Scaling Law[1]를 근거로 175B개 매개변수를 갖는 초거대 생성모델을 학습하여 2020년에 공개하였고, 이에 바탕을 둔 ChatGPT를 2022년 말에 공개하여 일반인에게 그 유용성을 인정받아 많은 기업 및 연구소에서 초거대모델 개발 경쟁에 뛰어들게 하는 계기가 되었다.

이러한 추세에서 구글은 매개변수를 늘려가면서 학습했을 때 상대적으로 작은 개수(8B)의 매개변수 언어모델에 비해 초거대에 해당하는 언어모델(540B)에 갑자기 많은 태스크를 수행할 수 있다는 Emergent Abilities라는 개념을 근거로 거대 모델의 유용성을 주장하였다[2].

참고문헌 [3]에서는 2019년 이후로 공개된 100억(10B)개 이상의 LLM에 대한 타임라인을 정리하며 2020년 GPT-3 공개 이후로 미국, 중국 등을 대표로 한국, 이스라엘 등의 글로벌 빅테크 기업들이 거대 언어모델 개발 및 공개에 몰두하고 있는 현황을 정리하여 공개하였다.

참고문헌 [3]이 공개된 이후로도 2023년 5월 구글에서 540B 규모의 PaLM2를 공개하였고, 2023년 7월에는 IBM이 watsonx.ai, watsonx.data, watsonx.governance로 구성된 Watson X를 개발하였다고 밝혔으며, 2023년 9월에는 Meta가 자체 개발한 생성형

AI인 Meta AI를 공개했다. Apple 또한 GPT와 유사한 언어모델을 개발, 2024년 초에 교육과 엔터테인먼트, 헬스케어 분야 등에 적용할 계획임을 알렸다.

관련하여 국내에서는 2023년 7월, LG가 멀티모달 모델인 EXAONE 2의 API를 우선적으로 LG 계열사에 공개하여 전자, 화학, 통신 등 LG 사업 전반에 초거대 AI를 적용한다는 방침이다. 네이버는 2023년 8월 생성형 AI를 표방하는 CLOVA X를 공개하며 한국 문화와 맥락을 가장 잘 이해하는 언어모델인 Hyper CLOVA X를 기반으로 하는 대화형 인공지능 서비스를 강화하겠다고 밝혔다.

최근에는 대상 데이터를 텍스트에 국한하지 않고 이미지, 동영상, 음성과 같은 멀티모달로 확장하는 대형멀티모달모델(LMM: Large Multimodal Model)도 글로벌 빅테크를 중심으로 개발되고 있다. 오픈 AI의 GPT-4V, 23년 연말에 공개 예정인 구글의 제미니(Gemini)가 대표적이며, 오픈소스인 LLava 1.5도 공개되어 대형 언어모델에서 대형 멀티모달 모델로 확장하는 연구를 가속화하고 있다.

이러한 초거대 모델의 성과에도 불구하고 몇 가지 쟁점 사항이 있다. 첫째는 고비용으로 인해 초거대 모델을 구축할 수 있는 국가와 기업이 국한되어 있어 인공지능 기술을 독점하는 문제로 이미 미국, EU 등에서는 관련한 규제를 준비하고 있는 단계이다. 둘째는 초거대 언어모델로 생성되는 자연어 문장은 의미 등을 고려하지 않고 학습 데이터에서 계산된 확률값에 의존하기 때문에 종종 혐오 발언, 개인정보 데이터 노출, 폭력적 및 차별적 편향 데이터 생성뿐만 아니라 환각(Hallucination)이라고 불리는 거짓 정보를 그럴듯하게 생성하는 문제가 발생하고 있고, 이러한 문제를 해결하기 위해 전 세계 연구기관에서 다각도로 연구를 모색하고 있다.

2. 소형 거대언어모델 기술 연구 동향

LLM의 한계는 역설적으로 그 파라미터 규모와 계산 요구량에 있다. 메타(구, 페이스북)에서 공개한 LLaMA[4] 언어모델의 경우, 65B 파라미터 규모의 모델을 1회 학습하는 데 약 1백만 GPU 시간이 필요한데, 클라우드 대여 시 미화 490만 달러(한화 약 65억 원; 아마존 웹 서비스, 미국 동부 리전, 23년 9월 기준)가 소요될 것으로 추산된다. 만약 학습에 사용될 데이터의 양이 늘어나게 되면 요구 비용 역시 선형적으로 증가한다. 이처럼 파라미터 규모가 커질수록, 모델에 필요한 메모리 공간과 계산 요구량을 뒷받침하기 위한 소요 비용은 점차 감당하기 어려운 수준으로 커지고 있다.

다소 모순적인 이름인 '소형(화된) 거대언어모델(sLLM: Small(er) Large Language Model)'이라 불리는 sLLM은 상대적으로 적은 비용의 컴퓨팅 자원으로 접근할 수 있는 3B~30B(30억~300억) 파라미터 규모의 모델로 정의할 수 있다. 보통 12~80GB 사이의 메모리 공간을 갖는 단일 GPU만으로 학습과 추론을 수행할 수 있는 파라미터 규모의 언어모델이 이 기준에 부합한다고 할 수 있다.

sLLM은 이들 LLM에서 나타나는 Emergent Ability를 유도하여, 한정된 응용 대상에서 초거대 언어모델에 준하거나, 그 이상의 품질을 획득하려는 모델 특화 연구에 집중되어 있다.

일반적으로 LLM의 응용을 구현하기 위한 전체 학습-추론 파이프라인은 (1) 사전학습, (2) 인스트럭션 조정(Instruction Tuning), (3) 정렬(Alignment) 개선, (4) 검증 및 서비스의 4단계로 구분할 수 있다. sLLM의 연구 역시 이 파이프라인을 구성하는 요소로 나뉘어 발전하고 있다.

가. sLLM 사전학습 연구 동향

sLLM 사전학습 연구는 Common Crawl과 같이 대중에게 개방된 웹 스크래핑 데이터 집합을 사용 및 정련하여, 사용된 데이터셋이 공개되지 않는 상용 LLM 성능에 근접하려는 범용 언어모델 학습 연구가 활발하게 진행되고 있다. 기업의 경우 메타의 OPT[5], LLaMA 모델[4], 국가 지원 연구소의 경우 TII의 Falcon 모델을 대표적인 예로 들 수 있다. 또한, LLM 연구 개발을 목표로 구성된 참여적 연구그룹 EleutherAI가 AI 기술 기반 기업 및 클라우드 서비스 기업으로부터 지원받아 GPT-J, GPT-NeoX[6], Pythia, 한국어 언어 생성 모델인 Polyglot-ko[7] 모델 등을 공개한 바 있다. 다른 갈래로는 마이크로소프트에서 공개한 1.3B 파라미터 급의 소형 코드 생성 모델인 phi-1[8]과 같이, 코드 생성(Code Generation)과 같은 특정 분야에 특화한 모델을 상대적으로 적은 양의 데이터로 구축하여 높은 성능을 달성하려는 연구들이 있다.

LLM의 등장 이후 주목할 만한 부분은 LLM 자체가 일종의 데이터 원천으로 활용되어 사람이 원하는 하위 태스크에 대응하는 텍스트 데이터의 합성을 수행하거나, 정련되지 않은 데이터에서 학습에 필요한 데이터인지 아닌지를 가려내는 정련 도구로의 역할을 담당했던 연구가 많은 비중을 차지하고 있다는 것이다. 이렇게 정련되거나 합성된 데이터 그 자체만으로 학습된 언어모델이 그 원천에 해당하는 LLM의 성능을 넘어설 수 있을지에 대해서는 아직 의견이 다소 나뉘고 있으나, 해당 데이터를 사람이 다시 별도의 방법으로 가공, 정련해서 사용하는 방법으로 sLLM의 성능 개선에 한발 더 나아가고 있다는 것에는 이견이 없을 것으로 판단된다.

나. sLLM 인스트럭션 조정 연구 동향

OpenAI의 ChatGPT/GPT-3.5 연구 결과[9]의 공개 이후, 지도학습을 통한 미세조정 데이터셋을 구축하여, 비지도 학습된 LLM의 성능에 준하는 작은 모델을 학습할 수 있다는 사실에 기초한 연구들이 나타나고 있다. 비지도(또는 자가지도) 학습으로 이루어지는 언어 모델링은 사용자가 원하는 응답을 반환하지 못한다. 사람이 원하는 응답을 얻기 위해서는 입력 조건을 예시 등을 포함해 구성할 필요가 있다. 인스트럭션 조정은 지도학습을 통해 사람의 의도를 더 잘 이해할 수 있게 미리 예시를 전달해 학습하는 것과 같다. 스탠포드대학에서 공개한 Alpaca, UC버클리, UCSD와 카네기멜론대학 공동 연구진이 공개한 Vicuna와 같은 모델들이 이에 속하며, 기업의 경우는 DataBricks가 자체 구축한 데이터를 통해 만든 Dolly 모델을 대표적 예시로 들 수 있다. 다양한 예시 중심의 개발은 오픈소스 커뮤니티에서 활발하게 이루어지고 있으며, 응답 및 추론 성능을 개선하기 위해 CoT(Chain-of-Thoughts)[10] 및 Least-to-Most[11]와 같은 프롬프팅(Prompting) 기법에 관한 연구도 이어지고 있다.

다. sLLM의 정렬 연구 동향

언어모델에서 정렬은 인스트럭션 조정을 통해 사용자의 의도를 잘 이해하도록 언어모델을 개선하고자 하는 것과 달리, 사용자가 선호하는 형식의 응답을 생성하거나 요구하는 추론 능력을 개선하도록 생성 정책을 최적화하여 사용자가 원하는 목표로 잘 생성하도록 하는 것을 의미한다. 사용자의 기호도를 반영하기 위해 이를 정량화한 모델로 만들어 강화학습을 수행하는 RLHF(Reinforced Learning from Human Feedback)[9]가 대표적이다. 초기에는 여러 개의 후보 응답을 두고 사람이 선호도를 매겨 출력의 선호도를 순위화하는 보상 모델에 기반한 강화학습 방법을 사용했으나, 점차 결과가 사람에 도움이 되는지 아닌지, 해가 없는지를 따지는 등 보상 모델의

관점을 세분화하는 연구와 사람의 선호도 응답을 AI 피드백으로 대체하는 RLAIF(Reinforced Learning from AI Feedback)[12] 연구로도 이어지고 있다.

한편, 피드백을 통해 모델 자체를 개선하려는 노력 외에도, 입력 프롬프트를 제어하여 사용자가 원하는 답변에 최적화하는 시도들이 있다. 이들은 GPT-3 등장 이후, 사전학습된 정보를 그대로 활용하여, ICL 능력을 극대화하려는 연구로 볼 수 있다. 이러한 접근으로는 주어진 입력이 어떤 태스크를 수행하고자 하는지를 검출하고, 해당 태스크를 잘 수행할 수 있는 프롬프트 문자열이나 벡터를 찾아서 포함시키는 프롬프트 강화 연구[13], 소규모의 학습 데이터를 사용해 언어모델의 파라미터가 고정된 상태에서 그 성능을 최적화하는 프롬프트 벡터 표현을 학습하는 prefix-tuning 기법[14]을 대표적 예로 들 수 있다.

라. sLLM 경량화 연구 동향

서비스에서 연산에 필요한 컴퓨팅 자원은 모두 비용으로 이어지기 때문에, 적은 공간의 메모리를 갖는 GPU 또는 통상의 CPU에서 더 빠르게 언어모델을 추론하거나 학습될 필요가 있다. 모델의 정밀도를 낮추어 속도와 메모리 사용량을 절감하는 양자화(Quantization) 연구가 한 축을 차지하고 있는데, 생성 언어모델의 연산 과정에서 적용할 수 있는 8비트 양자화 연구[15]와 그 오픈소스 구현체 bitsandbytes는 sLLM의 확산에 큰 영향을 미친 바 있다.

한편, 미세조정을 위한 학습 과정에서는 학습될 파라미터 수에 비례해 더 많은 메모리 공간이 필요하다. 적은 메모리 공간으로도 학습할 수 있도록 하기 위해 모델의 일부분만 갱신하여 높은 성능을 내고자 하는 방법들이 제안되었는데, 이 중 LoRA(Low-Rank Adaptation)[16], IA^3[17]가 많은 주목을 받았으며, 최근에는 이를 양자화한 QLoRA[18] 기법이 13B 이상의 모델 미세조정에 활용되기도 한다.

Ⅲ. 수학 추론 연구 동향

수학은 인간 지능의 핵심 능력으로, 세상 경험과 지식에 기반하여 공리와 정의를 이해하고 학습, 추론, 규칙적용, 풀이과정 및 풀이계획, 공간 및 물리적 규칙을 이해 및 활용, 지식을 기호화, 즉 개념과 사물에 이름을 붙이고 사용하는 능력 등의 인간의 복합적 지적 영역을 모두 포함한다. 또한 수학추론은 상식 및 지식 추론과는 달리, 정확한 답과 동의할 수 있는 풀이과정, 즉 추론과정을 가진다. 그렇기에 인공지능의 추론 능력을 검증하기에 매우 좋은 태스크이다.

1. 수학추론 태스크 종류 및 데이터셋

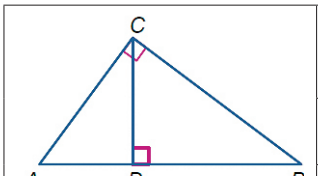
전통적 수학추론에는 서술형 수학 문제 풀이, 정리 증명, 기하 문제 풀이 및 수학 질의 응답 등이 있다. 서술형 수학 문제 풀이(MWP: Math Word Problem Solving)는 그림 1과 같이 사칙연산에 기반하여 하나 이상의 연산 단계를 거쳐 풀이되는 문제로서 기본적으로 상식을 포함한 수학적 추론과 언어 이해를 필요로 한다.

정리 증명(Theorem Proving)은 형식 논리(Formal Logic)를 사용하여 자동으로 수학 정리를 증명하는

Question: Bod has 2 apples and David has 5 apples. How many apples do they have in total?
Rationale: x = 2 + 5
Solution: 7

출처 Reproduced with permission from [19], CC BY 4.0.

그림 1 문장형 수학 문제 예



Question: In triangle ABC, AD = 3 and BD = 14. Find CD.	
Choices: (A) 6.0 (B) 6.5 (C) 7.0 (D) 8.5	
Answer: (B) 6.5	

출처 Reproduced with permission from [20], CC BY 4.0.

그림 2 기하 문제 예

태스크이다. 기하 문제 풀이(GPS: Geometry Problem Solving)는 다이어그램과 텍스트로 구성된 기하 문제의 정답을 구하는 태스크이다. 이를 위해서는 언어적 이해뿐만 아니라 다이어그램과 같은 이미지에 대한 이해도 요구된다. 그림 2는 Geometry3K 데이터셋의 기하 문제 예를 나타낸다[20].

수학 질의 응답(MathQA: Math Question Answering)은 텍스트가 주어졌을 때 주어진 텍스트를 이해한 후 수학적 추론을 통하여 질문에 맞는 답을 구하는 태스크이다. 다음 예는 수학 질의 응답 벤치마크인 DROP의 수학 질의 응답 문제 예를 나타낸다[21].

- Passage: Although the movement initially gathered some 60,000 adherents, the subsequent establishment of the Bulgarian Exarchate reduced their number by some 75%.
- Question: How many adherents were left after the establishment of the Bulgarian Exarchate?
- Answer: 15000

표 1은 수학 추론을 위한 대표적인 데이터셋으로 비교적 최근에 발표된 데이터셋을 대상으로 수학추론 태스크에 따라 분류하였다.

표 1 태스크별 수학 추론 문제 데이터셋

데이터셋	태스크	크기
GSM8K	MWP	8,792
MathQA-Python	MWP	23,914
TabMWP	MWP	38,431
NaturalProofs	TP	32,000
LeanStep	TP	21,606,000
Geometry3K	GPS	3,002
GeoQA	GPS	4,998
DROP	MathQA	96,567
NumGLUE	MathQA	101,835
ScienceQA	VQA	21,208

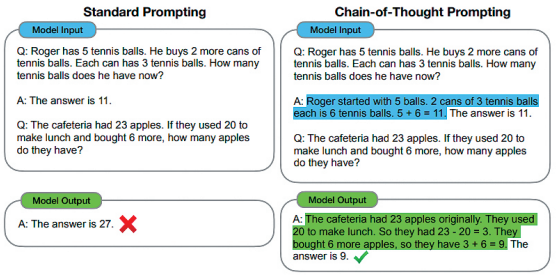
출처 Reproduced with permission from [19], CC BY 4.0.

2. LLM 기반 MWP 연구 동향

GPT, PaLM 같은 LLM의 등장은 MWP 연구에 새로운 장을 열었다. 기존 언어모델이 초대형화되면서 작은 언어모델에서는 보이지 않던 추론 능력이 발현되어 획기적인 성능 개선이 이루어졌기 때문이다[2].

LLM은 모델의 파라미터를 직접 업데이트시키는 미세조정을 대신하여 ICL 기법을 이용하여 모델을 목적 태스크에 최적화시킨다. 통상의 ICL은 몇 개의 입력과 출력의 쌍을 예제로 제공하여 여러 분야에서 좋은 성능을 보였으나 MWP에서는 별다른 성능 향상을 보이지 못했다. CoT 기법[10]이라 불리는 일련의 중간 추론 단계들을 few-shot 프롬프트로 제공하는 기법의 ICL을 이용하여 LLM의 추론 능력을 본격적으로 끌어올린 연구의 시발점이라 할 수 있다.

그림 3은 기존의 표준 프롬프팅 기법과 CoT 프롬프팅 기법의 차이를 보여준다. 기존 프롬프팅에서는 LLM의 입력으로 질문과 답변의 쌍만을 제공하였으나, CoT 프롬프팅에서는 질문에 대한 답으



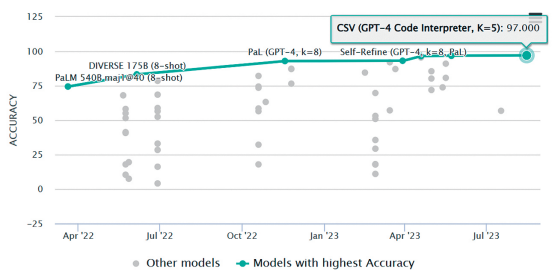
출처 Reprinted with permission from [10], CC BY 4.0.

그림 3 CoT 프롬프팅 기반 MWP 처리 예

로 그 답이 나온 추론 근거가 되는 문장들인 "Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls."를 프롬프트로 함께 제공하고 있음을 볼 수 있다.

그림 4는 2021년 11월 발표된 MWP 관련 대표 평가셋인 GSM8K 데이터셋의 리더보드상의 성능 추이를 나타낸다. 540B의 PaLM 기반 CoT 모델 이후 2023년 현재까지 SOTA를 달성한 [22-26]연구들은 모두 GPT-4에 기반하고 있다.

CoT의 성공 이후, 이에 고무된 후속 연구들은 주로 few-shot 환경에서 입력으로 어떤 프롬프트를 제공하는가가 성능 향상에 매우 중요한 요인이 되었다. Complexity-based Prompting은 CoT의 중간 추론 단계를 더 세분화하여 제공함으로써 정확도를



출처 Reprinted with permission from [28], CC BY-SA.

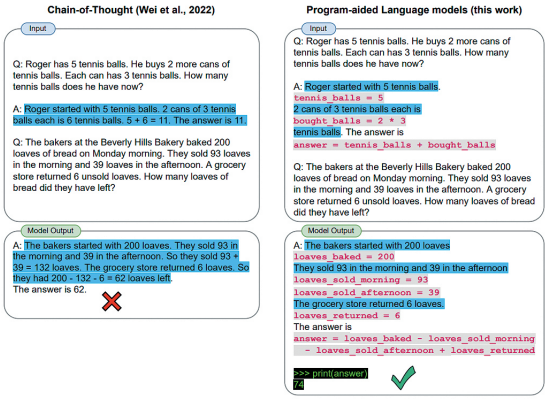
그림 4 GSM8K Arithmatic Reasoning 관련 성능

높일 수 있음을 보였다[27]. 한편, few-shot 프롬프팅 기법과 달리 별도의 예제를 제공하지 않고 수학 문제와 함께 "Let's think step by step"만을 프롬프트로 제공하는 zero-shot CoT로도 LLM이 수학 문제를 논리적으로 추론할 수 있음을 보였다[29].

추론 정확도를 높이기 위한 또 다른 방법으로 LLM을 결과 도출에 여러 번 활용하여 단계적으로 문제를 푸는 다단계 추론 기법들이 있다. Least-to-Most 프롬프팅[11]은 LLM에게 문제를 하위 문제로 분해하도록 질의하는 1단계와 분해된 하위 문제를 차례대로 풀어가도록 질의하는 2단계로 구성된다. 이때, 이전 하위 질문의 답변을 다음 하위 질문의 입력에 추가하여 LLM의 다음 하위 질문의 풀이를 돕는다.

PHP(Progressive-Hint Prompting)[23]에서는 통상의 프롬프팅 형태의 Base Prompting 단계와 PHP 단계로 나뉜다. PHP 단계에서는 LLM이 생성한 답을 다시 LLM에게 힌트로 입력해 주는 과정을 여러 번 반복하여 같은 답이 나올 때까지 답을 생성하게 함으로써 점진적으로 정확도를 높이고자 하였다. 참고문헌 [30]에서는 zero-shot 환경에서 문제풀이를 위한 단계별 계획을 세운 후 각 단계에 대해 풀이하도록 하는 다단계 추론 방법을 제안하였다.

이러한 노력에도 불구하고 CoT는 여전히 정답 추정에 실패하는 경우를 종종 보인다. 특히 큰 숫자 처리에 어려움이 있으며, 유사한 추론 문제에 대해서 일관성없는 결과를 내는 등의 약점을 보였다. LLM은 텍스트뿐만 아니라 프로그램 코드도 자동 생성할 수 있다. 코드의 논리적 완결성과 수처리 완벽성을 취하기 위해 LLM의 코드 생성 능력에 기반한 추론 연구들이 좋은 성능을 보이고 있다[22,31]. 그림 5는 대표적 코드기반 모델인 PaL의 추론과정이다. '파란색+검은색'으로 표시된 기존의 CoT와 달리 '회색+분홍색'으로 표시된 PaL은 중간 추론



출처 Reprinted with permission from [22], CC0 1.0.

그림 5 CoT와 PaL 기반 MWP 처리 예

과정을 Python 코드로 생성한다. 이후 모델 출력은 Python 인터프리터를 통해 '검은색+녹색'으로 표시된 최종결과를 생성한다[22].

텍스트 형태의 출력과 프로그램 형태의 출력은 장단점이 있다. 참고문헌 [25]에서는 CoT와 PaL의 장점을 모두 취하기 위해 두 모델을 모두 동작시킨 후 최종 결과를 자동으로 선택하는 방법을 제안하였다. 최근 수학 관련 여러 데이터셋의 SOTA를 갱신한 CSV(Code-based Self-Verification) 모델[26]은 GPT-4 코드를 기반으로 하여 명시적 코드 기반 자체 검증 프롬프트와 검증 결과에 기반한 weighted majority voting 기법으로 LLM의 코드 기반 수학적 추론 능력을 향상시켰다.

한편 sLLM 기반 MWP 연구의 경우, 초기 연구에서는 CoT 형태의 추론능력은 100B 이상의 LLM에서만 출현되며 그 이하 소형모델에서는 나타나지 않는다고 보았으나[10], 이후 계속된 LLM 기반 지식 정제, 모델 특화, 강화학습 기법 등을 이용한 연구를 통해 소형 언어모델에서도 다단계 추론이 가능해짐이 증명되었다.

참고문헌 [32]에서는 LLM의 여러 분야에 일반화된 능력치를 줄이고 특정 분야에 능력이 집중될 수 있도록 추가학습함으로써 250M~11B의 소형 언어모델도 CoT 형태의 다단계 수학추론 능력을 가질 수 있음을 보였다. 이를 위해 T5와 같은 사전학습된 베이스 LM을 기반으로, 인스트럭션 조정단계와 code-davinci-002로부터 추출한 MWP용 CoT 학습 데이터를 이용한 모델 특화 단계를 거친다. 최근에는 강화학습을 이용한 sLLM 연구가 괄목할 만한 성과를 보였다. 참고문헌 [33]에서는 RLEIF(Reinforcement Learning from Evol-Instruct Feedback)라는 방법을 제안하여 7B, 13B, 70B 규모에서 SOTA를 달성하였으며, 70B 모델에서는 ChatGPT-3.5보다 좋은 성능을 보이기도 하였다.

3. 멀티모달 수학추론 동향

수학 문제는 텍스트로 문제가 제시되는 MWP도 있지만, 일반적으로 기하학적 도형과 함께 텍스트 문제가 제시되는 GPS, 다양한 테이블과 함께 제시되는 TabMWP(Tabular Mathematical Reasoning), 차트나 다양한 유형의 그림과 함께 제시되는 VQA(Visual Question Answering) 등이 있다. 그림과 텍스트가 함께 제시되는 수학추론은 멀티모달 추론의 대표적인 태스크이다.

GPS 수학추론은 그림의 다양성과 이를 시각적으로 이해하기 위한 어노테이션의 세밀함 때문에 데이터의 규모가 커지고 있다. 데이터셋의 발전과 함께 GPS 수학추론 기술은 심볼릭 추론 방법의 Inter-GPS 모델[20], 그래프 기반 추론 방법의 PGDPNet 모델[34,35], Seq-to-Seq 추론방법의 Geoformer 모델[36]로 발전하고 있다.

TabMWP 수학추론은 테이블 질의응답 태스크에서 발전한 추론 태스크로, 테이블을 제시하고 테이블과 관련된 문제가 제시되면 정답을 추론하는 태

스크이다. 데이터셋은 TabMWP[37], FinQA[38], TAT-QA[39], MultiHiertt[40] 등이 있다. TabMWP를 위해서는 테이블의 구조적 정보를 텍스트 문제와의 의미적 관계 정보로 변환하는 것이 중요하다. BERT, RoBERTa와 같은 이해형 언어모델을 이용한 테이블 기계독해의 대표적인 모델로 TAPAS[41]가 있다. GPT 이후 생성형 언어모델을 이용한 ICL을 이용한 TabMWP 수학추론 모델은 Prompt-PG[37]가 있다. 차트에 대한 수학추론의 경우, 차트를 테이블로 변환하고, 텍스트 문제와 함께 초거대 언어모델에서 ICL을 수행하는 DePlot 모델[42]이 있다.

멀티모달 콘텐츠 분석과 추론기술의 개선을 위한 다양한 도구가 발전하면서, 기존에 개발되어있는 도구들을 Plug-and-Play 방식으로 복합추론(Compositional Reasoning)하는 도구증강형(Tool-augmented) 수학추론 방법들이 연구되고 있다. MWP 수학추론 단계별로 LLMs의 계산을 돕기 위한 세 가지 인터페이스 도구를 이용하는 DELI 모델[43]과 정답을 도출하기 위해 실행되어야 하는 도구의 순서를 예측하는 LLM-based planner를 이용한 도구증강형 모델인 Chameleon[44]가 대표적이다.

Ⅳ. 결론

LLM이 가져온 인공지능의 일반화를 더욱 발전시키기 위해서는 LLM에서 발현된 능력이 보다 작은 모델에서도 가능하게 하는 연구들이 필요하다. 이를 통해 LLM의 신경망 구조나 학습 방법 자체를 개선하여 LLM의 문제를 해소할 수 있을 것이다. 이러한 sLLM 연구들에서 효과적으로 그 성능을 검증할 수 있는 연구가 수학추론 연구이며, 또한 멀티모달 영역으로 확장하여 인공지능 일반화를 탐구할 수 있는 연구분야이기도 하다. 향후 멀티모달 수학추론 모델 연구는 현재 LLM의 한계를 극복하는 데 많은 시사점을 제시할 것으로 기대한다.

용어해설

사전학습(Pre-training) 특정 태스크의 신경망 모델에 대한 가중치 학습 이전에 초기 셋팅 가중치를 학습하는 방법으로, 대표적인 사전학습 언어모델은 레이블이 없는 원시텍스트를 대량으로 비지도(자가지도)학습하여 다른 언어처리 태스크의 초기 가중치 모델로 활용함

미세조정(Fine-tuning) 사전학습된 인공지능 모델의 가중치를 특정 태스크를 위해 해당 태스크의 학습데이터로 최소한의 가중치를 추가로 학습(미세조정)하는 방법

약어 정리

BERT	Bidirectional Encoder Representations from Transformers
EXAONE2	EXpert AI for everyONE 2
GPT	Generative Pre-trained Transformer
LLaVA	Large Language and Vision Assistant
PaLM	Pathways Language Model

참고문헌

[1] J. Kaplan et al., "Scaling laws for neural language models," arXiv preprint, CoRR, 2020, arXiv: 2001.08361.

[2] J. Wei et al., "Emergent Abilities of Large Language Models," arXiv preprint, CoRR, 2022, arXiv: 2206.07682.

[3] W.X. Zhao et al., "A survey of large language models," arXiv preprint, CoRR, 2023, arXiv: 2303.18223.

[4] H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv preprint, CoRR, 2023, arXiv: 2302.13971.

[5] S. Zhang et al., "OPT: Open pre-trained transformer language models," arXiv preprint, CoRR, 2022, arXiv: 2205.01068.

[6] S. Black et al., "Gpt-neox-20b: An open-source autoregressive language model," arXiv preprint, CoRR, 2022, arXiv: 2204.06745.

[7] H. Ko et al., "A technical report for Polyglot-Ko: Open-source large-scale korean language models," arXiv preprint. arXiv: 2306.02254. 2023.

[8] S. Gunasekar et al., "Textbooks are all you need," arXiv preprint, CoRR, 2023, arXiv: 2306.11644.

[9] L. Ouyang et al., "Training language models to follow instructions with human feedback," arXiv preprint,

CoRR, 2022, arXiv: 2203.02155.

[10] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," arXiv preprint, CoRR, 2022, arXiv: 2201.11903.

[11] D. Zhou et al., "Least-to-most prompting enables complex reasoning in large language models," arXiv preprint, CoRR, 2023, arXiv: 2205.10625.

[12] H. Lee et al., "Rlaif: Scaling reinforcement learning from human feedback with ai feedback," arXiv preprint, CoRR, 2023, arXiv: 2309.00261.

[13] T. Shin et al., "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," arXiv preprint, CoRR, 2020, arXiv: 2010.15980.

[14] X.L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," arXiv preprint, CoRR, 2021, arXiv: 2101.00190.

[15] T. Dettmers et al., "LLM.int8(): 8-bit matrix multiplication for transformers at scale," arXiv preprint, CoRR, 2022, arXiv: 2208.07339.

[16] E.J. Hu et al., "LoRA: Low-rank adaptation of large language models," arXiv preprint, CoRR, 2021, arXiv: 2106.09685.

[17] H. Liu et al., "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," arXiv preprint, CoRR, 2022, arXiv: 2205.05638.

[18] T. Dettmers et al., "QLoRA: Efficient finetuning of quantized LLMs," arXiv preprint, CoRR, 2023, arXiv: 2305.14314.

[19] P. Lu et al., "A survey of deep learning for mathematical reasoning," in Proc. ACL, (Toronto, Canada), Jul. 2023, pp. 14605-14631.

[20] P. Lu et al., "Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning," in Proc.. ACL&IJCNLP, (Online Only), Aug. 2021, pp. 6774-6786.

[21] D. Dua et al., "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs," in Proc. NAACL-HLT, (Minneapolis, MN, USA), June 2019, pp. 2368-2378.

[22] L. Gao et al., "PAL: Program-aided language models," arXiv preprint, CoRR, 2022, arXiv: abs/2211.10435.

[23] C. Zheng et al., "Progressive-hint prompting improves reasoning in large language models," arXiv preprint, CoRR, 2023, arXiv: 2304.09797.

[24] A. Madaan et al., "Self-refine: Iterative refinement with self-feedback," arXiv preprint, CoRR, 2023, arXiv: 2303.17651.

[25] X. Zhao et al., "Automatic model selection with large language models for reasoning," arXiv preprint, CoRR, 2023, arXiv: 2305.14333.

[26] A. Zhou et al., "Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification," arXiv preprint, CoRR, 2023, arXiv: 2308.07921.

[27] Y. Fu et al., "Complexity-based prompting for multi-step reasoning," arXiv preprint, CoRR, 2022, arXiv: 2210.00720.

[28] https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k

[29] T. Kojima et al., "Large language models are zero-shot reasoners," in Proc. NeurIPS, (Virtual Only), Nov. 2022, pp. 22199-22213.

[30] L. Wang et al., "Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models," in Proc. ACL, (Toronto, Canada), Jul. 2023, pp. 2609-2634.

[31] W. Chen et al., "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," arXiv preprint, CoRR, 2022, arXiv: 2211.12588.

[32] Y. Fu et al., "Specializing smaller language models towards multi-step reasoning," arXiv preprint, CoRR, 2023, arXiv: 2301.12726.

[33] H. Luo et al., "Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct," arXiv preprint, CoRR, 2023, arXiv: 2308.09583.

[34] Y. Hao et al., "PGDP5K: A diagram parsing dataset for plane geometry problems," in Proc. ICPR, (Montreal, Canada), Aug. 2022.

[35] M.L. Zhang et al., "Plane geometry diagram parsing," arXiv preprint, CoRR, 2022, arXiv: 2205.09363.

[36] J. Chen et al., "UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression," arXiv preprint, CoRR, 2022, arXiv: 2212.02746.

[37] P. Lu et al., "Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning," arXiv preprint, CoRR, 2022, arXiv: 2209.14610.

[38] Z. Chen et al., "Finqa: A dataset of numerical reasoning over financial data," arXiv preprint, CoRR, 2021, arXiv: 2109.00122.

[39] F. Zhu et al., "TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance," arXiv preprint, CoRR, 2021, arXiv: 2105.07624.

[40] Y. Zhao et al., "MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data," arXiv preprint, CoRR, 2022, arXiv: 2206.01347.

[41] J. Herzig et al., "TaPas: Weakly supervised table parsing

via pre-training," arXiv preprint, CoRR, 2020, arXiv: 2004.02349.

[42] F. Liu et al., "DePlot: One-shot visual language reasoning by plot-to-table translation," arXiv preprint, CoRR, 2022, arXiv: 2212.10505.

[43] B. Zhang et al., "Evaluating and Improving Tool-Augmented Computation-Intensive Math Reasoning," arXiv preprint, CoRR, 2023, arXiv: 2306.02408.

[44] P. Lu et al., "Chameleon: Plug-and-play compositional reasoning with large language models," arXiv preprint, CoRR, 2023, arXiv: 2304.09842.