# 생성형 대규모 언어 모델을 활용한 법률 분야 정관 검토 기계 독해 방법론*

정해인○ 김민주 연희연 임영선 구명완†
서강대학교 인공지능학과
haeindain@sogang.ac.kr, mjmjkk0307@sogang.ac.kr, yeen214@sogang.ac.kr, firefpswl@gmail.com,
mwkoo@sogang.ac.kr

# Machine Comprehension Methodology for Legal Document Review Using a Generative Large Language Model

Haeln Jung○ MinJu Kim HeuiYeen Yeen YoungSun Lim Myoung-Wan Koo
Department of Artificial Intelligence, Sogang University

**요 약**

생성형 대규모 언어 모델을 활용한 한국 법률 분야의 정관 검토 기계 독해 방법론을 제시한다. 정관 텍스트를 읽고 주어진 질문에 대한 적절한 답변을 할 수 있도록 다양한 프롬프팅 방법론을 실험한다. 이에 대한 대규모 언어 모델 성능 측정을 통해, 본 연구는 한국 법률 도메인에서의 GPT-3.5와 GPT-4의 성능을 최초로 제시한다. 문제 유형에 따른 성능 차이와 프롬프팅 방법의 장단점을 분석하여 답이 정형화되어 있는 문제 유형의 경우에는 단순한 프롬프팅 방식만으로 성능이 향상되지만, 그렇지 않은 경우에는 한계가 존재하는 것을 확인하여 추후 해당 분야에서의 고도화된 프롬프팅 방식 연구에 응용될 수 있기를 기대한다.

## 1. 서 론

대규모 언어 모델은 자연어 이해, 생성 등 다양한 자연어 처리 작업에서 뛰어난 성능을 보이며 그 활용성이 강조되고 있다. 이러한 대규모 언어 모델을 효과적으로 활용할 수 있는 프롬프트 엔지니어링에 대한 연구도 함께 활발히 진행되고 있다.

대규모 언어 모델 등장 이전에는 방대한 데이터를 기반으로 한 사전 학습 모델 파인튜닝으로 기계 독해 문제를 해결하는 연구가 주류를 이루었다. 그러나 이는 추가적인 모델 학습과 방대한 양의 데이터가 필요하다는 점에서 한계를 가진다.[1] 반면, 대규모 언어 모델을 프롬프트 엔지니어링으로 잘 활용하는 경우, 방대한 양의 학습 데이터 없이 일반화된 성능을 달성할 수 있다는 장점을 가진다.[2]

따라서 본 논문에서는 별도의 파인튜닝 없이 대규모 언어 모델을 활용한 프롬프트 엔지니어링을 통해 전문 지식을 기반으로 하는 답변이 필요한 법률 분야의 정관 검토 기계 독해 문제를 해결하는 방안을 제시한다. 정관이란, 법인의 목적, 조직, 업무 집행 따위에 관한 근본 규칙 혹은 그것을 적은 문서로, 변호사는 상법에 기반하여 회사 정관에 위법한 서술이나 누락된 요소가 없는지 검토하게 된다. 그러나 변호사의 정관 검토 과정은 통상적이고 반복적인 일이기 때문에, 이를 대규모 언어 모델을 통한 기계 독해 문제로 해결하여 변호사의 업무 효율성을 높이고자 한다.

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

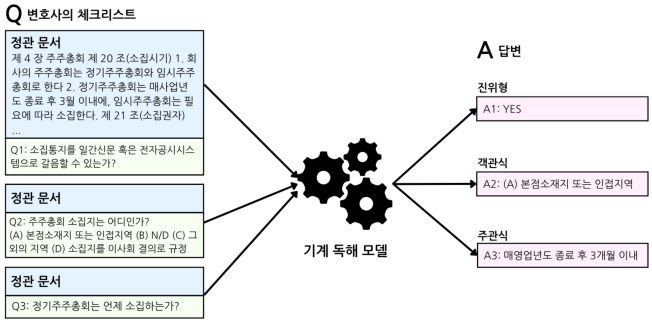


**그림 1** 정관 검토 기계 독해 문제 프로세스

대규모 언어 모델을 활용한 정관 검토 기계 독해 문제에 대한 프로세스는 그림 1과 같다. 특정 정관 문서가 주어지면 이에 해당하는 변호사가 검토해야 할 체크리스트 질문을 추출한 후, 대규모 언어 모델을 통해 질문에 대한 답변을 출력하는 문제이다. 정관 검토 기계 독해 문제를 해결하기 위해서는 법률 지식에 기반한 전문 용어를 이해하고, 정관 문서 안에서 질문에 대한 답을 추론해내는 능력이 필요하다. 그러나 현재 대규모 언어 모델은 뛰어난 한국어 법조문과 판례에 대해 충분히 학습되지 않아 한국의 법률 분야에 대한 기계 독해 성능은 아직 이에 미치지 못하기 때문에, 해당 문제를 극복하기 위해서는 전문 지식을 효과적으로 반영하여 답변을 생성하도록 하는 프롬프트 엔지니어링이 필요하다. 따라서 이를 위하여 법률 전문가 팀과 협업하여 정관 검토 기계 독해 문제를 위한 데이터셋을 구축하고,

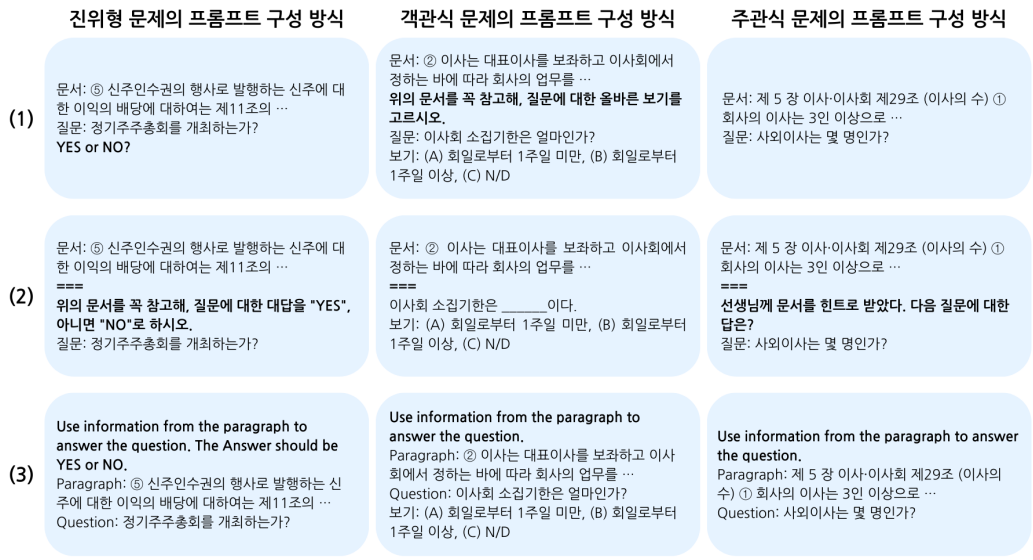


**그림 2** 정관 검토 기계 독해 문제 유형 별 프롬프트 구성 방식

문제 유형에 따라 다양한 프롬프트 구성 방식을 실험한다. 이에 대한 GPT-3.5와 GPT-4와 같은 대규모 언어 모델의 성능을 측정하여, 문제 유형에 따른 성능 차이와 프롬프팅 방법의 장단점을 분석한다. 추가적으로 파인튜닝 모델과의 성능 비교를 통해 대규모 언어 모델을 활용하는 프롬프팅 방법론이 더욱 효과적임을 확인하였다.
이로써, 전문 분야의 기계 독해 문제에 대해 대규모 언어 모델 프롬프트 엔지니어링 방법론이 효과적임을 보이며, 특히 한국 법률 도메인에서의 GPT-3.5와 GPT-4 성능을 최초로 비교하며 추후 해당 분야에서의 고도화된 프롬프팅 방식과 관련된 연구에 도움이 될 수 있기를 기대한다.

## 2. 정관 검토 기계 독해 문제에 대한 프롬프팅

### 2.1 정관 검토 기계 독해 문제 정의

변호사 및 법학 전문 대학원 교수와 학생들로 이루어진 법률 전문가 팀과 협업하여 정관 검토 기계 독해 문제에 대해 정의하고 프로세스를 구축하였다. 실제 상장 기업의 정관 문서가 주어지면 이에 대한 변호사의 체크리스트를 추출한 후, 정관 문서와 질문을 입력 시 모델로부터 답변이 출력되는 문제이다. 예를 들어, '주식과 주권'에 대한 정관 내용이 주어진 경우, 모델은 "회사의 주식은 어떤 방법을 발행할 수 있는가"와 같은 질문에 대해 정확한 답변을 정관 문서 안에서 찾아내게 된다.
정관 검토 기계 독해 문제의 유형은 3가지로, 진위형, 객관식, 주관식으로 구성되어 있다. 먼저, 진위형의 경우 질문에 대한 답변이 'Yes', 'No', 'N/D'로 구성되고, 이때 'N/D'는 Not Determined의 약자로, 답변을 주어진 정관 문서에서 찾을 수 없는 경우를 의미한다. 그리고 객관식은 답변이 될 수 있는 여러 개의 보기가 주어지는 질문 유형이다. 마지막으로, 주관식은 답변이 서술 형태인 질문 유형이다.

### 2.2 정관 검토 기계 독해 프롬프팅 방법론 연구

프롬프트는 문제, 질문, 설명으로 구성을 하여, 주어진 문제에 대해 대규모 언어 모델이 가장 알맞는 답변을 생성할 수 있는 설명을 추가해주는 것에 초점을 맞추었다. 이때, 설명은 다양한 자연어처리 분야에 활용될 수 있는 영어 프롬프트 데이터셋인 P3(Public Pool of Prompts)[3]를 참고하였다. P3의 Multiple Choice QA와 Natural Language Inference 데이터셋의 프롬프트 형태를 집중적으로 탐구하여 한국어 정관 도메인에 대해 좋은 성능을 보이는 프롬프트 형태를 각 문제 유형 별로 3가지 선정하여 실험하였다. 다양한 프롬프트 방식을 실험해보고자 설명을 한글과 영어로 나누어 구성해보았고, 구분자 "==="를 추가해주어 모델로 하여금 문서와 질문을 구분하여 인식할 수 있도록 돕고자 하였다. 특별히 보기가 주어진 객관식 문제 유형의 경우 질문에 빈칸을 삽입하여 모델이 빈칸을 채울 수 있도록 하는 프롬프팅 방식도 실험하고자 하였다. 문제 유형 별 구체적인 프롬프트 구성 방식은 그림 2과 같다.

## 3. 실험

### 3.1 데이터 셋 구축

정관 텍스트는 전자공시 시스템을 통해 상장 기업 2000여 개의 정관 문서 데이터를 크롤링 하여 수집되었다. 회사 정관 문서 데이터셋이 구축된 이후, 변호사팀이 '주식과 주권', '주주총회', '이사, 이사회, 감사 위원회'분야에 관해, 정관을 검토하며 꼭 확인해야 하는 체크리스트 질문-답변 쌍을 제작하였다. 이후 크라우드 소싱을 통해 각 정관 별 질문에 대한 답변을 어노테이션 하는 과정을 거쳐 최종적인 정관 검토 기계 독해 데이터셋이 완성되었다.
정관 데이터셋은 기계 독해 알고리즘 적용에 용이하도록 전처리 되었다. 최종적인 스키마는 '정관 텍스트', '질문', '답변', '답변의 근거', '답변의 근거 조항', '변호사 설명' 항목으로 구성되었다. 다만, 본 논문에서는 질문에 대한 답변만 사용을 하였고, 이후 연구에서 답변의 근거와 변호사 설명 항목을 활용하여 설명 가능한 정관 검토 시스템을 구축할 예정이다.
그리고 답변 유형에 따라 크게 세 종류(진위형, 객관식, 주관식)의 문제로 데이터를 분류하였다. 진위형 질문은 9163개, 객관식 질문은 12161개, 주관식 질문은 9576개이다.

### 3.2 사용 모델

Open AI의 GPT-3.5-turbo† 모델과 GPT-4‡ 모델에 문제

와 설명으로 구성된 프롬프트를 넣고 답변을 생성하게 하였다. GPT-3.5-turbo와 GPT-4 모델은 트랜스포머 디코더 기반의 GPT 모델로 입력 가능한 토큰 수는 각각 4096개, 8192개이다.

### 3.3 평가 방식

평가 지표로 정확도를 사용하여, 단답형과 객관식의 경우 예측 결과가 동일한 데이터 건수를 계산하여 전체 예측 데이터 건수로 나누어 주었다. 주관식의 경우 같은 내용이라도 다르게 설명할 수 있기 때문에 토큰을 기준으로 일치하는지 측정하였다.

## 4. 결과

각 정관 문제 유형에 대한 프롬프트 구성 방식 차이에 따른 GPT-3.5, GPT-4 모델 실험 결과는 표 1, 2, 3과 같다. 먼저 진위형과 객관식 유형의 결과를 살펴보면, 영어 설명을 프롬프트에 추가해주었을 경우에서 두 모델 모두 가장 좋은 성능을 보이는 것을 알 수 있다. 그리고 영어 설명을 사용했을 경우 두 모델의 성능은 동일하지만, 한국어 설명을 사용했을 경우에는 GPT-4 모델의 성능이 GPT-3.5 모델보다 더 높은 것을 확인할 수 있다. 반면 앞서 두 문제 유형과 달리 주관식 유형의 경우, 영어와 한국어 설명을 사용하였을 때의 성능 차이가 뚜렷히 보이지 않는 것이 발견되었다. 뿐만 아니라 GPT-3.5 모델과 GPT-4 모델의 성능 차이도 보이지 않는 것을 알 수 있다.
최종적으로 문제 유형 별 평가 점수를 평균 낸 결과는 표 4와 같다. P3 데이터셋의 프롬프트에 포함된 기본적인 설명 형식과 동일한 형태로 프롬프트를 구성한 결과 진위형, 객관식 유형에서 GPT-4 모델이 GPT-3.5 모델보다 더 향상된 성능을 보인 것을 알 수 있다. 반면에, 주관식 유형에서는 GPT-3.5와 GPT-4 모델의 성능 차이가 유사한 것을 볼 수 있다. 이는 진위형, 객관식 유형과 같이 답의 형태가 정형화되어 있고 모델이 보기를 보고 추론할 수 있는 경우에는 기본적인 설명만 프롬프트에 추가하여도 성능이 어느 정도 좋다는 것으로 분석된다. 그러나 주관식 유형과 같이 답의 형태가 정해져 있지 않고 모델로 하여금 논리적인 추론을 요하는 경우는 간단한 설명만으로 모델이 정확한 답변을 생성해내도록 하는 것에 한계가 있다고 분석할 수 있다. 그러므로 주관식 유형의 문제에 대해서는 보다 더 고도화된 프롬프트 엔지니어링 방식을 사용한 실험이 필요할 것으로 생각된다.

**표 1** 진위형 유형에 대한 프롬프트 구성 방식 별 GPT-3.5와 GPT-4 성능 비교

| 프롬프트 구성 방식 | GPT-3.5 | GPT-4 |
|---|---|---|
| 진위형 프롬프트 (1) | 0.894 | 0.925 |
| 진위형 프롬프트 (2) | 0.914 | 0.944 |
| 진위형 프롬프트 (3) | 0.956 | 0.956 |

**표 2** 객관식 유형에 대한 프롬프트 구성 방식 별 GPT-3.5와 GPT-4 성능 비교

| 프롬프트 구성 방식 | GPT-3.5 | GPT-4 |
|---|---|---|
| 객관식 프롬프트 (1) | 0.519 | 0.696 |
| 객관식 프롬프트 (2) | 0.546 | 0.654 |
| 객관식 프롬프트 (3) | 0.758 | 0.758 |

**표 3** 주관식 유형에 대한 프롬프트 구성 방식 별 GPT-3.5와 GPT-4 성능 비교

| 프롬프트 구성 방식 | GPT-3.5 | GPT-4 |
|---|---|---|
| 주관식 프롬프트 (1) | 0.689 | 0.695 |
| 주관식 프롬프트 (2) | 0.702 | 0.682 |
| 주관식 프롬프트 (3) | 0.691 | 0.691 |

**표 4** 문제 유형 별 GPT-3.5와 GPT-4 성능 비교

| 문제 유형 | GPT-3.5 | GPT-4 |
|---|---|---|
| 진위형 | 0.921 | 0.942 |
| 객관식 | 0.608 | 0.703 |
| 주관식 | 0.694 | 0.689 |

추가적으로 작은 모델의 파인 튜닝 성능과 대규모 언어 모델의 성능을 비교하였다. 표 5는 진위형 36652개, 객관식 48640개, 주관식 38301개의 문제를 한국어 T5-base 모델이 각 문제 유형 별로 학습하고 앞서 대규모 언어 모델 실험 시 사용했던 평가 데이터로 성능을 계산한 결과를 나타낸다. 진위형, 객관식 형식의 경우 T5-base 모델의 성능이 더 높지만, 주관식 형식의 경우 대규모 언어 모델의 성능이 크게 향상된 것을 확인할 수 있다. 이는 주관식 형식과 같이 데이터나 답의 형태가 정형화되어 있지 않고 다양한 경우 파인튜닝보다 프롬프트 엔지니어링 방법이 더욱 효과적임으로 분석된다.

**표 5** 파인튜닝 모델과 대규모 언어모델 성능 비교

| 문제 유형 | T5-base Accuracy | GPT-4 Accuracy |
|---|---|---|
| 진위형 | 0.991 | 0.942 |
| 객관식 | 0.775 | 0.703 |
| 주관식 | 0.366 | 0.689 |

† https://platform.openai.com/docs/models/overview
‡ https://platform.openai.com/docs/models/overview

## 5. 결론 및 향후 연구

본 논문에서는 정관 프롬프팅 방법론을 활용한 대규모 언어 모델의 한국 법률 도메인 문제 풀이 성능을 확인하였다. 실험 결과 정관 프롬프팅 방법론에 대하여 GPT-4 모델이 진위형, 객관식 문제 유형에 대하여 GPT-3.5 모델보다 강점을 보였으나 주관식 문제 유형에 대하여는 성능 향상을 보이지 않았다. 이와 같이 진위형, 객관식 유형과 같이 모델이 답변의 보기를 참고할 수 있는 경우에는 짧은 설명만을 추가한 기본적인 프롬프팅 방법론으로 충분히 성능 향상을 이끌어 낼 수 있지만, 주관식 유형과 같이 답이 정형화되어 있지 않은 경우에는 단순한 설명만으로 모델이 논리적인 추론을 하도록 만드는 것에는 한계가 존재한다. 따라서 향후에는 주관식 유형에 대하여 수학 문제나 상식 추론 문제에서 대규모 언어모델이 높은 성능을 낼 수 있도록 유도하는 Chain of thought 프롬프트 방식이나 문제를 단계별로 나누어 해결하도록 하는 프롬프팅 방식과 같이 더 고도화된 프롬프트 엔지니어링 방법론에 대한 실험과 연구가 필요한 것으로 보인다.

## 참 고 문 헌


[1] LESTER, Brian; AL-RFOU, Rami; CONSTANT, Noah. The power of scale for parameter-efficient prompt tuning.*arXiv preprint arXiv:2104.08691*, 2021.

[2] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.*ACM Computing Surveys*,*55*(9), 1-35.

[3] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization.*arXiv preprint arXiv:2110.08207*.