“거대언어모델(LLM)의 추론 능력 어디까지 왔을까?”
GIST, LLM 추론 능력 정량적 평가 방법 개발

- AI융합학과 김선동 교수팀, 사고 언어 가설(LoTH) 기반 LLM 추론능력평가 프레임워크 개발... LLM, 일부 추론 능력 있지만 인간과 비교하면 여전히 상당한 한계 보여
- AI의 논리적 사고와 문제 해결 과정 분석... 인간 수준의 추론 능력 개발에 기여 기대
- 국제학술지《ACM Transactions on Intelligent Systems and Technology》게재

▲ (앞줄 왼쪽부터 시계 반대방향으로) 김선동 교수, 이승필 학생, 신동현 학생, 김세진 연구원

인공지능은 인간이 가진 추론 능력*을 어디까지 모방할 수 있을까? 오픈AI가 챗GPT에 적용된 거대언어모델(LLM)*인 GPT-4는 언어 능력과 기억력에서 큰 발전을 이루었지만 **실제 논리적 사고나 추론 능력은 여전히 제한적**이라는 평가를 받고 있다.

특히 LLM의 추론 능력에 대한 정의가 모호할 뿐만 아니라 기존 평가 방법은 주로 결과 중심적이어서 LLM이 어떻게 사고하고 추론하는지를 객관적·포괄적으로 평가하는 방법이 명확하지 않았다.

* **추론 능력(Reasoning):** 대규모 언어 모델의 성능을 평가하는 중요한 요소 중 하나로서, 이는 모델이 주어진 정보를 바탕으로 논리적 결론을 도출하고, 문제를 해결하며 복잡한 질문에 대한 답변을 생성할 수 있는 능력을 평가한다.

* **거대언어모델(Large Language Model, LLM) :** ChatGPT, Claude 등 대규모 파라미터를 이용한 생성형 언어 모델을 지칭한다.

광주과학기술원(GIST, 총장 임기철)은 AI융합학과 김선동 교수 연구팀이 **LLM의 추론 능력을 정량적으로 측정할 수 있는 새로운 프레임워크를 개발**했다고 밝혔다. (https://llm-on-arc.pages.dev/).

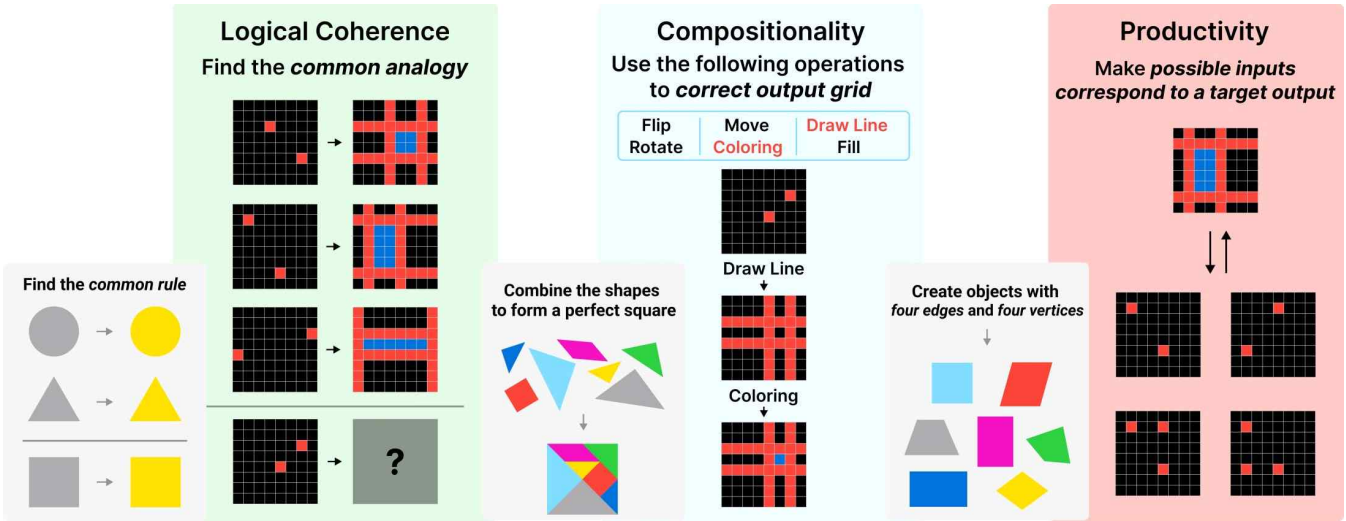
연구팀은 인간의 인지 과정이 '사고 언어'로 매개된다는 인지심리학의 **'사고 언어 가설(Language of Thought Hypothesis, LoTH)'을 기반으로 LLM의 추론 과정을 평가하는 방법을 제시**했다.

이 가설에 따르면, 인간의 추론 과정은 ▲논리적 일관성 ▲구성성 ▲생성성의 세 가지 특징*을 가진다. 이 세 가지 요소에 초점을 맞춘 연구팀은 벤치마크* 데이터세트 **ARC*를 통해 프로세스 중심 방식으로 LLM의 추론 및 문맥 이해 능력을 평가하는 새로운 접근 방식을 도출**했다.

* ▲**논리적 일관성**은 추론 과정과 결과에서 논리적 일관성을 유지하는 능력을 ▲**구성성**은 단순한 요소들을 조합하여 복잡한 아이디어를 구성할 수 있는 능력을 ▲**생성성**은 관찰된 데이터에서 보이지 않은 새로운 표현을 무한히 생성할 수 있는 능력을 의미한다.

* **벤치마크(Benchmark)**: 생성형 인공지능(AI) 분야에서 LLM의 성능을 평가하는 기준 데이터세트를 의미하며, 벤치마크 점수라고 하면 특정 LLM이 벤치마크 데이터세트를 얼마나 정답에 가깝게 산출해 내는지를 평가한 수치다.

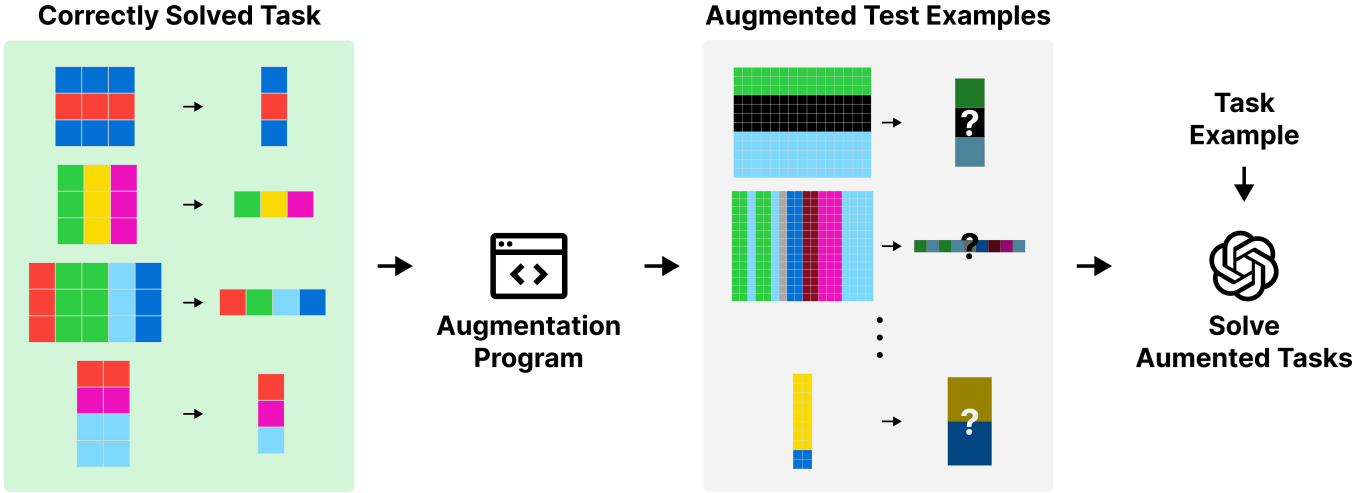
* **ARC(AI2 Ressoning Challenge):** 인공지능의 추론 능력만을 공정하게 측정하기 위해 개발된 벤치마크로, 30X30개의 그리드와 10개의 색상만 사용한 입출력 이미지로부터 규칙을 추론하는 것이 특징이다.



▲ LoTH의 세 가지 핵심 개념과 ARC를 통해 이를 검증한 실험들의 개요. 표시된 작은 그림은 논리적 일관성, 구성성, 생산성이 무엇인지를 나타내는 그림(왼쪽)과 ARC 벤치마크를 활용해 어떻게 실험했는지를 나타내는 그림(오른쪽).

먼저, 논리적 일관성을 측정하기 위해 LLM이 문제를 해결할 때 일관된 정답을 도출하는지를 실험했다. 연구팀은 동일한 문제를 변형한'증강 문제'를 만들어 LLM이 변형된 문제에서도 동일한 논리를 유지하는지를 분석했다. [그림2] 이를 통해, LLM의 논리적 일관성이 프롬프팅 방법*에 따라 차이를 보인다는 점을 확인했다.

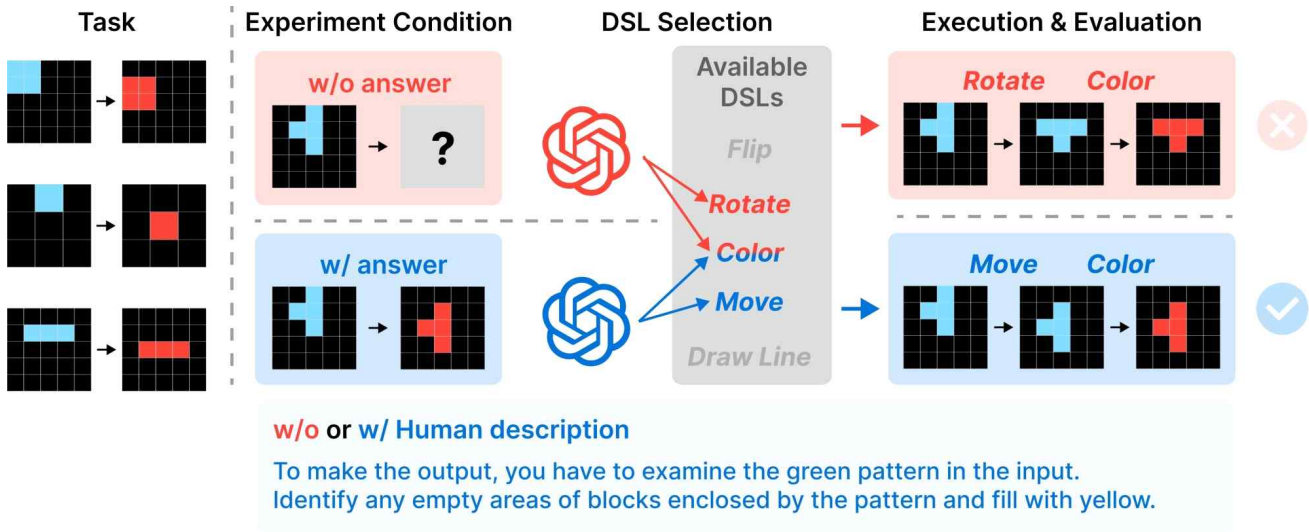
* **프롬프팅 방법**: AI 모델에게 내리는 지시 사항. ▲생각의 연결고리(Chain of Thoughts, CoT) ▲생각의 나무(Tree of Thoughts, ToT) ▲점진적 심화법(Least to Most, LtM) 등



▲ **LLM의 논리적 일관성(Logical Coherence)를 측정하는 방법과 문제 증강 예시**. 입력으로 주어진 과제와 동일한 규칙을 공유하는 평가 데이터를 생성 후 LLM에게 증강 데이터에 대해 측정.

다음으로, 구성성(조합 능력)을 평가하기 위해 LLM이 문제를 해결하는 데 필요한 개념들을 얼마나 효과적으로 조합하는지를 실험했다*. 전체 과정을 고려해 개별 개념을 조합하는 인간에 비해 LLM은 **조합해야 할 단계가 많아질수록 정확도가 떨어지는 모습**을 보였다.

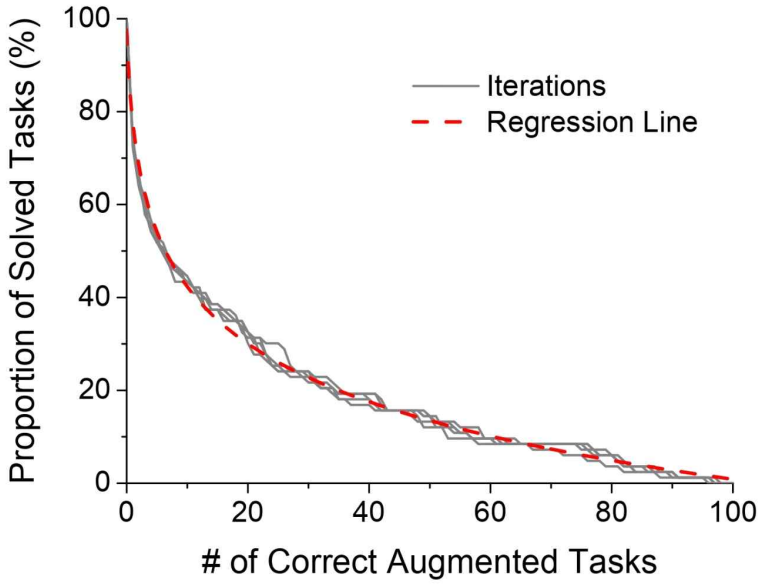
* ARC 벤치마크를 해결할 수 있는 하위 함수 집합(DSL, Domain Specific Language)을 제공하였다.



▲ **LLM의 구성성(Compositionality)을 시험하는 방법**. 사람의 설명을 제공한 후 적합한 파이썬 함수를 고르는 방식의 프로그램 합성법을 처음 도입.

마지막으로, **LLM의 생성성을 평가하기 위해 제약 조건에 맞는 유효한 결과를 얼마나 많이 생성하는지를 실험**했다. 이를 위해 연구팀은 **ARC 문제를 여러 개의 카테고리로 나누고, 역방향의 새로운 프롬프팅 방식을 제시**했다.

또한 연구팀은 **LLM의 추론 능력을 과정 중심으로 분석하는 실험법을 제시**하였으며, **이 과정에서 LLM뿐만 아니라 추론 AI 개발에 필요한 LLM을 활용한 프로그램 합성법, 프롬프팅 기법을 통한 데이터 증강법 등을 제안**하였다.



▲ **증강 데이터에 대한 LLM의 정확도.** LLM이 맞춘 30개의 원본 과제에 대해 각 100 문제를 증강하여 실험한 결과, 정확도가 지수적으로 감소함을 확인. 이는 같은 유형의 문제의 일부만 변형되어도 LLM이 풀 수 없음을 시사함.

LLM의 추론 능력을 정량적으로 측정한 결과, 논리적 일관성 부문에서 증강(변형) 문제에 대해 평균 18.2%의 정확도를, 구성성 부문에서 조합 과제에 대해 5~15%의 정확도를, 생성성 부문에서는 17.12%의 생성 타당도를 보였다.

연구 결과에 관하여 연구팀은 **LLM이 일부 추론 능력을 보이지만 계획 단계가 길고 입출력 이미지가 복잡해지면 단계적인 추론을 거치지 못하여 이 세 가지 측면(논리적 일관성, 구성성, 생성성)에서 한계**를 보이며, 인간과 비교했을 때 추론 능력은 여전히 뒤쳐져 있다고 설명했다.

김선동 교수는 "이전의 LLM 평가 방식이 특정 벤치마크에 의한 성능 측정에 치중한 반면, 이번 연구는 **LLM의 추론 과정과 인간의 차이를 분석한 것이 특징**"이라며 "향후 AI 로봇을 비롯한 **인공지능 시스템이 인간 수준의 추론 능력을 갖추는 데 기여할 것으로 기대**한다"고 말했다.

AI융합학과 김선동 교수가 지도하고 이승필 학사과정생, 심우창 석사과정생, 신동현 석사과정생이 수행한 이번 연구는 한국전자통신연구원 연구개발지원사업, 정보통신기획평가원 디지털혁신기술 국제공동연구사업, 한국연구재단 중견연구자지원사업의 지원을 받았으며, 국제학술지 《ACM Transactions on Intelligent Systems and Technology(TIST)》에 2025년 1월 20일 온라인 게재됐다.

논문의 주요 정보

1. 논문명, 저자정보

- 저널명 : ACM Transactions on Intelligent Systems and Technology (TIST) (IF: 7.2, 2023년 기준)
- 논문명 : Reasoning Abilities of Large Language Models: In-Depth Analysis on the Abstraction and Reasoning Corpus
- 저자 정보 : 이승필(제1저자, 전기전자컴퓨터공학과), 김선동(교신저자, AI융합학과), 심우창(제1저자, AI융합학과), 신동현(제1저자, AI융합학과), 서원규(제4저자, 전기전자컴퓨터공학과), 박지원(제5저자, 전기전자컴퓨터공학과), 이석기(제6저자, 전기전자컴퓨터공학과), 황산하(제7저자, 전기전자컴퓨터공학과), 김세진(제8저자, 전기전자컴퓨터공학과)