

일반논문

법률영역에서 GPT-4 활용 가능성과 시사점^{†*}

- 법학적성시험(LEET) 중심으로 -

Potential Applications and Implications of GPT-4 in Legal Inference Using Korean Legal Aptitude Test (LEET)

박 성 미(Sungmi Park)** · 박 지 원(Jeewon Park)*** · 안 정 민(Jungmihn Ahn)****

목차

- I. 서론
- II. GPT-4 아키텍처 분석
- III. 연구 대상 및 방법
- IV. 연구 결과 및 분석
- V. 결론

<국문초록>

2022년 11월 OpenAI에서 공개한 초거대 언어모델 (Large Language Model, LLM) 기반 대화형 인공지능 ChatGPT가 전세계적으로 큰 반향을 일으키고 있다. 특히 가장 최근 모델인 GPT-4의 경우 LSAT의 성적이 상위 12%의 점수를 기록하는 등 혁신적인 성과를 보여줌으로써 언어와 추론에 크게 의존하는 법률 분야에서 법률 문서 해석, 계약서 작성 등 모든 법률 업무에서 사람을 대체할 수 있는 수준까지 얼마 남아 있지 않겠다는 예측이 있다.

본 연구는 이러한 막연한 우려에 대한 근거를 객관적인 실험을 통해 증명해보고자 하였다. 이를 위해 본 연구에서

는 OpenAI가 발표하였던 미국의 LSAT 실험 결과를 재확인해보고, 이와 유사한 우리나라의 법학적성시험인 LEET 시험의 문제풀이 분석을 통해 법률영역에서 GPT-4가 가지고 올 영향을 평가해 보고자 하였다. 이를 위해 GPT-4의 문제풀이가 논리적이고, 타당하며, 정확한 추론을 내릴 수 있도록 다양한 프롬프트 디자인을 통한 실험을 진행하였다. 이러한 실험 방법을 통해 GPT-4의 추론 과정에서 발생한 문제점을 분석하여 적시하였다. 연구결과에 의하면 GPT-4 모델이 LEET 시험 문제를 풀 때 피상적인 분석만 가능하며 새로운 사실 적용이 어렵고, 개방형 질문에 대한 답변이 일관되지 않으며, 암시적 사실은 단순히 없는 것으로 판단하는 경우 등이 나타나 아직까지 GPT-4는 합리적인 추론모델로 인정하기 어려웠다. 다만, 법률 말뭉치를 활용한 사전 학습, 법률전문가 피드백 시스템 구축, 법적 지원 기술 개발 등의 방식으로 추론영역에서 사람의 의사결정을 지원할 수 있는 가능성은 충분하다고 판단된다.

I. 서론

지난 2022년 11월, OpenAI에서 GPT-3.5 버전에 기반한 초거대 언어모델(Large Language Model, LLM)인 대화형 인공지능 챗봇 ChatGPT를 공개하여 전 세계적으로 큰 반향을 일으키고 있다. OpenAI는 2018년 GPT-1을 처음 출시한 이후, GPT-3.5를 탑재한 ChatGPT는 공개된지 단 몇 달만에 100만

† 투고일자 2023.4.25. 심사일자 2023.5.25. 게재 확정일자 2023.5.27

* 이 논문은 2023년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: AI 기반 범죄수사 지원 / 연구개발과제번호: PR10-02-000-21]

** 고려대학교 정보보호대학원 박사과정, 주저자

*** 한림대학교 정보법학과 석사과정

**** 한림대학교 정보법학과 교수, 교신저자

명의 사용자를 유치하였고, 3월 말 현재에는 유료 회원을 포함하여 1억명이 넘는 가입자가 이용하는 자연어처리 관련 분야에서 최강자로 자리잡았다.) ChatGPT를 개발한 OpenAI사에 직접 투자를 한 마이크로소프트(MS)는 인공지능(AI) 챗봇 기술을 적용한 검색엔진 ‘빙(Bing)’을 내세워 검색의 ‘절대 강자’인 구글을 추격하고 있다.²⁾ 나아가 MS사가 자사의 소프트웨어인 Microsoft 365에 GPT-4를 탑재하여 사용할 수 있는 Copilot을 공개하는 등 GPT 기반 상용화 서비스가 제공되기 시작하면서³⁾, 인공지능 전문 기업뿐만 아니라 회계법인⁴⁾, 리컬리 서치⁵⁾ 등 특정 도메인 전문 기업들도 OpenAI 또는 유사 인공지능 업체와의 협업을 통해 인공지능 시장에 뛰어드는 추세를 보이고 있다. OpenAI는 지난 3월 14일 불과 몇 달 전에 출시한 ChatGPT 성능을 배로 향상시킨 GPT-4 버전을 출시하였다. 이미 각계에서 뛰어난 성능을 보여주고 있는 GTP-4는 곧 법률 분야에서도 두각을 나타낼 것으로 예측되고 있다.⁶⁾

1) 정영훈, ‘1억명 이용’ AI 챗GPT...“월 20달러 유료화”, 2023. 02.02., KBS 뉴스, <<https://news.kbs.co.kr/news/view.do?ncd=7596001>> (최종방문일 23.04.01).

2) 차병섭, MS검색엔진 Bing, 오픈AI 업고 구글 추격...방문자 수 16%↑, 2023. 03. 23. 연합뉴스, <<https://www.yna.co.kr/view/AKR20230323084400009>> (최종방문일 23.04.01).

3) 전남혁, ‘PPT 만들어줘’ 몇 초만에 똑똑...MS, 워드, 엑셀에도 AI 탑재, 2023.03.17., 동아일보, <<https://www.donga.com/news/IT/article/all/20230317/118379443/1>> (최종방문일 23.04.01).

4) “PwC announces strategic alliance with Harvey, positioning PwC’s Legal Business Solutions at the forefront of legal generative AI” [PwC, Harvey와 전략적 제휴 발표, PwC의 법률 비즈니스 솔루션을 법적 생성 AI의 최전선에 배치], 2023.3.15., PwC, <<https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-announces-strategic-alliance-with-harvey-positioning-pwcs-legal-business-solutions-at-the-forefront-of-legal-generative-ai.html>> (최종방문일 23.04.01).

5) Caroline Hill, “Casetext unveils AI legal assistant powered by OpenAI’s ‘most advanced’ large language model” [Casetext, OpenAI의 “가장 진보된” 대규모 언어 모델로 구동되는 AI 법률 비서 공개], 2023.03.01., Legal Insider, <<https://legaltechnology.com/2023/03/01/casetext-unveils-ai-legal-assistant-powered-by-openais-most-advanced-large-language-model/>> (최종방문일 23.04.01).

6) 이재환, 변호사도 인정한 GPT-4 적용된 AI 법률 비서 2023. 03. 16. GTT코리아, <<https://www.gttkorea.com/news/articleView.html?idxno=4454>> (최종방문일 23.04.06).

불과 2개월 전인 2023년 1월, 전신인 GPT-3.5-turbo⁷⁾ (ChatGPT 3.5의 모델)가 미국의 법학전문대학원 입학시험인 LSAT(Law School Admission Test)을 통과하여 사람과 유사한 법적 추론능력의 가능성을 보이긴 했으나, 당시 성적은 하위 40%정도에 그쳐 여전히 인간의 추론을 따라갈 수 없다는 결론을 내리기도 했었다.⁸⁾ 그러나 뒤따라 출시된 GPT-4의 경우 LSAT의 성적이 상위 12%(88th percentile)의 점수를 기록하는 등 혁신적인 성능 향상을 보여줌으로써 인간 추론의 평균 수준을 상회하는 것으로 평가받는 수준에 이르렀다.⁹⁾ 이는 GPT-4가 법률 도메인에 특화되지 않음에도 불구하고 정답을 추론해낼 수 있는 일정 수준의 법적 추론능력을 가지고 있음을 시사하는 것이어서 더욱 충격을 주었다. ChatGPT가 이런 수준이라면 언어와 추론에 크게 의존하는 법률 분야에서 법률 문서 해석, 판례 분석, 계약서 작성 등 모든 법률 업무에서 전문가를 대신할 정도의 수준에 이르기까지 얼마 남아 있지 않음을 의미하는 것이기도 하며 많은 기대와 우려가 교차하고 있다.¹⁰⁾

법률영역에서는 고도의 정확성과 추론능력이 요구될 뿐만 아니라 결론에 이르는 사고과정이 합리적으로 제시되어야 한다. 이미 잘 알려져 있다시피

7) OpenAI의 공식 소개글에서 ChatGPT에 사용된 모델과 동일하다고 명시하고 있으나 ChatGPT Plus(유료 버전) 기준, Default 모델은 text-davinci-002-render-sha, Legacy는 text-davinci-002-render-paid 모델을 사용하고 있다. 각 모델의 성능 차이, 학습 정보에 관하여 자세한 정보는 공개되지 않았다.

8) Bommarito II., M., Katz, D. M., “GPT Takes the Bar Exam.”, SSRN, 2022, p.8.

9) 그뿐만 아니라 스탠퍼드 인공지능-법률 연구소인 CodeX와 리컬테크 회사 Casetext가 진행한 미국 변호사시험에서, 상위 10%의 점수를 획득하였다. GPT-4 Technical Report, p.24.

10) 2023년 5월 말 현재 시점, 초거대 언어 모델(Large Language Model, LLM) 연구는 OpenAI의 GPT 모델에 의존하는 것이 아닌 Meta의 Llama, EleutherAI의 GPT-J, bigscience의 BLOOM 등 다른 공개 언어 모델을 활용하여 GPT-4를 대체할 수 있는 모델 개발 방향으로 가고 있다. LoRA(Low-Rank Adaptation of Large Language Models) 등 메모리 사용을 최소화하며 거대 언어 모델을 학습할 수 있도록 여러 저성능 학습 환경에서의 학습 방안이 제시되고 있으며, 관련 코드 등을 Github(소스 코드 관리 및 공유 플랫폼)에 무료로 공개하여 접근성을 높이고 있는 추세다. 이러한 동향을 봤을 때, 도메인 최적화가 완료된 인공지능의 모범하는 현대 사회에서 피할 수 없는 패러다임 시프트로 볼 수 있다.

GPT-4 모델은 ‘생성 사전학습 트랜스포머(Generative Pre-trained Transformer)’ 계열에 속하고, 자연어처리 분야의 성장을 촉진한 BERT모델과 같이 구글이 개발한 트랜스포머에 기반하고 있다.¹¹⁾ GPT-4 모델에서 생성되는 단어나 문장들은 사전 학습된 내용에서 그 다음에 올 문장이 무엇인지를 확률에 의존하고 있다는 점에서 이러한 확률값(probability)이 추리 논증의 타당성(validity)을 보유했다고 볼 수 있을지에 대한 의문이 생긴다.¹²⁾ 다시 말해 OpenAI에서 공개한 시험점수를 통해 실제 GPT-4가 사람과 동일한 또는 사람보다 우수한 역량을 가졌다고 하더라도 이것이 곧바로 법률영역에서 필요한 추론능력을 충분히 갖추었다고 보기에는 어려울 수도 있다는 것을 의미이다.

ChatGPT에 대해서 기대와 우려가 공존한다. 특히 변호사시험이나 의사시험과 같은 전문시험을 통과했다는 사실에 앞으로 해당 분야에 대한 AI의 대체 가능성까지 회자되고 있는 상황이다.¹³⁾ 본 연구는 이러한 ChatGPT의 열풍 속에서 가장 최근 공개된 OpenAI의 GPT-4의 한국어 법적 추론 성능을 실험하고 검토해서 추리 논증 영역에서 GPT-4의 활용 가능성을 알아보려고 하였다. 이를 위해서 흔히 문제점으로 언급되는 잘못된 정보의 제시 여부, 나아가 단순 질문이 아닌 복잡한 예제를 GPT-4가 어떻게 처리하는지를 확인해보고자 하였다.

그 동안 GPT모델의 성능과 관련하여 많이 지적되고 있는 ‘환각(hallucination)’ 문제는 특히 정확성과 타당성이 요구되는 법률영역에서 큰 영향을 끼칠 수 있다는 점에서 확실히 분석되어야 할 연구 대상이다. 환각문제는 옳지 않거나 편향된 문장을

사실처럼 생성하는 텍스트 생성 모델의 오류로 OpenAI의 ChatGPT는 잘못된 정보 제공으로 인해 여전히 구글 등 검색 특화 엔진의 성능에 못 미친다는 평가를 받고 있어 이에 대한 확인이 필요하다고 보았기 때문이다. 또한, 프롬프트에 따른 성능 차이를 분석해보았다. 지시문(prompt, 이하 프롬프트)이 구체적일수록 원하는 답을 정확하게 생성한다면 ChatGPT의 활용 가능성은 높다고 말할 수 있고, 단지 이용자가 좋은 프롬프트를 이용하면 되는 것이다. 즉, 생성 사전학습 트랜스포머 계열의 언어 모델(Language Model)이 추리 논증을 충분히 학습하였다면 질문의 형식 프롬프트가 체계적이면 체계적일수록 생성되는 답변의 타당도(validity)가 높아질 것을 기대할 수 있을 것이고,¹⁴⁾ 이는 법률영역에서 활용될 가능성을 높게 될 것이다.

추론영역에서의 GPT-4 성능평가와 관련하여 또 중요하게 다루어야 할 점은 주어진 프롬프트와 해결하고자 하는 문제의 복잡도에 따라 달라지는 GPT의 성능이다. GPT 계열의 생성 모델은 단순하고 일련의 사고방식을 포함하는 작업에서는 탁월한 성능을 보였지만 여러 단계를 거쳐야 하는 복잡한 문제를 다룰수록 성능이 급감하는 모습을 보여준다. GPT와 같은 언어모델은 수치적으로 변환된 수많은 언어 데이터의 관계성을 학습하여 주어진 입력값에 가장 적합한 다음 단어를 예측(inference)하는 것이지 사람이 생각하는 것처럼 지식을 기반으로 실제 문제를 이해하고 답변을 도출하지 않기 때문이다. 본 연구에서는 이런 한계점을 명확히 둔 채 GPT 모델이 문제에 대한 답을 맞추는 것이 아니라 합리적인 추론 과정을 모방할 수 있는지를 살펴보고자 하였다. 특히, 법적 추론이라는 특정한 문제 해결을 위한 GPT 모델의 활용 가능성을 보기 위해서는 모델 답변 품질 향상을 위한 양질의 데이터 생산, 미세조정(fine-tuning)을 위한 추가적인 학습, 사용자를 통해 지속적인 검증을 진행하는 ‘휴먼

11) 최재식, KAIST 최재식 교수, “챗GPT시대, 어떻게 AI 시장 주인 공으로 살 수 있을까?”, 2023.3.31, 인공지능 신문, <<https://www.aitimes.kr/news/articleView.html?idxno=27683>> (최종 방문일 23.04.01).

12) 추리논증의 경우 수학적 접근과 마찬가지로 해답에 이르는 과정(reasoning)이 논리적이고 불변임을 전제로 한다는 점에서 타당성(validity)의 역할은 매우 중요하다.

13) 최창현, GPT-4 발표한 OpenAI ‘샘 알트만’, “AI가 사회를 재구성할 것이며 위험을 인정합니다, 이것이 조금 두렵습니다”, 2023.3.19., 인공지능신문, <<https://www.aitimes.kr/news/articleView.html?idxno=27595>> (최종방문일 23.04.01).

14) 여기서 생성AI인 GPT-4는 확률기반을 전제하더라도 추리논증에 대해 충분한 학습을 했다고 가정했을 경우 추론과정에 대한 표현은 프롬프트의 품질에 따라 향상될 수 있을 것으로 보인다.

인 더 루프(human-in-the-loop)’ 방식 등으로 제시하는 답변의 타당성과 정확성을 확인해 볼 필요가 있다.

이와 같이 본 연구는 법적 추론능력이라는 관점에서 GPT-4의 성능을 실험하고 검토해 보고자 하였다. 이를 위해 OpenAI의 연구진에서 공개한 GPT-4 실험 중 미국의 로스쿨 입학시험인 LSAT에 뛰어난 성능을 보였다는 점에서 착안하여 그와 유사한 한국의 법학능력시험 LEET를 사용하여 GPT-4의 추론 및 논증 능력을 평가하기 위한 실험을 하였다. 두 시험은 모두 추리논증에 대한 문제가 중심을 이루고 있어 GPT-4의 추론능력 평가가 적합한 것으로 보였다. 이하에서는 가장 최근 2022년 시행된 LEET시험 문제를 대상으로 LSAT와 관련하여 OpenAI에서 제공한 프롬프트를 포함한 다양한 방식의 프롬프트를 디자인하여 GPT-4 모델의 능력을 검토하고 추론과 관련하여 모델이 보여주는 특성을 분석한다. 여기에서는 언어적인 차이에서 발생하는 GPT-4의 능력을 확인하기 위하여 GPT-4를 영어가 아닌 타언어로 논증과 추론 과정에 사용하였을 때 발생하는 문제점들을 함께 해석하였다. 이렇게 분석한 결과를 토대로 GPT-4가 가지고 있는 법적논증 추론에서의 문제점 및 향후 개선 가능성 등을 제시해보고자 한다.

II. GPT-4 아키텍처 분석

1. 사용자 선호도에 최적화된 InstructGPT

GPT(Generative Pre-trained Transformer)는 사담처럼 자연스러운 글을 쓸 수 있도록 대용량의 데이터를 학습시킨 언어모델이다. 2017년 구글의 ‘Attention is all you need’¹⁵⁾라는 논문을 통해 사전 학습(pre-training) 과정을 거친 거대한 언어모델

을 목적에 맞도록 파인튜닝(fine-tuning)해 사용하는 트랜스포머 모델을 공개하며 자연어 태스크를 수행하는 인공지능 언어모델(Large Language Model, LLM)은 비약적인 발전을 거듭해왔다. 그럼에도 불구하고 이러한 언어모델들은 학습된 문제를 성공적으로 해결하는 과정에 적게는 수천 개, 많게는 수만개의 예시(학습데이터)를 요구할 뿐만 아니라 하나의 태스크만 집중적으로 해결하는 등 유연성 부족의 문제가 있었다. GPT-2의 ‘인컨텍스트 러닝(in-context learning)’ 학습 방식은 이러한 한계를 극복하기 위해 개발된 것으로, 태스크 설명과 몇 개의 예제를 모델 입력값으로 두고, 같이 뒤에 올 적절한 답변을 생성하여 문제 해결을 하는 방식이다.¹⁶⁾ 예를 들어, “Dog = 강아지”, “Cat = 고양이”, “Bird = 새”라는 예제를 두고 “Lion”을 프롬프트로 준다면 “사자”를 생성하는 방식이다.

그 동안의 인공지능이 데이터를 분석하고 패턴을 식별하는 분석형 인공지능이라고 한다면 대화형 인공지능 또는 생성형 인공지능 추론과 의사결정 능력을 보유하게 되어 더욱 우리가 원하는 결과를 도출해줄 수 있게 된 것이다. 그러나 GPT-2가 공개되었던 당시, ‘인컨텍스트 러닝(in-context learning)’ 방식으로 학습된 모델은 일반적인 파인튜닝 기법의 성능에 미치지지는 못하였다.¹⁷⁾ 그 후 공개된 GPT-3는 GPT-2와 같은 모델 구조를 지녔으나 더 많은 데이터를 사용하고 파라미터¹⁸⁾ 수(1750억, GPT-2의 약 100배 이상)를 늘려 파인튜닝 과정 없이 몇 개의 예제(few shot learning)만으로 더 나은 성능을 보이게 되었고¹⁹⁾, OpenAI는 2022년 11월 GPT-3를 보완한 ChatGPT(ChatGPT)를 출시하게 된다.

15) Vaswani, Ashish, et al., “Attention is all you need.”, Advances in Neural Information Processing Systems 30, 2017.

16) Brown, Tom, et al., “Language Models are Few-Shot Learners.”, Advances in Neural Information Processing Systems 33, 2020, p.4.

17) Radford, Alec, et al., “Language Models are Unsupervised Multitask Learners”, 2019, p.9.

18) 학습과정에 최적화되어야 하는 가중치(weight), 편향(bias) 등 매개변수를 말하는 것으로 모델의 성능과 예측 능력에 영향을 끼친다.

19) Brown, Tom, et al., 앞의 글, p.8.

그러나 이러한 초거대 GPT 모델은 특정한 태스크의 예제를 학습하는 것이 아니라 이미 학습된 확률에 의한 답을 제시하는 것일 뿐이기 때문에 모델에 지시하는 명령 프롬프트 구성이 곧 모델 결과의 품질에 큰 영향을 미치게 되었다. OpenAI에서는 이러한 문제점을 해결하고자 인간의 평가를 반영하여 강화학습을 진행하는 방식을 GPT 모델에 적용하였는데 이를 ‘인간에 의한 강화학습(Reinforcement learning with human feedback, RLHF)’ 방식이라고 한다.²⁰⁾

RLHF 방법은 (1) 먼저 평가자가 주어진 프롬프트에 대해 원하는 결과물을 작성하고, 그렇게 생성된 데이터를 수집하여 파인튜닝한 모델인 지도학습 파인튜닝(SFT, Supervised finetuning) 모델을 확보한다. 그 후 (2) 주어진 프롬프트에 대해 모델의 결과물을 여러 개 출력하고 인간 평가자는 그 결과물 간의 순위를 비교한다. 그 데이터를 파인튜닝한 모델은 보상 모델(reward model)이라고 한다. (3) 마지막으로 근접 정책 최적화(PPO, Proximal policy optimization) 알고리즘을 사용하여 보상 모델의 값을 통해 정책을 업데이트함으로써 모델은 인간 평가자의 피드백을 반영하게 된다.²¹⁾ OpenAI에서는 사용자의 일반화된 선호도에 맞춰 결과값을 도출하도록 모델을 학습하는 과정을 ‘정렬(alignment)’라고 지칭하며 여러 가지 측면에서 사용자들의 선호도 평가를 진행하였다.²²⁾ InstructGPT를 통해 생성된 결과는 특히 ‘유용성’ 측면에서 GPT-3보다 더 사용자들에게 선호도가 높았으며 잘못된 정보를 사실인 것처럼 답하는 환각오류(hallucinations)도 기존 GPT모델에 비해 월등히 감소한 것으로 나타났다.²³⁾ InstructGPT의 최대 의의는 사용자의 대표적인 선호도를 도출하여 사용자가 원하는 결과물 도출에 최적화되었다는 점이다.²⁴⁾ 본 연구에서 활용

할 ChatGPT는 이러한 InstructGPT의 형제모델로 유사한 학습 방식²⁵⁾을 사용하였다.

2. InstructGPT 형식 대화형 모델 ‘ChatGPT’와 업그레이드 버전 GPT-4

우리가 일반적으로 사용하고 있는 챗봇(chatbot)은 음성 혹은 문자로 대화하는 기능을 가지고 특정 작업을 수행하도록 제작된 프로그램이다. 1964년 MIT(Massachusetts Institute Technology) 인공지능연구소 교수인 조셉 웨이젠바움(Joseph Weizenbaum)이 역사상 최초의 챗봇인 Eliza를 발명한 것을 기점으로, 구글의 Siri, Amazon의 Alexa 등 스마트 챗봇이 급부상하기 시작했다.²⁶⁾ 일상적인 대화를 하며 해답을 주는 초기 버전 챗봇은 기대했던 인간 대 인간의 대화 수준을 제공하지 못했을 뿐 아니라 간단한 질의응답(Question & Answering) 태스크조차 수행하지 못해 대화에 맞지 않는 답변을 제공하는 소통 불가 상태를 초래했다.

그러나 InstructGPT의 탄생은 사용자의 의도에 맞춘 ‘정렬(align)’된 결과값을 도출을 가능하게 하였고, ChatGPT는 이를 기반으로 사용자의 선호도를 추론하여 그에 알맞은 텍스트를 생성하는 능력을 갖추게 되었다. 기존 대화형 인공지능의 대화 수준을 월등히 뛰어넘을 뿐만 아니라, 사용자의 대화 내용 및 전체 맥락을 고려하며 사용자가 입력한 텍스트를 상황에 따라 분석하며 대화를 이어가는데

20) Ouyang, Long, et al., “Training Language Models to Follow Instructions with Human Feedback.”, *Advances in Neural Information Processing Systems* 35, 2022, p.3.

21) 위의 글, p.3.

22) 위의 글, p.9.

23) 위의 글, p.3.

24) OpenAI의 Completions API인 text-davinci-002, text-davinci-003 등은 모두 RLHF기법으로 강화학습한 InstructGPT 모델이다. ChatGPT의 모델인 gpt-3.5-turbo가 API로 공개되기 전, 가장 대표적인 모델이었으며 스탠퍼드에서 발표한 Alpaca 7B 모델도 데이터 생성에 text-davinci-003을 사용하였다.

25) 평가자의 선호도를 반영하는 학습 방식은 같으나 학습 데이터 수집 방식에 차이가 존재한다. ChatGPT의 경우, 대화 형태의 데이터가 필요하여 인간 평가자가 사용자와 인공지능 모델 간의 대화를 제공하고 보상모델 학습 시에는 챗봇과 대화를 하게 하여 돌아온 답변을 순위화하였다.

26) Brian X. Chen, Nico Grant, Karen Weise, “How Siri, Alexa, and Google Assistant Lost the A.I. Race”, 2023.03.15., *The New York Times*, <<https://www.nytimes.com/2023/03/15/technology/siri-alexa-google-assistant-artificial-intelligence.html>>(최종방문일 2023.04.05).

것이 특징이다. 실제로 성능 검증을 위해 사람을 대상으로 설계된 다양한 시험을 풀게 하였을 때 시험의 출제유형별 출제의도를 파악하여 대다수의 인간 응시자보다 높은 점수를 얻은 경우가 많았다.²⁷⁾

OpenAI에서 제공하는 GPT 모델 서비스는 API를 사용하는 방식과 OpenAI에서 제공하는 대화방에서 ChatGPT를 사용하는 방식이 있다.²⁸⁾ GPT-4를 공개한 이후 GPT-4의 성능을 바로 테스트할 수 있는 방법은 ChatGPT Plus 가입 방법밖에 없기 때문에 유료 가입자들의 수가 폭발적으로 증가하였다.²⁹⁾ 현시점에서 GPT-4 API의 경우, 대기자 명단에서 기다려야 하며 OpenAI 데이터 평가 등에 기여한 사용자들이 우선적으로 제공받고 있다. GPT-4의 구체적인 모델 구조 또는 학습 방법에 대해 공개된 자료는 제공되지 않으며,³⁰⁾ RLHF 방식을 활용하여 더 많은 데이터에 강화학습을 하였을 것이라는 사실만 추측하고 있는 상황이다. OpenAI에서 공개한 자료를 기반으로 확인할 수 있는 점은 GPT-3.5에 비해 GPT-4가 대다수의 태스크에서 성능 향상을 보이며 사용자의 의도를 더욱 이해하고 결과에 반영하는 것에 더욱 최적화되었다는 점이다.

이렇게 지속적으로 발전하는 모습을 보이는 ChatGPT와 GPT-4에 대한 관심과 활용 가능성에 대해서는 놀라움, 걱정, 기대, 우려 등 다양한 반응이 일어나고 있다. 그러나 ChatGPT를 직접 개발한 OpenAI에서도 다음과 같은 한계를 내세우면서 GPT 모델의 신뢰도(reliability) 문제를 인정하고 있다. ChatGPT 모델의 RLHF 학습에 사용되는 데이

터는 사용자의 선호도를 표현하는 만큼 ‘진실된 정답값(ground truth)’이 없으며, 충분히 학습되었다면 모델의 성능으로 보아 풀 수 있는 것에 대해서도 답변 자체를 하지 못하는 상황이 발생하는 문제가 있음을 밝혔다. 또한, GPT모델은 입력값의 미세한 차이에 예민하여 같은 질문이라도 사용하는 질문(프롬프트)에 따라 답을 맞힐 수도 오답을 말할 수도 있고, 이러한 문제는 사용자가 반복해서 같은 프롬프트를 입력할 때도 발생하여 프롬프트 사용에 따라 결과가 쉽게 달라지는 문제가 발생한다고 하였다(이 문제는 아래 제3장의 실험을 통해 직접 확인하였다).

3. GPT모델의 신뢰성 향상을 위한 프롬프트 디자인

프롬프트에 따라 제시되는 답의 정확도가 달라지면서 최적화된 프롬프트에 대한 의존도는 프롬프트 디자인, 프롬프트 엔지니어라는 새로운 직종을 탄생시킬 정도로 그 중요도가 부각되고 있다. 프롬프트 디자인은 모델의 최적 성능 달성을 위해 필수적으로 활용되어야 될 요소이다. 프롬프트 디자인 방식은 여러 가지가 있으나 크게 (1) 단순 프롬프트(Simple prompt), (2) 연쇄사고 프롬프트(Chain-of-Thought, CoT), (3) 퓨샷 프롬프트(Few-shot examples) 등 총 세 가지로 분류할 수 있다. 단순(Simple) 프롬프트 방식은 별다른 기법 없이 문제를 풀게 하는 방식으로, 기본(baseline) 프롬프트로 주로 사용되며 일상적으로 사용하는 방법이다.³¹⁾

Chain-of-thought(CoT)³²⁾은 GPT에게 따라야 할 절차를 세부적으로 명시하여 정한 순서에 따라 문제를 풀도록 유도하는 방식으로, GPT가 문제를 쪼개어 해석하며 생각을 연쇄적으로 풀이해나감에 그

27) Katz, Daniel Martin, et al., “GPT-4 Passes the Bar Exam”, SSRN, 2023.

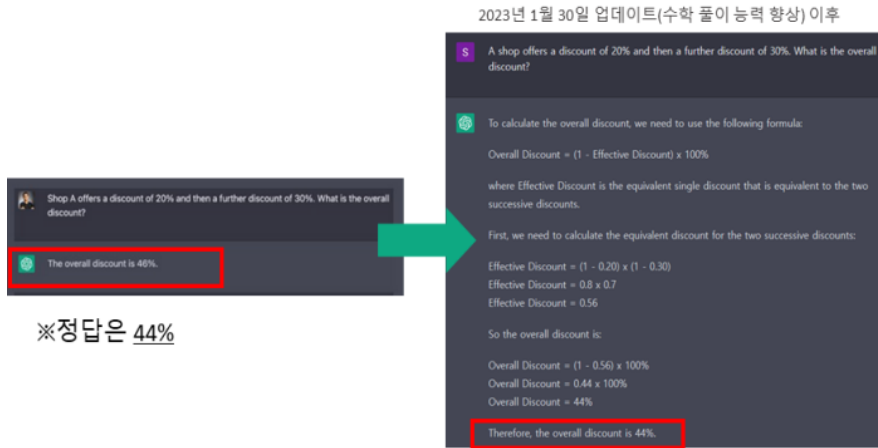
28) 현재 API로는 GPT 3.0대와 ChatGPT와 동일한 모델인 GPT-3.5-turbo 호출이 가능하며, 대화방을 사용하는 ChatGPT는 기본적으로 무료이나 안정된 접속과 빠른 답변 생성을 위해서는 월 \$20의 유료 서비스를 제공하고 있다.

29) 단, 현 시점에서 ChatGPT Plus사용자라도 GPT-4는 3시간에 25개의 대화 메시지만 보낼 수 있다. GPT-4가 공개된 날에는 4시간에 100건이었으며 다음날에는 50건, 현재 3시간 당 25건으로 축소되었다.

30) OpenAI측은 GPT-4 기술문서에서 경쟁 상황과 초거대 모델의 안정성 문제를 고려하여 아키텍처, 하드웨어, 학습 컴퓨팅, 데이터셋 구축, 학습 방법 등 정보를 공개하지 않겠다고 밝혔다. GPT-4 Technical Report, p.2.

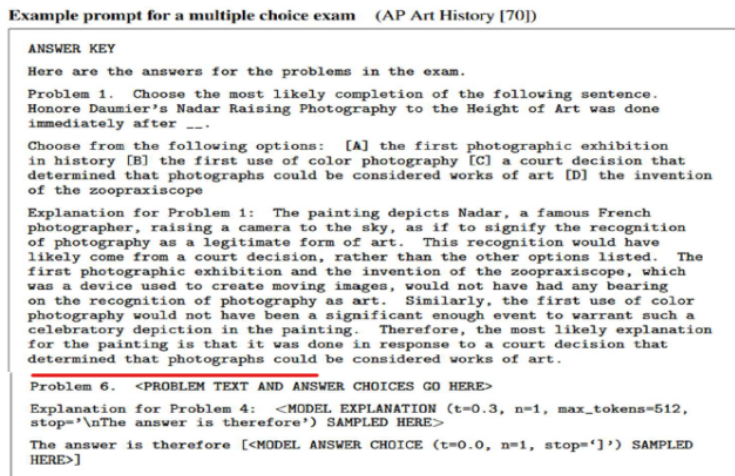
31) Katz et al.(2023) 에서 에세이 작성에 가장 좋은 성능을 보였다. 단, 단순히 ‘문제를 풀어라’에 추가적으로 ‘판례를 참조할 것’, ‘거짓말을 하지 말 것’ 등 지시를 추가한 것으로 보인다.

32) Wei, Jason, et al., “Chain of Thought Prompting Elicits Reasoning in Large Language Models.”, Advances in Neural Information Processing Systems 35, 2022.



※정답은 44%

〈그림 1〉 GPT모델 연쇄사고 프롬프트의 풀이 능력 향상 예시(오답인 46%에서 정답인 44%로 답변)



〈그림 2〉 OpenAI 팀에서 공개한 GPT-4 성능 검증 프롬프트 예시
(역사문제에 대해서 예제를 제공하고, 예제에 따라 답변을 유도하는 방식)

정확성을 높이는 방안 중 하나이다. 아래 예를 보면 Simple 방식으로는 오답이 나왔던 수학 문제를 연쇄 사고 프롬프트 방식으로 입력하자 정답을 맞추는 것을 볼 수 있다.³³⁾ 이러한 사실은 연쇄 프롬프트를 이용한 다른 연구에서 “Let’s think step by step”을 추가하는 것만으로 수학 문제 데이터셋 정확도를 18%에서 79%로 증가했다는 것을 확인할 수 있다.³⁴⁾

OpenAI는 GPT-4의 다지선다형 문제들을 풀기 위하여 퓨샷 프롬프트(Few-shot examples)를 사용하였는데, 이는 사용자가 프롬프트에 문제풀이 예시를 추가해서 질의하는 방식으로 다지선다형 영역에서 높은 성과를 보였다. 아래는 공개한 GPT-4 성능 검증 프롬프트 예시³⁵⁾이다. 본 연구에서는 이리

33) 2023년 1월 30일 업데이트 이후 ChatGPT가 절차를 따라 문제 풀이를 진행하는 모습을 확인할 수 있다.

34) Kojima, Takeshi, et al. “Large Language Models are Zero-Shot Reasoners.”, Advances in Neural Information Processing Systems 35, 2022.

35) GPT-4 Technical Report, p.24. 원본에는 Problem 1부터 5가지 총 5개의 문항 및 풀이가 있었으나 임의로 생략하였다.

한 모든 프롬프트 방식을 사용하고 그 외에 다양한 프롬프트를 디자인하여 실험을 실시하였다.

Ⅲ. 연구 대상 및 방법

1. 연구대상 선정 및 LSAT에 대한 GPT-4 결과 재현

본 연구의 목적은 GPT-4의 법적 추론에 대한 성능을 알아보고자 하는 데 있다. 이를 위해 본 연구는 법학적성시험문제(LEET)를 분석대상으로 선정하여 심층분석하는 한편, 미국의 법학대학원 시험인 LSAT(Law School Admission Test)에 대한 GPT-4의 성능 결과를 재현하여 그 결과를 OpenAI 발표와 비교 분석하였다. 본 연구에서 한국의 LEET시험을 분석대상으로 선정한 이유는 다음과 같다. LEET 시험은 (i) 법적인 추론 내용을 중심으로 하며, (ii) 법학전문대학원협의회에서 공식 풀이가 제공되었기 때문에³⁶⁾ (iii) 연구자가 GPT-4가 제시하는 답의 정답 여부를 확인할 수 있고, OpenAI는 LSAT에 대한 GPT-4의 결과만 제시하였고 사용한 프롬프트나 실제 세부적인 추론 과정을 제시하지 않았기 때문에 본 연구에서는 성능평가를 위해 LSAT 모의고사에 프롬프트를 직접 개발하여 OpenAI가 제시했던 LSAT 시험 결과를 재현하였다.

가. 미국 - LSAT 문제구성 분석 및 GPT-4 성능평가

OpenAI는 GPT-4의 LSAT성능이 GPT-3.5에 비해 크게 향상된 것으로 발표하였다(환산점수 기준 149/180, 백분율 석차 40%에서 163/180, 88%로 향상). 그러나 최종적인 수치만 공개했을 뿐 대상 문제조차 공개하지 않아, 동일한 시험 문제로는 실험할 수 없어, 본 연구에서는 웹상에 공개된 LSAT

모의고사³⁷⁾를 활용하여 발표 내용을 확인하였다. LSAT은 논리게임 위주인 Section I, 문장 단위의 추론 문제인 Section II & III, 긴 지문을 읽고 추리하는 언어추리 문제 Section IV로 구성되어 있다. OpenAI의 기술문서에 따르면 다지선다형(Multiple-choice) 문제는 예제를 포함하는 few-shot 방식으로 진행되었다고 하나,³⁸⁾ 본 연구에서는 필요한 다양한 예제를 구할 수 없었기 때문에 예제없이 풀이과정을 단계별로 정리하여 제시하는 Zero-Shot + Chain of Thought(CoT) 방식³⁹⁾을 채택하였다. 그 예시는 다음과 같다.

위에서 Q1(그림3 왼쪽)를 풀기 위하여 먼저 프롬프트는 1-4 단계로 구분하여 단계적(step by step) 방식으로 명령하여 답을 제시하도록 하였다. 시험문제(Q1)은 5개의 숫자 [0,1,2,3,4,5]로 이루어진 코드가 있고 첫 번째 위치의 숫자가 1일 때 ‘참’인 제시문을 고르는 내용이다. 이 문제에 대해 GPT-4는 풀이 단계(그림3 오른쪽)에서 임의로 알파벳 A-E를 부여하고 부여한 알파벳을 사용하여 문제를 순차적으로 해결하는 모습을 보여주고 있다. (1)지문에서 조건을 추출하고 (2) 그에 대한 값을 대입, (3) 반대가설을 생성하고, (4) 추출한 조건을 확인하여 정답을 추론한다. 이 실험에서 GPT-4의 성능은 OpenAI에서 공개한 성적과 동일한 점수(환산 점수: 163점)를 획득한 것을 확인할 수 있었다. 한편, 본 연구에서는 LEET시험문제 구성과 비교하기 위하여 섹션별 정답률을 다음과 같이 확인하였다.

그림 4와 같이 논리게임이 다수인 Section I은 정답률이 60% 정도이지만 긴 지문을 읽고 정보를 추출하는 Section IV의 경우 매우 높은 정답률을 보였다.

36) 공식해설은 법학전문대학원협의회, 법학적성시험 문제 해설: LEET 추리논증 I (2023- 2019), 에피스테메(방송대출판문화원), 22-107면 참조.

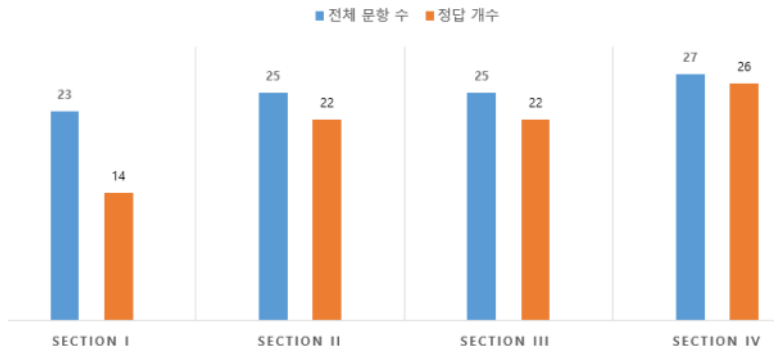
37) The Official LSAT Prep Test 2007, Law School Admission Council.

38) OpenAI, "GPT-4 Technical Report", 2023, p.24.

39) OpenAI의 Cookbook에서 CoT 예제 참조.

<p style="text-align: center;">Prompt</p> <p>Use the following clues to answer the following multiple-choice question, using the following procedure: (1) First, convert the condition to simple logical representation (2) Second, apply the options to the conditions (3) Third, generate counter-facts (4) Fourth, verify whether an option meets ALL the conditions from (1) Context:</p> <p>A company employee generates a series of five-digit product codes in accordance with the following rules: The codes use the digits 0, 1, 2, 3, and 4, and no others. Each digit occurs exactly once in any code. The second digit has a value exactly twice that of the first digit. The value of the third digit is less than the value of the fifth digit.</p> <p>Question: 1. If the last digit of an acceptable product code is 1, it must be true that the (A) first digit is 2 (B) second digit is 0 (C) third digit is 3 (D) fourth digit is 4 (E) fourth digit is 0</p> <p>Work through your logic step-by-step before answering. Solution: (1) First, convert the condition to simple logical representation:</p>	<p>LSAT question</p>
<p>(1) First, convert the condition to simple logical representation: Let's denote the five digits of the product code as A, B, C, D, and E. Conditions: A, B, C, D, E ∈ {0, 1, 2, 3, 4} All digits are unique. B = 2 * A, C < E, Given: E = 1</p> <p>(2) Second, apply the options to the conditions: (A) A = 2, (B) B = 0, (C) C = 3, (D) D = 4, (E) D = 0</p> <p>(3) Third, generate counter-facts: (A) If A = 2, then B = 2 * A = 4. The code would be 241_3. 조건 부합 확인 This satisfies all the given conditions. (B) If B = 0, then A = B / 2 = 0. This doesn't satisfy condition 2 (All digits are unique). (C) If C = 3, then the code would be _30_1. The remaining digits are 2 and 4. If A = 2, B = 4, but this doesn't satisfy condition 4 (C < E), as C = 3 and E = 1. If A = 4, B = 2, but this doesn't satisfy condition 3 (B = 2 * A). (D) If D = 4, then the code would be _3_41. The remaining digits are 0 and 2. If A = 0, B = 0, which doesn't satisfy condition 2 (All digits are unique). If A = 2, B = 4, but this doesn't satisfy condition 3 (B = 2 * A). (E) If D = 0, then the code would be _3_01. The remaining digits are 2 and 4. If A = 2, B = 4, but this doesn't satisfy condition 4 (C < E) as C = 3 and E = 1. If A = 4, B = 2, but this doesn't satisfy condition 3 (B = 2 * A).</p> <p>(4) Fourth, verify whether an option meets ALL the conditions from (1): (A) If the first digit is 2, then the code will be 241_3, which satisfies all the given conditions. Therefore, the correct answer is (A) first digit is 2.</p>	

〈그림 3〉 CoT 프롬프트를 사용한 LSAT 풀이예시



〈그림 4〉 GPT-4를 활용한 LSAT 섹션별 정답 개수

나. 한국 - 법학적성시험(LEET)의 대상 데이터 선정 및 구성분석

국내 법학적성시험인 LEET는 법학전문대학원 입학에 위한 시험으로 크게 ‘언어이해’ 영역과 ‘추리논증’ 영역으로 구분된다.⁴⁰⁾ 본 연구는 GPT-4의 추리능력을 평가함에 목적을 두고 있어 ‘추리논증’ 영역을 연구대상으로 선정하였다.⁴¹⁾ 추리논증 영역

의 유형은 언어추리, 논증평가, 논증분석, 논쟁반론, 논리게임 등 5개 유형으로 구성되어 있고, 이를 모두 실험대상으로 삼았다.

정하였다. GPT-4가 사전에 문제유형과 풀이까지 학습했을 가능성이 있으나, GPT-4는 2021년 9월까지의 데이터를 통해 학습을 진행하였기 때문에 대상 데이터 중 2022년 7월에 시행된 2023학년도 LEET 시험을 학습에 사용했을 가능성이 없다고 판단하였다. 본 연구에서는 2023학년도 LEET 추리논증(홀수형)을 대상으로 다양한 프롬프트를 사용하여 그 결과를 분석해보고자 한다.

40) 법학적성시험 홈페이지 참조 (<https://leet.uwayapply.com/>)

41) 본 연구는 LEET 시험 중에 2023년도 문제를 분석대상으로 선

〈표 1〉 문제 유형 비교

LSAT Section II 문제	LEET 시험 문제
<p>19. Historian: The Land Party achieved its only national victory in Banestria in 1935. It received most of its support that year in rural and semirural areas, where the bulk of Banestria's population lived at the time. The economic woes of the years surrounding that election hit agricultural and small business interests the hardest, and the Land Party specifically targeted these groups in 1935. I conclude that the success of the Land Party that year was due to the combination of the Land Party's specifically addressing the concerns of these groups and the depth of the economic problems people in these groups were facing. Each of the following, if true, strengthens the historian's argument EXCEPT:</p> <p>(A) In preceding elections the Land Party made no attempt to address the interests of economically distressed urban groups.</p> <p>(B) Voters are more likely to vote for a political party that focuses on their problems.</p> <p>(C) The Land Party had most of its successes when there was economic distress in the agricultural sector.</p> <p>(D) No other major party in Banestria specifically addressed the issues of people who lived in semirural areas in 1935.</p> <p>(E) The greater the degree of economic distress someone is in, the more likely that person is to vote.</p>	<p style="text-align: center;">〈보 가〉</p> <div style="border: 1px solid black; padding: 5px;"> <p>ㄱ. A에 따르면, 법에 시위 진압에 관한 근거가 없는 경우, 교통 편의를 위해 시위를 진압할 필요가 있다라도 행정부는 집회의 자유권을 제한하는 시위진압행위를 해서는 안 된다.</p> <p>ㄴ. B에 따르면, 구호품 지급에 관한 사항이 국민에게 이익이 되더라도 법에 그 내용이 규정되어 있지 않으면 행정부는 재난 시 이재민에게 구호품을 지급할 수 없다.</p> <p>ㄷ. C에 따르면, 초등학교 무상급식 정책이 개인과 공공에 영향을 미치는 중요한 사항인 경우, 이 정책은 권리를 제한하지 않는 행정이어도 그 시행에 있어 사전에 법적 근거가 필요하다.</p> </div> <p>① ㄱ ② ㄴ ③ ㄱ, ㄷ ④ ㄴ, ㄷ ⑤ ㄱ, ㄴ, ㄷ</p>

다만, LSAT의 경우 오지선다형의 문제로 지문과 가장 적합한(appropriate) 1개의 답을 고르는 방식이지만 LEET 시험은 지문에 맞는 답을 모두 고르는 형식으로 이루어져 있다. 5개 중 가장 정답에 근접한 하나를 고르는 LSAT는 답변의 개수가 정해져 있는 반면, LEET의 경우 정답의 개수가 정해져 있지 않아 각각의 문항에 대하여 ‘참, 거짓’을 판단해야 정답을 제시할 수 있다는 점에서 정답의 확률에서 큰 차이를 보이고 있다.⁴²⁾

2. GPT-4 최적화를 위한 LEET 시험문제 전처리(Pre-Processing)방법

가. 이미지 파일을 텍스트형식으로 변환

LEET 시험 문제는 텍스트뿐만 아니라 그래프나 표 형식으로 출제되는 경우도 포함하고 있어서 LEET 데이터를 GPT-4에 입력 대상으로 사용하기 위해서는 별도의 전처리 과정이 필수적이다.⁴³⁾ 본

42) LSAT의 경우 정답은 1/5의 확률이나, LEET에서는 정답인 경우의 수가 8개라는 점에서 1/8의 확률로 낮아진다. 실제로 LEET에서는 이런 유형의 문제에서 모두 오답일 경우는 없으나 모델 결과값은 답은 반드시 있다는 조건에도 ‘정답 없음(No correct answer)’을 최종 답변으로 출력하는 경우도 있었다. 반면, 오지선다형 문제에서는 풀이 내용에서 정확한 답을 도출하지 못하는 경우도 선택지 중 하나를 최종 답변으로 도출하였다.

43) GPT-4 Technical Report, p.25에서 OpenAI측은 전처리 과정에서 GPT-4를 사용하면서 ‘최대만 객관적으로 서술(as objectively

연구에서는 GPT-4가 독해가 가능하도록 표의 경우 행 단위로 “열 이름:값”이 들어갈 수 있도록 변경하여 테스트를 진행하여 독해 여부를 확인하였다(아래 표 1-대상유형/표 참조). 또한, 문제 안에 이미지가 그래프가 있을 경우 형태에 대한 간략한 설명문을 작성 후 GPT-4에 설명대로 ‘문자로 표현하기’를 요구하여 독해 여부를 확인하였다(표1-이미지/그래프 참조). 이 과정에서 ChatGPT에서 표현한 문자 표현(ASCII representation)이 문제에 나온 이미지와 유사할 때, 모델이 이해할 수 있는 수준의 설명으로 보고 제시문에 그림 대신 사용하였다. 또한, 지문 일부에 “다음 ㉠이 의미하는 바는 옳은 것은?” 같은 지시기호를 두어 제시한 문제의 경우, <보기>에서 해당 문구로 교체하여 사용하였다(표1-지시기호 참조).

나. 프롬프트 토큰수 최적화 작업

GPT의 경우 영어가 아닌 언어를 사용하는 경우 기술적인 제한이 있다. GPT-3.0가 처리할 수 있는 토큰 수는 2,049개, GPT-3.5는 4,096개이고, GPT-4는 각각 8,192개(GPT-4-8k)와 32,768개(GPT-4-32k)의 토큰을 사용할 수 있다.⁴⁴⁾ 그러나 한국어와 같

as possible)’을 하여 진행하였다고 설명하고 있으나 구체적인 서술방식은 공개하지 않았다. 또한 OpenAI의 경우 이미지 인식이 가능한 멀티모달 GPT-4도 사용하였다고 밝혔다.

〈표 2〉 시각자료 전처리 예시

대상유형	변환 전	변환 후								
표	<table border="1"> <tr> <td>A등급</td> <td>B등급</td> <td>C등급</td> <td>D등급</td> </tr> <tr> <td>2,000만 원</td> <td>1,500만 원</td> <td>1,000만 원</td> <td>500만 원</td> </tr> </table>	A등급	B등급	C등급	D등급	2,000만 원	1,500만 원	1,000만 원	500만 원	A등급: 2,000만원. B등급: 1,500만원. C등급: 1,000만원. D등급: 500만원.
A등급	B등급	C등급	D등급							
2,000만 원	1,500만 원	1,000만 원	500만 원							
이미지/ 그래프		<p>〈Description of graph〉 In this graph, the y-axis represents the utility and the x-axis represents the budget. The curve is shaped similar to a normal distribution. The mean value is x^*, which is marked with an asterisk (*). The two points x_0 and x_1 are on the x-axis, symmetrically placed around the x^* point.</p>								
지시기호 (밑줄)	<p>선출직과 임명직 공무원의 정채 결정 과정이 다른 경우는 흔하다. 선출직의 경우 장래 선거를 고려하여 ㉠ 주민 효용 극대화를, 임명직의 경우 조직의 확대를 고려하여 ㉡ 예산 극대화를 추구한다. 다음 상황을 생각해 보자.</p> <p style="text-align: center;">〈보 기〉</p> <p>7. 갑이 ㉠을 추구하고 기존 예산이 x_1이면, 신규 예산안은 주민 투표에서 통과될 것이다.</p>	<p>(a) 갑이 “주민 효용 극대화”를 추구하고 기존 예산이 x_1이면, 신규 예산안은 주민투표에서 통과될 것이다.</p>								

은 고유문자를 사용하는 경우 같은 내용의 글이라도 사용 토큰 수가 3배로 증가할 수 있다⁴⁵⁾. 더구나 API없이 GPT-4를 사용시 약 4,000토큰만을 사용할 수 있어 사용가능한 토큰수가 더욱 제한적

이다.⁴⁶⁾

따라서 프롬프트와 답변이 합쳐서 4,000토큰 이내로 유지하기 위해 문제가 포함된 프롬프트의 최대 길이를 2,700 토큰⁴⁷⁾으로 제한하였으며, 초과할 경우 제시문을 수정하는 방법으로 진행하였다. 또한 토큰 수의 제한을 회피하기 위하여 LEET문항에 대한 지시문은 영어로 제시하고, 결과는 simple 프롬프트를 제외하고는 모두 영어로 답변하도록 하였다.⁴⁸⁾

44) OpenAI 제공 모델, <<https://platform.openai.com/docs/models/overview>>: 이는 프롬프트와 답변을 합친 숫자로 OpenAI에 의하면 영문자 4개 또는 영단어 3/4개가 1토큰 정도로 2,800개 토큰은 약 1,500 단어가 된다.

45) OpenAI에서 제공하는 토큰라이저(tiktoken 라이브러리, GPT-3.5-turbo 모델 기준)로 테스트해 본 결과 영문으로 2토큰을 소비하는 ‘hello world’와 동일한 의미를 가진 타언어로 토큰 수를 계산할 시, 독일어(‘Hallo Welt’) 2토큰, 일본어(‘こんにちは世界’) 4토큰, 중국어 간체(‘你好世界’) 5 토큰, 한국어(‘안녕 세상’) 7개의 토큰이 사용된다. 언어에 따라 영어 토큰 수의 15.77배 가량 차이가 나는 경우도 있어 1k 토큰에 일정한 사용금액을 부과하는 OpenAI의 서비스를 감안했을 때 언어에 따라 발생하는 차별이 존재한다. <<https://denyslinkov.medium.com/why-is-gpt-3-15-77x-more-expensive-for-certain-languages-2b19a4dc4bc>>.

46) 초과할 경우 제출한 메시지의 길이가 너무 길어 대화를 새로 시작하거나 짧은 메시지를 입력하라는 오류 메시지가 발생한다는 점을 사용하여 실험해본 결과, 3,300 토큰 이상 입력 시 해당 오류가 발생하는 것을 볼 수 있었다. (2023.4월 기준)

47) OpenAI의 tiktoken BPE 토큰라이저(gpt-3.5-turbo)를 사용하였다.

48) LEET의 한 문항이 평균적으로 약 1,000 토큰임을 감안할 때

3. 프롬프트 디자인 방식

본 연구에서는 앞서 언급한 프롬프트 디자인을 활용하여 다음과 같이 프롬프트를 디자인하였다.

〈표 3〉 연구에 사용된 프롬프트 유형 및 특징

프롬프트 유형	지시문 언어	출력 언어	대화 세션
Simple	영어	한국어	동일 세션
CoT	영어	영어	동일 세션
CoT	영어	영어	문항 당 새로운 세션
CoT + one-shot	영어	영어	문항 당 새로운 세션

가. 단순 프롬프트(Simple)

가장 기본적인 Simple 유형은 출력 포맷에 대한 정보 외 ‘문제를 풀고 풀이 과정을 설명하십시오 (Solve the problem and provide explanations for your solution)’만을 입력하여 문제풀이를 진행하였다. 원문의 형태는 위의 템플릿과 다르게 질문(question)이 가장 위에 위치하고 있으나, 랜덤으로 선택한 다른 연도의 문제를 사용하여 테스트해본 결과 LSAT과 유사하게 질문이 보기(statements) 바로 위에 있는 경우 모델의 정확도가 향상되는 경향을 보여서 데이터를 분리한 후 사용하였다. 또한, 보기에서 옳은 것을 고르는 유형이 아닌 오지선다형의 문제인 경우, 결과 형태를 ‘정답- {choice}. 해설- (1) {explanation}, (2) {explanation}...’ 으로 일괄 변경하여 모델이 무작위로 답변을 출력하는 것을 방지하였다.

2023 LEET 추리형(홀수) 3번 문항:

〈passage〉

Alice : 형사절차에서 추구해야 할 진실은 사건의 진상, 즉 ‘객관적 진실’이다. 그리고 객관적 진실을 발견하기 위해서 사건당사자(피고인, 검사) 못지않게 판사의 적극적인 진실발견의 활동과 개입이 필요하다. 따라서 진실 발견을 위해 필요한 경우, 중대한 절차 위반이 없다면 판사가 사건 당사자의 주장이나 청구에 제약을 받지 않고 직접 증거를 수집하거나조사하는 것도 가능하다. (생략)

〈question〉

다음 논쟁에 대한 분석으로 옳은 것만을 〈statements〉에서 있는 대로 고른 것은?

〈statements〉

- (a) 범죄를 조사하기 위해 구속기간 연장의 횟수 제한을 없애자는 법률개정안에 대해 Alice과 Charlie은 찬성할 것이다.
- (b) ‘법이 정한 적법한 절차를 위반하여 수집된 증거는 설사 그것이 유죄를 입증할 유일하고 명백한 증거라 하더라도 예외 없이 유죄의 증거로 사용할 수 없다’는 법원칙에 대해 Bob은 찬성하지만, Alice은 반대할 것이다.
- (c) ‘피고인이 재판에 출석하지 아니한 때에는 특별한 규정이 없으면 재판을 진행하지 못한다’는 법원칙에 대해 Bob과 Charlie은 찬성할 것이다.

〈choices〉

- ① (a) ② (b) ③ (a),(c) ④ (b),(c) ⑤ (a),(b),(c)

Solve the problem and provide explanations for your solution. Select the most appropriate option from 〈choices〉 to answer the 〈question〉. Do not repeat the statements. Answer only in Korean using this format: 정답- {choice}. 해설- (a) {explanation} (b) {explanation} (c) {explanation}.⁴⁹⁾

OpenAI에서 사용했던 방식대로 풀이가 포함된 예제 5개를 주는 것이 불가능하였다. 예제가 함께 들어가는 few-shot 프롬프트는 예제 문항 및 풀이 1건과 풀어야 하는 문항만으로 대다수 2,000 토큰을 초과하여 일반적인 풀이 절차를 추가하는 것으로 대신하였다.

49) 번역문: ‘문제를 풀고 풀이과정을 설명하십시오. <question>에 답할 수 있는 <choices>에서 가장 적합한 선지를 선택하십시오. statements를 반복하지 마시오. 다음 포맷을 사용하고 한국어로 답하십시오: 정답 - {선택}. 해설 - (a) {해설}. (b). {해설}. (c). {해설}.’

나. 연쇄사고 프롬프트(Chain-of-Thought, CoT)

연쇄사고 프롬프트(CoT) 단독으로 사용한 경우 (1) 제시문 분석, (2) 보기 분석, (3) 최종 결과 추론 등 문제풀이 절차 방식을 설명하여 문제를 풀도록 진행을 하도록 하였다. 문제풀이와 관련하여서는 LEET의 공식 해설집을 참조하여 풀이 과정을 간추렸으며 2023년 문제를 사용하기 전 연도의 데이터를 활용하여 자주 발생하던 문제점들을 방지하기 위해 ‘제시문에 명시되지 않았기 때문에 옳은 추론이 아니다’, ‘제시문을 반복하지 말기’ 등 주의사항을 명시하였다. 또한 ‘step-by-step’을 강조하여 모델의 추론 과정을 중심으로 설명하도록 유도하였다.

2023 LEET 추리형(홀수) 3번 문항:

<passage>

Alice : 형사절차에서 추구해야 할 진실은 사건의 진상, 즉 '객관적 진실'이다. 그리고 객관적 진실을 발견하기 위해서 사건당사자(피고인, 검사) 못지않게 판사의 적극적인 진실발견의활동과 개입이 필요하다. 따라서 진실 발견을 위해 필요한 경우, 중대한 절차 위반이 없다면 판사가 사건 당사자의 주장이나 청구에 제약을 받지 않고 직접 증거를 수집하거나조사하는 것도 가능하다. (생략)

<question>

다음 논쟁에 대한 분석으로 옳은 것만을 <statements>에서 있는 대로 고른 것은?

<statements>

- (a) 범죄를 조사하기 위해 구속기간 연장의 횟수 제한을 없애자는 법률개정안에 대해 Alice과 Charlie은 찬성할 것이다.
- (b) '법이 정한 적법한 절차를 위반하여 수집된 증거는 설사 그것이 유죄를 입증할 유일하고 명백한 증거라 하더라도 예외없이 유죄의 증거로 사용할 수 없다'는 법원칙에 대해 Bob은 찬성하지만, Alice은 반대할 것이다.
- (c) '피고인이 재판에 출석하지 아니한 때에는 특별한 규정이 없으면 재판을 진행하지 못한다'는 법원칙에 대해 Bob과 Charlie은 찬성할 것이다.

<choices>

- ①(a) ②(b) ③(a),(c) ④(b),(c) ⑤(a),(b),(c)

Verify the validity of each statement. A statement is valid if it is true based on the information given in the <passage> and reasoning. Follow these steps:
Step 1. Analyze the <passage> and extract relevant criteria such as key differences, supporting facts for statements, and exceptions.

Step 2. Analyze each <statement> one by one. Verify its correctness using the extracted criteria in Step 1, and use logical inference and reasoning to infer new information, including counter-facts. Do not disregard statements or hypotheses due to unclear passage information. Avoid using unsuitable reasons such as 'The statement is invalid because it is not mentioned in the passage.' Format the result as follows: (a) {Valid/Invalid}, {Explanation}, (b) {Valid/Invalid}, {Explanation}, {Valid/Invalid}, {Explanation}.

Step 3. Based on the results from Step 2, choose the correct answer from the <choices>.

Answer in English. Do not repeat the statements. Let's think step by step.⁵⁰⁾

50) 번역문: '각 진술의 타당성을 검토하십시오. <지문>에 제공된 정보와 추론에 따라 참이면 진술은 타당하다. 다음 과정을 따르시오: Step 1. <지문>을 분석하고 주요 차이점, 진술에 대한 뒷받침 사실 및 예외와 같은 관련 기준을 추출.

Step 2. 각 <statement>를 하나씩 분석. 1단계에서 추출한 기준을 사용하여 정확성을 확인하고 논리적 추론과 추론을 사용하여 반대 사실을 포함한 새로운 정보를 추론하십시오. 지문이 불명확하여도 진술이나 가설을 무시하지 마시오. '문구에 언급되지 않았기 때문에 진술이 타당하지 않습니다.'와 같은 부적절한 이유를 사용하지 마시오. 다음과 같이 결과를 출력하십시오: (a) {옳음/옳지않음}, {설명}, (b) {옳음/옳지않음}, {설명}, (c) {옳음/옳지않음}, {설명}.

Step 3. Step 2의 결과를 바탕으로 <보기>에서 정답을 선택하십시오. 영어로 대답하십시오. 진술을 반복하지 마시오. 순차적으로 생각하십시오.'

다. 연쇄사고 프롬프트 + 원샷 프롬프트 결합 (CoT + one-shot examples)

해당 프롬프트는 토큰 수 제한의 입력한계로 인해 OpenAI의 실험조건인 5개의 예시를 추가하기 어려워 영어로 번역 후 예제를 1개만을 추가하여 사용하였다.⁵¹⁾ One-shot의 형태는 최대한 OpenAI에서 공개한 다지선다형 문제의 프롬프트에 맞춰 작성하였으며 프롬프트에 포함된 예제는 2022년 이전에 시행된 LEET의 문항 중 문항 유형이 같고 내용영역이 같거나 문제풀이 방식이 유사한 문제를 선정하여 자동 번역 후 사람이 재검수하여 사용하였다.

2023 LEET 추리형(홀수) 3번 문항:

ANSWER KEY. Here are the answers for the problems in the exam : Problem 16. We generally distinguish strictly between judgments of fact and judgments of ought. (생략)

<passage>

Alice : 형사절차에서 추구해야 할 진실은 사건의 진상, 즉 '객관적 진실'이다. 그리고 객관적 진실을 발견하기 위해서 사건당사자(피고인, 검사) 못지않게 판사의 적극적인 진실발견의활동과 개입이 필요하다. 따라서 진실 발견을 위해 필요한 경우, 중대한 절차 위반이 없다면 판사가 사건 당사자의 주장이나 청구에 제약을 받지 않고 직접 증거를 수집하거나 조사하는 것도 가능하다. (생략)

<question>

다음 논쟁에 대한 분석으로 옳은 것만을 <statements>에서 있는 대로 고른 것은?

<statements>

- (a) 범죄를 조사하기 위해 구속기간 연장의 횟수 제한을 없애자는 법률개정안에 대해 Alice과 Charlie은 찬성할 것이다.
- (b) 법이 정한 적법한 절차를 위반하여 수집된 증거는 설사 그것이 유죄를 입증할 유일하고 명백한 증거라 하더라도 예외없이 유죄의 증거로 사용할 수 없다'는 법원칙에 대해 Bob은 찬성하지만, Alice은 반대할 것이다.
- (c) '피고인이 재판에 출석하지 아니한 때에는 특별한 규정이 없으면 재판을 진행하지 못한다'는 법원칙에 대해 Bob과 Charlie은 찬성할 것이다.

<choices>

- ①(a) ②(b) ③(a),(c) ④(b),(c) ⑤(a),(b),(c)

Verify the validity of each statement. A statement is valid if it is true based on the information given in the <passage> and reasoning. Follow these steps:

Step 1. Analyze the <passage> and extract relevant criteria such as key differences, supporting facts for statements, and exceptions.

Step 2. Analyze each <statement> one by one. Verify its correctness using the extracted criteria in Step 1, and use logical inference and reasoning to infer new information, including counter-facts. Do not disregard statements or hypotheses due to unclear passage information. Avoid using unsuitable reasons such as 'The statement is invalid because it is not mentioned in the passage.' Format the result as follows: (a) {Valid/Invalid}. {Explanation},(b) {Valid/Invalid}. {Explanation}, {Valid/Invalid}. {Explanation}.

Step 3. Based on the results from Step 2, choose the correct answer from the <choices>.

Answer in English. Do not repeat the statements. Let's think step by step.⁵²⁾

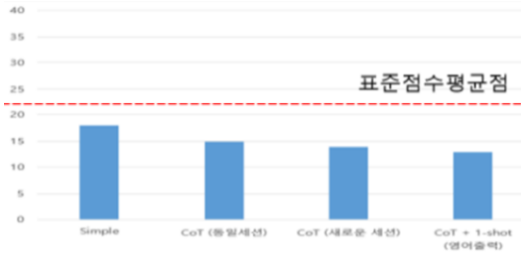
51) 또한, 세션을 새로 사용하는 것과 한 세션을 최대한 길게 사용하는 방법 2가지를 사용하여 그 차이를 검토하였다. 실제로 API를 사용할 경우 호출할 때마다 새로운 대화를 시작하는 것이나 다름 없어 이전 대화를 다시 입력값으로 넣는 등의 별도의 설정 없이 이전 대화가 기록되지 않는다. 즉, API를 활용할 경우 이전의 정보가 GPT의 응답에 반영되지 않으나, 본 연구를 진행한 GPT-4의 ChatGPT 대화방 환경에서는 한 세션에서 대화가 이어지면 이전 대화의 내용이 기억된다는 특징이 있다. 이에 따라 앞서 문제의 방식을 기억하는 장점도 존재할 수 있겠으나 예제를 추가한 CoT+one-shot의 경우, 독립적인 문제임에 불구하고 이전 문항의 내용을 문제풀이에 추가하여 풀이를 작성하는 경우도 있어 CoT+one-shot 테스트에서는 제외하였다.

52) 각주 49와 동일.

IV. 연구 결과 및 분석

1. 연구결과

2023년 LEET를 GPT-4를 활용하여 풀어본 결과는 아래 그림과 같다.



〈그림 5〉 프롬프트 유형별 정답 개수 (40문항)

〈표 4〉 프롬프트 유형별 정답 개수

프롬프트 유형	문항별 정답 개수	선택지별(o,x) 정답 개수*
Simple	18/40	73/96
CoT (동일 세션)	15/40	62/96
CoT (문항 당 새로운 세션)	10/40	58/96
CoT + one-shot (영어 출력)	13/40	62/96

* 선택지별 정답 총 개수: <보기>에서 모두 고르는 문제 = 32 x 3 제시문(Γ, L, D). 각 선택지가 o,x를 맞추는 문제라고 가정하였을 때 점수로, 8개의 오지선다 문제는 제외하고 계산하였음.

4가지 유형의 프롬프트를 활용한 실험결과에 의하면 전체에서 50% 이하의 정답수를 보여주고 있다. 특히 CoT를 단독으로 적용한 프롬프트와 원샷(one-shot)을 혼합한 프롬프트 형식에서는 20%-32.5%에 불과한 정답률을 보여주고 있다. 4가지 프롬프트 유형 중 가장 좋은 성과를 보인 것은 Simple 프롬프트였으며 40문항 중 18개의 정답을 추론하였다. 이는 2023년 기준 표준점수로 환산하였을 때, 49.4점 정도로 누적비율 89.7%에 해당한다.⁵³⁾ 상

위 10%의 결과를 보인 미국 변호사시험 (UBE, Uniform Bar Exam)⁵⁴⁾, LSAT⁵⁵⁾의 결과와 대비하자면 상당히 저조한 점수라고 볼 수 있다.⁵⁶⁾ 각각의 보기에 대한 ‘참과 거짓’의 선택형으로 구분하여 GPT-4가 제시한 답안을 분석하였을 때 전체 평균은 66.4%에 이르렀다. 이 중에서 simple 프롬프트는 76%에 도달하여 최고의 성적을 보여주고 있다. 그 동안 OpenAI가 GPT-4의 신뢰성을 높이기 위하여 프롬프트 디자인 방식을 강조했다라는 점에서 Simple 프롬프트가 다른 유형의 프롬프트 보다 높은 정답률을 보이는 것은 특이한 점이다. 그러나 이러한 현상은 법률영역에서 이미 언급된 바가 있다. GPT-4 이전 버전인 ChatGPT를 미국의 한 법학전문대학원 교내시험 다지선다형 문제에 적용해본 연구⁵⁷⁾에서 ChatGPT가 CoT와 정답 후보 순위화(Ranking order)기법에 비해 예시가 주어지지 않은 Zero-shot 프롬프트(simple형과 유사)로 더 좋은 성능을 보인 적이 있다. 특히 해당 연구에서 ChatGPT는 서술형 법률문제 답안지를 작성할 때 관련 문제를 식별하는 데 어려움을 겪었고, 학생들과는 달리 피상적인 사실만에 규칙을 적용하는 특성이 나타났으며, 개방형 질문을 답할 시 성능이 급감하였다. 이러한 미국의 실험결과는 LEET 시험에서 정답을 모두 고르는 개방형 질문에 대한 저조한 정답률을 보인 실험결과와 일치한다.

html?idxno=739218>.

54) 미국 변호사시험 공식 홈페이지, <https://www.americanbar.org/groups/legal_education/resources/bar-admissions/bar-exams>.

55) 미국 법학대학원 공식 홈페이지, <<https://www.lsac.org/lsat>>.

56) OpenAI에서 진행한 실험의 경우, 그 공정성과 무결성에 대해 의혹이 발생하고 있다. 실제로 LSAT 문제를 포함한 평가 데이터가 이미 GPT-4의 학습 데이터에 속해 있었으며 OpenAI가 데이터 오염(data contamination)에 대비해서 사용한 문자열 매칭은 부족하다는 의견과 특정 문제는 제외하고 낸 결과이므로 공정한 결과가 아니라는 의견도 있다., <<https://towardsdatascience.com/the-decontaminated-evaluation-of-gpt-4-38a27fc45c30>>. 다만 본 실험연구에서는 정답을 맞추었다고 해도 실제로 GPT-4를 활용해 확보하고자 하는 것은 단순히 예측만 잘하는 것이 아닌 그 예측에 대한 설명, 즉 법적 추론이 가능한 모델이므로 그 문제풀이를 분석하고 한계점을 명확히 함에 있다.

57) Choi, Jonathan H., et al., "ChatGPT goes to law school.", SSRN Electronic Journal, 2023, p.4.

53) 2023년 표준점수, <<http://www.lec.co.kr/news/articleView>>.

2. 연구결과를 통해 확인한 GPT-4의 문제점 분석

본 연구는 GPT-4가 LEET 문제를 풀이하는 과정에서 아래와 같은 몇 가지 유형의 문제점을 도출할 수 있었다.

가. 상식 추론과 관련한 능력 부재

GPT-4는 제시문에서 명시되지 않은 암시적 사실을 추론하지 못하거나, 한 개의 주장을 이해하여 새로운 사실(결론)을 도출하는 추론 방식에는 약점을 보여주었다. 예를 들어 아래 제시문과 같은 2023년 LEET 문항 공식해설에서는 “구속기간 연장의 횟수 제한을 없애자는 법률개정안에 대해서, 객관적 진실의 발견을 강조하는 Alice는 찬성할 것이나,”라고 표현되어 Alice가 법률개정안에 대해 찬성임을 알 수 있으나, GPT-4는 “Alice는 판사의 적극적인 진실 발견 활동과 개입이 필요하다고 주장하지만, 구속기간 연장에 대한 의견은 밝히지 않았다.”라고만 이해하여 법률개정안에 찬성 여부를 추론해내지 못하였다. 이처럼 GPT-4는 일반 상식과 추론을 통해 알아내는 새로운 지식을 도출하지 못함을 알 수 있다.

오류	[오류] 3번 문항 풀이 (Simple 프롬프트 사용):
정답	공식 해설 ⁵⁸⁾ : (b), (c) GPT-4: (b)
해설	논거 (a)에 대한 해설: 공식 해설: (a) 구속기간 연장의 횟수 제한을 없애자는 법률 개정안에 대해서, 객관적 진실의 발견을 강조하는 Alice는 찬성할 것이나, 객관적 진실의 발견이 중요하지만 인권침해를 통해서 얻어낸 객관적 진실은 정당성을 획득할 수 없다고 주장하는 Charlie는 반대할 것이다. (a)는 옳지 않은 분석이다. GPT-4: (a) Alice는 판사의 적극적인 진실 발견 활동과

개입이 필요하다고 주장하지만, 구속기간 연장에 대한 의견은 밝히지 않았다. Charlie는 판사가 원칙적으로 중립적인 판단자의 역할을 수행하되, 인권침해를 감시하고 바로잡는 역할을 해야 한다고 주장했지만, 구속기간 연장에 대한 의견은 밝히지 않았다. 따라서 (a)는 해당하지 않는다.
--

나. 일관성과 관련한 능력 부재

GPT-4는 동일한 제시문에 대해서 같은 프롬프트를 사용하여도 다른 답을 도출하는 문제점을 노출 시킨다. 아래 예시에서는 동일한 프롬프트에 대해 각각 다른 정답과 풀이 결과를 볼 수 있다. 추리 논증의 경우 수학 문제풀이 과정과 같이 일관된 결과와 풀이 과정이 신뢰성과 밀접한 관련이 있다. GPT-4는 같은 프롬프트를 반복하여 사용할 경우 답이 달라질 수 있어 일관성을 기대할 수 없는 문제가 발생한다.

오류	[오류] 5번 문항 풀이 (Simple 프롬프트 사용):
정답	{GPT-4} 첫 번째 풀이 ⁵⁹⁾ : (b), (c) {GPT-4} 두 번째 풀이: (a), (c)
해설	논거 (a)에 대한 해설: 첫 번째 풀이: (a) P에 속한 회사들의 부동산 합산 가격이 5억 원 이하라면, 견해1에 의하여 과세하든 견해 2에 의하여 과세하든 과세 총액이 달라지지 않는다고 주장하지만, 실제로는 견해1에 의한 합산과세와 견해2에 의한 개별과세에서 과세 총액이 달라질 수 있다. 따라서 이 설명은 선택지에 해당하지 않는다. 두 번째 풀이: (a) P 그룹의 부동산 합산 가격이 5억 원 이하인 경우, 두 견해 모두 0.5% 세율이 적용되므로 과세 총액이 같다.

58) 법학전문대학원협의회, 앞의 책, 22-107면.

59) 위의 책, 22-107면.

다. 판단기준 제시와 관련한 문제

GPT-4는 경우에 따라서는 답변에 대한 어떠한 판단기준도 제시하지 못하는 경우가 있어 그 근거를 신뢰할 수 없는 문제점을 야기하고 있다. 아래 예시는 GPT-4가 판단의 기준이나 선택지의 주제가 될 수 있는 정보를 특정하지 않고 지문에 해당 관련 사실이 없으므로 분석이 불가하다는 결론을 내린 것을 보여주고 있다. 또 다른 분석결과에서는 선택지에서 ‘만약 X가 참이라고 한다면’ 등 가정문이 있을 때 가정을 참이라고 여기거나 무엇이 참이고 거짓인지 알 수 없어 선택을 포기하는 경우도 있었다. 이런 문제점은 논리적 기본원리를 충분히 학습하지 못한 결과에서 기인한 것으로 추측된다.

오류	[오류] 17번 문항에 대한 답 (Simple 프롬프트 사용):
정답	공식 해설 ⁶⁰⁾ : (a), (b) GPT-4:(c)
해설	<p>논거 (a)에 대한 해설: 공식 해설: (a) ㉔은 두 문장 ‘철수는 노트북 또는 핸드폰을 가지고 있다.’와 ‘후식으로 커피 또는 녹차를 드립니다.’를 구성하는 개별문장들이 모두 참인 경우를 고려하고 있으며, 이때 원래의 두 문장의 진리치가 다르다는 것을 근거로 ‘또는’의 의미가 다르다고 주장하고 있다. 어떤 주장이 있을 때, 그 주장이 전제하는 것은 그 주장과 상충할 수 없으므로, ㉔은 옳은 분석이다.</p> <p>GPT-4: (a) 이 설명은 문제에 대한 분석이 아니기 때문에 옳은 분석이 아니다.</p>

라. 정답-풀이 근거의 타당성 문제

GPT-4가 제시한 답변에서 일견 공식해설과 매우 유사하게 풀이하여 정답을 맞춘 것 같지만, 실

제로 반대로 해설하는 사례가 확인되었다. 아래 제시문을 보면 정답을 유추하기 위해 풀이를 <보기>의 논거에 대입하여 올바르게 결론을 도출하였음에도 선택지를 고르는 과정에 오류가 있음을 알 수 있다. 예시에서 도덕적으로 옳지 않은 의무가 사라지는 것까지는 올바르게 추출하였으나, 선택지에서 제외되어야 하는 보기를 선택하는 오류를 범하고 있다. 이는 문제풀이 뒤 정답을 확인하는 과정에서 정답과 풀이가 연계되지 않아 발생할 수 있는 오류로 보인다.

오류	[오류] 12번 문항에 대한 답 (Simple 프롬프트 사용):
정답	공식 해설 ⁶¹⁾ : (a) GPT-4: (a),(b)
해설	<p>논거(b)에 대한 해설: 공식 해설: (b) 갑은 일반적으로 약속을 한 사람은 그 약속을 지켜야 할 의무가 있지만, 살인 약속은 살인 자체가 나쁘기 때문에 그 약속을 지켜야 할 의무가 없다고 말하고 있다. 을 역시 일반적으로 약속은 그 약속을 지켜야 할 의무를 부과하지만, 살인과 같이 X가 도덕적으로 옳지 않고 X를 하지 않을 의무가 X를 하기로 한 약속을 지키는 의무보다 더 강할 때 그 약속을 지켜야 할 의무가 사라진다고 보고 있다. 따라서 (B)은 옳지 않은 판단이다.</p> <p>GPT-4: (b) A가 B에게 한 약속을 지킬 의무가 있는지에 대해서 Alice와 Bob은 의견을 달리할 것이다. Alice는 도덕적으로 옳지 않은 약속일 경우에 그 약속을 지킬 의무가 생겨나지 않는다고 주장하며, Bob은 X가 도덕적으로 옳지 않고 X를 하지 않을 의무가 X를 하기로 한 약속을 지키는 의무보다 더 강할 때 그 약속을 지켜야 할 의무가 사라지는 것이라고 주장한다. 따라서 이 설명은 선택지에 해당한다.</p>

60) 위의 책, 22-107면.

마. 핵심조건 적용과 관련된 문제

GPT-4가 답을 분석할 때 기억해야 할 조건이 2개 이상일 경우 두 번째 조건을 때때로 대입하지 못하는 것을 확인할 수 있었다. LEET에서는 규범, 원칙 등이 주어지고 사례에 대입하는 유형이 많은 부분을 차지하고 있다. 그러나 아래 예시가 보여주듯이 중요한 조건을 성공적으로 추출하더라도 적용 단계에서 추출한 모든 조건을 적용하지 못하는 경우가 나타났다. 공식해설과 GPT의 답변을 비교 분석한 결과, 두 가지 조건 간의 상관관계를 포함시켜 풀이를 해야하나 GPT-4의 경우 이에 대한 어려움을 가지고 있는 것으로 보인다.

오류	[오류] 05번 문항에 대한 답 (Simple 프롬프트 사용):
정답	공식 해설 ⁶²⁾ : (b) GPT-4: (b), (c)
해설	논거 (a)에 대한 해설: 공식 해설: (a) 무도장업자 갑이 주류판매로 2019. 6. 20. 영업정지 1개월을 받은 후, 이를 알고 있는 을에게 2020. 6. 30. 그 영업을 양도하고 신고를 마쳤는데, (1) 을이 2020. 7. 25. <u>접대부 고용과 주류판매로 적발되었다면</u> . (2) 행정청은 을에게 영업정지 3개월의 처분을 한다. GPT-4: (a) 옳지 않다. Alice는 주류 판매로 2019. 6. 20. 영업 정지 1개월을 받았다. 그러나 양도한 이후, Bob은 접대부 고용과 주류 판매로 적발되었다. (1) <u>접대부 고용의 처분 기준은 영업 정지 2개월이고, 주류 판매의 처분 기준은 영업 정지 1개월이다. 서로 다른 위반 사항이므로 무거운 처분 기준인 영업 정지 2개월을 적용해야 한다.</u>

바. 소결론

이상과 같이 GPT-4를 통해 LEET의 문제를 풀어봄으로써 모델의 추론 과정을 확인해보고 일반사람들이 추상적으로 언급하고 있는 GPT-4의 문제점을 실증하였다. 아직까지 GPT-4가 추리논증에 대한 충분한 학습이 없었거나 논증구조를 잘 이해하지 못하고 있는 것으로 보인다. 왜냐하면 논증구조를 학습하여 이해하고 있다면 질문의 형식이 체계적일수록 생성되는 답변의 타당성이 더 높아져야 하나, 실제로는 그렇지 않았기 때문이다. GPT 계열의 언어모델(Language Model)은 확률을 기반으로 하고 있기 때문에 지속적인 데이터 학습을 통해 정답률은 향상될 수 있을 것으로 예상된다. 그러나 GPT-4가 진정 사람처럼 생각하고 이해할 수 있어 법률전문가를 대체하는 AGI(Artificial General Intelligence)가 되기 위해서는 단순한 정답 제시를 넘어 타당한 근거를 기반으로 하는 논증과정을 보여줄 필요가 있는데, 확률에 기반한 정답의 향상이 논증구조의 이해로 연결될지에 대해서는 의문이 든다.

V. 결론

본 연구는 2021년부터 2023년까지 시행된 LEET 시험문제를 대상으로 OpenAI에서 제공한 프롬프트를 포함한 다양한 방식의 프롬프트를 디자인하여 GPT-4 모델의 추론능력을 검토하고 각 프롬프트에 따른 특성을 분석하였다. 특히 판단기준이나 근거를 적용해야 하는 논증에 대한 인공지능의 해석 능력을 중심으로 평가를 하였으나 확률 기반 모델이 논증 타당성(validity)을 대체하기에는 아직까지 부족하다는 결론이다. 그럼에도 불구하고 우리는 GPT-4와 같은 언어 모델의 결점을 최대한 보완하며 발전시켜 나가야 할 필요성도 보인다.

현재 인공지능 서비스는 단순히 연구자나 인공지능 회사들이 눈여겨보는 기술만이 아니고, 일반

61) 위의 책, 22-107면.
62) 위의 책, 22-107면.

사람들의 일상생활에서 누구나 사용하고 누구나 활용할 수 있는 서비스로 인식되어 가고 있다. 특히 GPT-4가 공개됨에 따라 상대적으로 성과가 부진했던 자연어 분야의 인공지능이 큰 관심을 받고 있으며, 인간과 유사한 의사결정처리에 활용하기 위해 다양한 연구가 공개되고 있다.⁶³⁾ OpenAI의 GPT-4를 통해 인공지능의 무한한 가능성을 발견한 현재 지금 GPT모델을 대체할 수 있는 언어모델 개발 역시 급부상하고 있는 연구 주제다. 스탠퍼드 대학의 연구팀은 OpenAI의 모델을 통해 양질의 학습 데이터를 확보하여 학습한 Alpaca 7B 모델을 공개하였으며, 국내에서는 네이버 지식인 데이터를 기반으로 생성된 데이터를 학습한 KoAlpaca⁶⁴⁾나 국내 법률 데이터를 학습한 LawGPT⁶⁵⁾ 등 한국어에 최적화된 언어 생성 모델 개발이 활발히 진행되고 있다.

이렇게 각각 분야에서 초거대 언어 모델 구축 및 시스템 개발에 매진하는 상황 속에서 다른 전문가들은 모델 구축 과정에서 발생할 수 있는 불분명한 데이터 처리과정, 프라이버시 침해, 범죄 활용 가능성 및 AI로 인한 일자리 상실 등 AI가 가져올 격변에 대한 우려를 표하고 있다. 사회 유력자 단체들

은 제도적인 안전장치 없이 인공지능의 무분별한 발전을 사회적 문제로 보고 6개월간 개발 중지를 요구하기도 하였다.⁶⁶⁾ 우리 정부 역시 ChatGPT와 같은 사기업의 인공지능 서비스를 사용 시 정보 유출 가능성을 주요사안으로 보며 정보 주체의 권리 강화, 개인정보보호 생태계 조성과 같은 여러 가지 대책안을 고려하고 있는 중이다.⁶⁷⁾ 공공정보 유출을 우려하여 공공기관에서의 GPT 사용을 금지한 것과 민간기업에서 전면적인 사용금지로 이어진 것도 정보 유출에 대한 위험이 상당히 높기 때문인 것으로 보인다.⁶⁸⁾

그러나 그 어느 분야보다 빠른 속도로 혁신을 거듭하는 인공지능 연구의 현실을 봤을 때, 개발된 기술의 사용을 금지하여 발전의 속도를 늦추는 것보다 기술적 한계를 극복할 수 있는 정확한 설계 및 정책적 고려를 통해 인공지능을 발전시키면서 공존하는 방향으로 나아가야 할 것으로 보인다.

이미 법률영역에서도 GPT-4의 성능에 기반하여 모든 법률 업무에서 전문가를 대신할 것이라는 기대와 변호사 대체라는 걱정이 교차되고 있는 상황이다. 그러나 본 연구 결과에 비추어 본다면 GPT-4가 법률영역에서 사람 대신 활용되기까지는 상당한 시간이 소요될 것으로 보인다. 왜냐하면 GPT-4 등 초거대 언어 모델이 법률전문가를 대체하기 위해서는 일관적이고 객관적인 추론적 사고가 요구되나, 이 정도의 합리적인 판단 능력까지 도달하지 못하고 있기 때문이다. 그럼에도 불구하고 이번 연구는 초거대 인공지능이 논리학 등의 기초 영역에서 충분한 학습을 거치면 사람의 추론을 지원하는 성능을 발휘할 수 있다는 가능성을 보여 주었다. 이미

63) 그 일례로 자동 문제해결 시스템 Auto-GPT (<https://github.com/Significant-Gravitas/Auto-GPT>)가 있다. Auto-GPT는 GPT-4를 활용하여 사용자가 설정한 목표에 따라 필요한 수행 내용을 생성하고, 인터넷 검색 등 모듈을 자체적으로 실행하여 문제를 해결한다. 2023년 3월에 공개된 BabyAGI (<https://github.com/yoheinakajima/babyagi>) 역시 GPT-4 모델의 연계를 통해 인간의 사고 흐름 모방하는 시스템이다. 이렇듯 모델 자체를 사용하는 데 그치지 않고 여러 기능을 가진 모듈들과 결합하여 GPT-4의 시스템적 한계를 극복하려는 시도가 이루어지고 있는 실정이다.

64) 스탠퍼드 연구팀에서 공개한 Alpaca에서 영향을 받은 프로젝트로 1.0의 경우 Alpaca의 학습 데이터를 한국어로 번역하여 사용하였으며 1.1에서는 네이버 지식인 데이터를 기반으로 자체 생산하여 학습하였다. 완성된 모델의 기초가 되는 파운데이션 모델은 1.0에는 Alpaca와 동일한 Llama로 테스트하였으나 영문 데이터가 대부분인 Llama 모델로 성능의 한계가 있어 1.1에는 EleutherAI에서 공개한 Polyglot-ko-12.8B 모델을 사용하였다. Polyglot-ko-12.8B 모델은 한국어 데이터 863GB를 학습한 언어모델이며 이를 공개한 EleutherAI는 인공지능 발전의 투명성을 주장하며 오픈소스를 중시하는 비영리 연구 단체로 OpenAI의 대척점에 서 있다고 볼 수 있다.

65) 국내 리걸테크회사인 인텔리콘이 개발한 모델로 300만 건 이상의 법령, 판례, 법률논문 등을 학습하였으며 관련 법률 조항과 판례까지 제공해주는 시스템으로 개발하였다고 한다.

66) 정원식, 본격화되는 'AI 디스토피아' 논쟁...“안전장치 마련 전까지 AI 차세대 모델 중단해야”, 2023.4.02., 경향신문, <<https://n.news.naver.com/article/032/0003214744?sid=104>>, (최종방문일 23.04.04).

67) 박은주, 양날의 검 '챗GPT'에 베이지 없으면... AI 보안위협 선제 대응 필요, 2023.05.18., 보안뉴스, <<https://www.boanews.com/media/view.asp?id=118177>>, (최종방문일 23.05.29).

68) 강일용, [단독] 정부 "챗GPT 쓰면 공공 정보 유출 우려"...우선 국정원 허락 받아라, 2023.05.03., 아주경제, <<https://www.ajunews.com/view/20230503112621151>>, (최종방문일 23.05.29).

법률영역의 인공지능 활용을 위해 단순히 데이터만 추가하는 방식이 아니라 보다 근본적인 활용방법을 모색하기 위한 많은 연구가 진행 중이다. 투명한 설계 및 가공 과정을 거친 법률 말뭉치를 활용한 사전학습을 통한 이해도 향상⁶⁹⁾, 법률전문가와와의 협력을 통한 효과적인 피드백 시스템 구축⁷⁰⁾, 법 영역에 특화된 기술 개발⁷¹⁾을 추진함으로써 GPT 활용 가능성은 더욱 높아질 것이라 예상되며 그 개발에 대한 구체적인 논의는 추후 연구과제로 제시하고자 한다.

<참고문헌>

국내 문헌

[단행본]

법학전문대학원협의회, 「법학적성시험 문제 해설: LEET 추리논증 I (2023-2019)」, 에피스테메(방송대출판문화원), 2022.

외국 문헌

[논문]

Bommarito, Michael James and Katz, Daniel Martin, “GPT Takes the Bar Exam.”, SSRN, 2022.
Brown, T., Mann, B., Ryder, N., Subbiah, M.,

Kaplan, J. D., Dhariwal, P., & Amodei, D., “Language Models are Few-Shot Learners.”, Advances in Neural Information Processing Systems, 33, 2020.

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. B., “ChatGPT Goes to Law School.”, SSRN, 2023.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P., “GPT-4 Passes the Bar Exam.”, SSRN, 2023.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y., “Large Language Models are Zero-Shot Reasoners.”, Advances in Neural Information Processing Systems 35, 2022.

OpenAI, GPT-4 Technical Report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R., “Training language models to follow instructions with human feedback.”, Advances in Neural Information Processing Systems 35, 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., “Language Models are Unsupervised Multitask Learners”, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I., “Attention is all you need.”, Advances in Neural Information Processing Systems 30, 2017.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D., “Chain of Thought Prompting Elicits Reasoning in Large Language Models.”, Advances in Neural Information Processing Systems 35, 2022.

69) 국내에서는 리걸테크 기업인 로엔컴퍼니가 적용법조를 예측하는 한국어 법률 AI 데이터셋 KLAID(<https://klaid.net/>)를 배포하였으며 Lbox는 태스크용 사건명, 법령명, 요약 태스크 관련 법률 데이터셋을 공유하고 사전학습 모델을 공개 (<https://github.com/lbox-kr/lbox-open>)하였다. 이처럼 양질의 한국어 법률 자료 공유가 활성화된다면 앞으로 인공지능 모델의 법률 도메인 최적화가 충분히 가능할 것이다.

70) InstructGPT의 특징은 인간 평가자의 피드백을 기반으로 강화학습 하였기 때문에 인간인 사용자의 보편적인 선호도에 부합한 결과를 도출한다는 점이다. 이와 같이 법률 도메인에서는 법률전문가의 피드백이 법률 모델의 기능 향상에 필수적인 요소이다.

71) 국외에서는 자동 계약 분석(Automated Contract Review), 인공지능을 활용한 관련 법률/판례 검색 시스템 등 리걸테크기술의 개발이 활발히 진행되고 있다. 국내에서는 법적인 한계, 기존 법조계의 반발 등 여러 가지 요소로 인해 발전이 지연되었음에 불구하고 지속적으로 관련 기술이 개발되고 있는 추세다. GPT-4 공개에 따라 여러 산업계에서 도입을 고려하고 있는 만큼 인공지능 모델의 법적 추론능력을 활용한 지원도구를 개발하기 위해 관련 기술도 발전해나가야 할 것이다.

기타 문헌

- 강일용, [단독] 정부 “챗GPT 쓰면 공공 정보 유출 우려”...우선 국정원 허락 받아라, 2023.05.03., 아주경제, <https://www.ajunews.com/view/20230503112621151>, (최종방문일 23.05.29).
- 박은주, 양날의 검 ‘챗GPT’에 베이지 않으려면... AI 보안위협 선제 대응 필요, 2023.05.18., 보안뉴스, <https://www.boannews.com/media/view.asp?idx=118177>, (최종방문일 23.05.29).
- 전남혁, ‘PPT 만들어줘’ 몇 초만에 딱딱...MS, 워드, 엑셀에도 AI 탑재, 2023.03.17., 동아일보, <https://www.donga.com/news/It/article/all/20230317/118379443/1>, (최종방문일 23.04.01).
- 정영훈, ‘1억명 이용’ AI 챗GPT...“월 20달러 유료화”, 2023.02.02., KBS 뉴스 <https://news.kbs.co.kr/news/view.do?ncd=7596001>, (최종방문일 23.04.01).
- 정원식, 본격화되는 ‘AI 디스토피아’ 논쟁...“안전 장치 마련 전까지 AI 차세대 모델 중단해야”, 2023.4.02., 경향신문, <https://n.news.naver.com/article/032/0003214744?sid=104>, (최종방문일 23.04.04.).
- 차병섭, MS검색엔진 Bing, 오픈AI 업고 구글 추격... 방문자 수 16% ↑, 2023.03.23., 연합뉴스, <https://www.yna.co.kr/view/AKR20230323084400009>, (최종방문일 23.04.01).
- 최재식, “챗GPT시대, 어떻게 AI 시장 주인공으로 살 수 있을까?”, 2023.3.31., 인공지능 신문, <https://www.aitimes.kr/news/articleView.html?idxno=27683>, (최종방문일 23.04.01).
- 최창현, GPT-4 발표한 오픈AI ‘샘 알트만’, “AI가 사회를 재구성할 것이며 위험을 인정합니다, 이것이 조금 두렵습니다”, 2023.3.19., 인공지능 신문, <https://www.aitimes.kr/news/articleView.html?idxno=27595>, (최종방문일 23.04.01).
- Caroline Hill, Casetext unveils AI legal assistant powered by OpenAI’s “most advanced” large language model [Casetext, OpenAI의 “가장 진보된” 대규모 언어 모델로 구동되는 AI 법률 비서 공개], 2023.03.01., Legal Insider, <https://legaltechnology.com/2023/03/01/casetext-unveils-ai-legal-assistant-powered-by-openais-most-advanced-large-language-model/>, (최종방문일 23.04.01).
- OpenAI 제공 모델, <https://platform.openai.com/docs/models/overview>.
- PwC announces strategic alliance with Harvey, positioning PwC’s Legal Business Solutions at the forefront of legal generative AI [PwC, Harvey와 전략적 제휴 발표, PwC의 법률 비즈니스 솔루션을 법적 생성 AI의 최전선에 배치], 2023.3.15., PwC, <https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-announces-strategic-alliance-with-harvey-positioning-pwcs-legal-business-solutions-at-the-forefront-of-legal-generative-ai.html>, (최종방문일 23.04.01).

<ABSTRACT>

Potential Applications and Implications of GPT-4 in Legal Inference Using Korean Legal Aptitude Test (LEET)

Sungmi Park · Jeewon Park · Jungmihn Ahn

In November 2022, OpenAI released ChatGPT, a conversational AI chatbot, which caused a global sensation. The most recent model, GPT-4 is expected to replace experts in various legal tasks, including interpreting legal documents and drafting contracts, due to its impressive performance enhancements, such as achieving top 12% on LSAT.

This study aims to objectively prove the concerns surrounding GPT-4 through experiments. To achieve this, the study confirmed the results of the LSAT experiment announced by OpenAI and evaluated the potential impact of GPT-4 on the legal field by analyzing the problem-solving of LEET, a legal aptitude test similar to LSAT, in Korea. The study conducted various prompt design experiments to ensure that GPT-4's problem-solving is logical, valid, and able to draw accurate inferences. The study analyzed the problems that arose during GPT-4's inference process through this experimental method. The research results showed that GPT-4 can only perform superficial analysis when solving LEET test questions, has difficulty applying new facts, returns inconsistent answers to the same question, and often ignores implicit facts. Therefore, it is currently difficult to recognize GPT-4 as a rational inference model. However, it has potential to aid human inference in law by pre-training with legal data, incorporating expert feedback, and developing specialized legal technology.

Keywords : GPT-4, 법적추론(Legal Reasoning), 법학적성시험(LEET), 프롬프트 엔지니어링(Prompt Engineering), 추론오류(Inference Error)