GPT의 법적 통찰력 평가:

법률 분야에서의 GPT의 역량과 한계 탐구*

- LSAT를 중점으로 -

박지원[O1] 이진헌[1] 최영철[†2]
[1]한림대학교 정보법과학과
[2]한림대학교 글로벌학부, 교신저자
kiddotwix@gmail.com, dlwlsgjs321@gmail.com, cychoeacedu@gmail.com

Assessing the Legal Acumen of GPT:

Exploring its Capabilities and Limitations in the Field of Law

- Focused on LSAT -

JeeWon Park[O1] JinHeon Lee[1] YoungCheol Choe[†2]
[1]Legal Informatics and Forensic Science, Hallym University
[2]School of Global Studies, Hallym University

요 약

OpenAI에서 발표한 GPT-4 모델은 미국의 법학대학원 입학시험인 LSAT에서 전체 상위 12%라는 높은 성적을 기록하였다. 하지만 이를 공개한 OpenAI 측에서는 별다른 문제풀이 과정을 공개하지 않아 본 연구에서는 각각 GPT-3.5와 GPT-4를 활용하여 2007년도에 출제된 LSAT ‘The Official PrepTest June Sample’을 풀이한다. 이를 통해 LSAT 시험에 대한 분석 추론 및 독해력을 사람(human)이 직접 평가하여 GPT의 법률영역 적합성 능력을 증명하고자 한다. 본 연구는 동일 문제와 프롬프트를 기준으로 GPT 모델에 따른 성능 차이를 비교하고, GPT가 법률 분야에서 활용될 수 있는 방안을 탐구하는 것을 목적으로 한다.

1. 서 론

인공지능(Artificial Intelligence, AI) 기술이 발전함에 따라 자연어처리(Natural Language Processing, NLP) 분야에서 단순 언어모델(Language Model, LM)을 넘어선 초거대 언어모델(Large Language Model)의 기술이 나날이 발전하고있다. 생성형 AI인 GPT(Generative Pre-trained Transformer)는 LLM 중에서도 사람과 분간할 수 없을 정도의 성능을 지니고 있으며, 다양한 태스크를 수행하고 있다. 나아가 OpenAI에서는 지난해 GPT-3.5를 이어 2023년 3월, OpenAI에서 초거대언어모델인 GPT-4를 공개했다. GPT-4가 발표된 이후 OpenAI 측에서 본래 사람을 대상으로 설계된 다양한 분야의 전문적인 시험 풀이를 통해 GPT-3.5보다 GPT-4는 모든 시험 영역에서 높은 성적을 거두어 그 성능을 입증했다. 비록 OpenAI 측에서 GPT-4가 미국의 법학대학원 입학시험인 LSAT에서 (전체 상위 12%)을 차지한 것으로 확인할 수 있었으나[1], 해당 결과의 성능을 입증하기 위한 GPT-4의 문제풀이 과정 및 추론 과정에 대한 언급이 존재하지 않았다. 따라서, 본 연구를 통해 동일 시험 문제를 기준으로 GPT-3.5와 GPT-4가 풀이과정을 비교하여 풀이방식에 따른 추론과정 내지 추론능력을 검증한 뒤 평가하고자 한다. 이를 통해 GPT-4의 논리적 사고 능력과 법률 분야에서의 적용 가능성에 대한 평가를 진행하며, 그 한계와 오류에 대한 분석을 수행한다. 이러한 연구결과는 GPT-4의 법률 도메인에서의 미래적 활용 가능성을 평가하는 데에 기여할 것으로 기대된다.

* 이 논문은 2021년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: AI 기반 범죄수사 지원 / 연구개발과제번호: PR10-02-000-21]

2. 관련 연구

본 연구에서는 각각 GPT-3.5 모델과 GPT-4 모델을 활용하여 미국 법학대학원 입학시험인 LSAT(Law School Admission Test) 중 2007년도에 출제된 ‘The Official LSAT PrepTest June Sample 2007’을 풀이한다. 미국 법학대학원 입학시험 LSAT은 로스쿨에서 필요한 핵심 기술인 비판적 사고, 분석적, 논리적 및 설득적 추론을 요하는 시험으로, 법학적 사고 능력의 잠재력을 평가하는 시험이다[2]. 이는 논리적 추론, 분석적 추론, 그리고 독해력 시험을 다음과 같이 총 4가지 섹션으로 구성되어 있다: ‘SECTION I : Analytical Reasoning’, ‘SECTION II : Logical Reasoning’, ‘SECTION III: Logical Reasoning’, ‘SECTION IV: Reading Comprehension’. 본 연구에서는 4가지 섹션 중 상대적으로 어려운 섹션인 ‘SECTION I : Analytical Reasoning’과 2007

년도 기점으로 역대 LSAT 시험 중 처음으로 등장한 'SECTION IV : Reading Comprehension'을 중점적으로 해석 및 분석하고자 한다. 분석적 사고 및 추론(Analytical Reasoning, AR) 섹션은 일련의 사실과 규칙을 고려하여 지문 내 사실과 규칙이 주어졌을 때 참과 거짓을 판별하여 이를 결정하는 능력을 평가하도록 고안된 시험으로, 법학도가 잠재적으로 갖추어야 할 법률문제 해결 능력을 파악한다. 나아가 독해력(Reading Comprehension) 섹션은 2007년도에 등장한 이후 현재 시점까지 중요시되고 있는 시험 영역으로 평가되어, 이를 기준으로 GPT를 사용하여 풀이함으로써 로스쿨에서 일반적으로 접하는 길고 복잡한 문장과 유사한 예시를 제시하여 이해와 통찰력을 요하는 독해 능력을 측정하는 것을 목적으로 한다.

GPT는 다수의 연구에서 없는 사실을 사실인 것처럼 거짓된 정보를 제공하는 ' 환각(hallucination) 현상 '이 발생하는 사례가 등장했다[3]. 이와 같은 문제로 GPT의 신뢰성이 논란이 되며, GPT의 신뢰성 제고를 위해 GPT에게 '프롬프트(prompt) 디자인'을 설정한다. 프롬프트 디자인이란, 인공지능 모델에게 특정 태스크 수행 절차를 세세하게 제공하여 모델의 역량을 최대로 발휘할 수 있도록 결과를 도출하는 방법이다[4]. 따라서, 이러한 방식으로 GPT 모델에게 LSAT 풀이 절차를 주어 프롬프트 설정의 유무에 따른 GPT의 LSAT 풀이 과정을 분석함으로써 논리적 및 법적 추론 능력을 평가하여 이를 검증하고자 한다.

3. 실험방법

GPT-3.5와 GPT-4를 활용하여 LSAT 시험 문제마다 새로운 세션을 통해 문제 풀이를 진행했으며, 이에 따라 총 두 가지 프롬프트를 사용한 실험방법은 다음과 같다. (1)제로샷 프롬프트(Zero-shot prompt): 지문과 문제 외 별다른 예시 없이 문제풀이 절차만을 제시하여 문제를 풀게 한 방식과, (2)Zero-shot prompt에 사고 흐름 순서를 제공하는 방안인 사슬 사고 프롬프트(Chain-of-Thought prompt, CoT)[5]를 결합한 Zero-shot+CoT 프롬프트 설정을 통한 문제풀이 방식이다. 이는 Zero-shot과 이전 연구에서 발표된 'Let' s think step by step' [6] CoT 프롬프트를 접목한 것으로, GPT가 문제풀이 시 순서에 따르도록 구체적인 추론 과정을 요하는 프롬프트이다. 아래 <표1>은 (1)Zero-shot prompt를, <표2>는 (2)Zero-shot+CoT prompt의 예시이다.

표 1 Zero-shot prompt 예시

Zero-shot prompt: 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 : A company employee generates ...
Question : 1. If the last digit of an acceptable product code is 1, it must be true that the ...
GPT Answer :

표 2 Zero-shot + CoT prompt 예시

Zero-shot prompt 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 : A company employee generates ...
Question : 1. If the last digit of an acceptable product code is 1, it must be true that the ...
CoT Prompt Work through your logic **step-by-step** before answering. Solution:
GPT Answer :

4. 결과

본 연구의 실험을 진행해본 결과, GPT-4 모델의 성능이 섹션 및 프롬프트와 무관하게 GPT-3.5 모델보다 우수한 것으로 나타났다. 아래 <표3>을 살펴보면, SECTION I의 전체 문항 수는 23개, SECTION IV의 전체 문항 수는 27개로 구성되어 있으며, 섹션 및 프롬프트 별 GPT 모델에 따른 정답 개수를 비교하였다.

표 3 프롬프트 및 섹션에 따른 정답 개수

프롬프트	GPT 모델	SECTION I	SECTION IV
Zero-shot	GPT-3.5	5/23	19/27
	GPT-4	8/23	**23/27**
Zero-shot + CoT	GPT-3.5	4/23 (-1)	16/27 (-3)
	GPT-4	4/23 (-1)	**26/27 (+4)**

GPT-3.5와 GPT-4 모두 분석적 추론 과목인 SECTION I에 비해 독해력 과목인 SECTION IV의 정답 개수가 현저히 높은 것을 확인할 수 있다.

표 4 SECTION I 오답 유형 정리표

프롬프트	GPT 모델	SECTION I	오류 개수/전체 오류
Zero-shot	GPT-3.5	제시 프롬프트 조건 미충족	11/18
		조건적용 실패	7/18
Zero-shot	GPT-3.5	제시 프롬프트	10/19

+ CoT		조건 미충족	
		조건적용 실패	9/19
Zero-shot	GPT-4	제시 프롬프트 조건 미충족	10/16
		조건적용 실패	6/16
Zero-shot + CoT	GPT-4	제시 프롬프트 조건 미충족	11/19
		조건적용 실패	8/19

표 5 SECTION IV 오답 유형 정리표

프롬프트	GPT 모델	SECTION IV	오류 개수/전체 오류
Zero-shot	GPT-3.5	제시 프롬프트 조건 미충족	4/8
		조건적용 실패	4/8
Zero-shot + CoT	GPT-3.5	제시 프롬프트 조건 미충족	6/11
		조건적용 실패	5/11
Zero-shot	GPT-4	제시 프롬프트 조건 미충족	3/4
		조건적용 실패	1/4
Zero-shot + CoT	GPT-4	제시 프롬프트 조건 미충족	1/1
		조건적용 실패	**0/1**

<표4>와 <표5>는 SECTION I 및 SECTION IV에서 나타난 오답 유형 및 오답 개수를 나타낸다. 오답 유형은 (1)제시 프롬프트 조건 미충족, (2)조건적용 실패 오답으로, 총 2가지 오류가 발생하였다. 조건적용 실패 오류의 경우 지문 및 문제에서 두 가지 이상 조건이 주어졌을 때, 모든 조건을 올바르게 적용 시키지 못하여 문제풀이에 실패한 오류이다. 연쇄적인 조건이 순차적으로 연결짓지 못하고 첫 번째 조건만 풀이한 채 다음 조건에 적용시키지 못하는 등, 풀이 과정에 모든 조건들을 갖추지 못한 오류로 판단된다.

표 6 제시 프롬프트 조건 미충족

제시 프롬프트 조건 충족 예시	제시 프롬프트 조건 미충족 예시
(1)Convert the conditions to a simple logical representation: C1: ¬J4 (Jamaica will not be its destination in week 4) (생략) (2)Apply the options to the conditions: (A) Guadeloupe, Jamaica, Martinique, Trinidad, Guadeloupe, Martinique, Trinidad (생략)	(1)Let's represent the conditions in a logical form: Jamaica ≠ week 4(생략) (2)Now, let's apply the options to the conditions: Option (A): Guadeloupe, Jamaica, Martinique, Trinidad, Guadeloupe, Martinique, Trinidad **(3)**
(3)Generate counter-facts: (A) Fails C4: G is not scheduled before J (Week 1) (생략) (4)Verify whether an option meets ALL the conditions from (1):	(4)Thus, only option (A) satisfies all the given conditions. Therefore, the answer is (A) Guadeloupe, Jamaica, Martinique, Trinidad, Guadeloupe, Martinique, Trinidad.

본 실험에서 가장 흔히 발생한 오류는 '제시 프롬프트 조건 미충족'으로, <표1>과 <표2>와 같이 GPT 모델에게 Zero-shot 프롬프트에서 문제풀이 절차를 제공하였으나, 프롬프트를 따르지 않고 문제풀이를 진행한 오류이다. 절차 중 '*(3) Third, generate counter-facts*'을 제외하고 문제풀이를 진행한 것을 <표6>에서 확인할 수 있다.

5. 결론

본 연구는 두 가지 프롬프트를 사용하여 GPT-3.5와 GPT-4 모델에 따른 LSAT 문제풀이 추론 과정을 평가하였다. Zero-shot+CoT 프롬프트를 적용한 경우, 모든 섹션에서 GPT-3.5의 성능이 오히려 떨어진 것을 <표3>을 통해 확인할 수 있다. GPT의 문제풀이 과정을 분석해본 결과, 프롬프트 디자인은 GPT를 활용하는 데 큰 영향을 끼치며, 이외에 발견된 제시 프롬프트 조건을 미충족하는 GPT의 잠재적 오류를 <표6>를 통해 확인할 수 있다.

본 연구는 이와 같은 오류를 통해 GPT의 추론능력은 아직 개선의 여지가 있다고 시사한다. 따라서, 향후 연구는 모델이 한계를 극복할 수 있도록 보다 효과적인 프롬프트를 개발하는 데 초점을 맞춰 GPT 모델을 법률영역에서 활용하고자 한다[7].

참 고 문 헌

[1] 조성미, 사람 같다는 챗GPT 뜨는데…우리 AI 기술 현주소는?, 매일경제, 2023.01.31.
[2] LSAT 공식 홈페이지 참조:https://www.lsac.org/about?gad=1&gclid=CjwKCAjwscGjBhAXEiwAswQqNAsov552-fPkDqztmewZ86I1-m_Ve1seNtMwIvgZAKU8fpH3V8QrBhoCuxUQAvD_BwE&gclsrc=aw.ds.
[3] OpenAI, GPT-4 Technical Report, 2023.
[4] Liu, Vivian, and Lydia B. Chilton., Design guidelines for prompt engineering text-to-image generative models.., CHI Conference on Human Factors in Computing Systems, 2022.
[5]Wei, Jason, et al., Chain of thought prompting elicits reasoning in large language models, 2022.
[6] Kojima, Takeshi, et al., Large language models are zero-shot reasoners, 2022.
[7] 임영익, 임영익 인텔리콘 대표변호사, GPT-4는 변호사를 대체 하는가? 인공지능신문, 2023.03.26., https://www.aitimes.kr/news/articleView.html?idxno=27645.