규제연구 제34권 제1호 2025년 6월 http://dx.doi.org/10.22954/ksrs.2025.34.1.002

법령 정보 확인 및 규제 판별에서의 생성형 AI 활용 가능성: ChatGPT 기반 탐색적 분석*

박 정 원**

본 연구는 생성형 인공지능(Generative AI), 특히 ChatGPT를 활용하여 법령 정보 확인 및 규제 여부 분류에 대한 적용 가능성을 탐색적으로 분석하고자 하였다. 이를 위해 법령명, 조문 번호, 조문 제목 등 조문 정보를 입력값으로 하여 해당 내용을 생성하도록 한 과제(프롬프트①)와, 조문 원문을 바탕으로 규제 여부를 분류하도록 한 과제(프롬프트②)를 구성하였다. 각 과제에 대해 분류 정확도, 내용 유사도, 실행 일관성, 비교 변수에 따른 성능 차이 등을 중심으로 평가를 수행하였다. 분석 결과, 조문 생성 과제에서는 정보 누락, 허위 생성, 실행 간 불일치 등의 문제가 확인되었으며, 기준 조문과의 유사도도 전반적으로 낮은 수준을 보였다. 반면, 규제 여부 분류 과제에서는 평균 정확도 66.9%, 높은 실행 일관성과 신뢰도를 보여, 실무적 보조 도구로서의 가능성을 일부 확인할 수 있었다. 다만 비규제 조문에 대한 오분류 경향은 주요한 한계로 나타났다. 본 연구는 생성형 AI의 법령·규제 분석 기능을 초기적으로 평가한 시도로서, 향후 도메인 특화 AI와의 성능 비교 및 정책적 활용 가능성에 대한 실증적 논의를 위한 기초 자료를 제공한다는

* 이 논문은 2021년 대한민국 교육부와 한국연구재단의 일반공동연구지원사업의 지원을 받아 수행된 연구임 (NRF-2021S1A5A2A03064391)

** 국립경국대학교 사회과학부 행정학전공 부교수, 경상북도 안동시 경동로(jwpark@gknu.ac.kr)

논문의 질적 향상을 위해서 건설적인 논평을 해주신 익명의 심사위원께 진심으로 감사드립니다.
접수일: 2025/3/26, 심사일: 2025/4/5, 게재확정일: 2025/5/7

점에서 의의가 있다.

핵심 용어 : 인공지능, ChatGPT, 법령 정보, 규제 판별

I. 서론

최근 인공지능(AI) 기술의 급격한 발전은 다양한 산업 및 분야에서 혁신적 변화를 촉진하고 있으며, 특히 생성형 AI는 자연어 처리(Natural Language Processing, NLP) 기술을 바탕으로 인간과 유사한 대화 및 텍스트 생성 능력을 구현하면서 주목받고 있다. Open AI의 ChatGPT, Google의 Gemini, xAI의 Grok 등 대표적인 생성형 AI 모델은 법률, 의료, 비즈니스 등 다수의 분야에서 정보 검색, 분석, 예측을 통한 의사결정 지원 도구로 활용될 가능성을 제시하고 있다. 이러한 흐름 속에서, 행정 및 산업 전반에서 규제 관련 법적 판단의 신속성과 정확성을 제고하기 위한 새로운 방법론으로서 생성형 AI의 적용 가능성에 대한 관심이 증가하고 있다.

본 연구는 규제 판별 과정에서 생성형 AI의 활용 가능성을 학술적으로 고찰하고자 한다. 규제 판별은 복잡한 법률 문서를 분석하고, 관련 법령 및 규제 조문을 식별한 후, 이를 바탕으로 규제 여부를 판단하는 일련의 절차를 포함한다. 이러한 과정은 행정 실무뿐만 아니라 피규제자의 입장에서도 중요한 의미를 가진다. 신제품 출시나 서비스 도입 시, 규제의 존재 여부, 준수 의무 및 요건, 집행 절차 등을 정확히 파악하는 것은 필수적이기 때문이다.

그러나 기존의 규제 판별 방식은 주로 소수 전문가의 해석과 판단에 의존해 왔으며, 법령 및 규제의 증가와 복잡성 심화로 인해 이러한 방식의 효율성과 판단 일관성에 한계가 제기되고 있다.[1] 특히, 법령·규제의 내용이 전문적이고 난해하여 비전문가인 국민(피

1) 한국의 법령(법률, 대통령령, 총리령·부령 포함) 수는 2000년 3,525개에서 2024년 5,144개로 24년 만에 46% 증가하였으며(법제처, n.d.-b), 미국에서도 모든 연방 부서 및 기관에서 발행한 규정을 성문화한 연방규정집(Code of Federal Regulations, CFR)의 페이지 수가 2000년부터 2023년까지 52,211페이지(38%)가 증가하였다

규제자)이 이를 직접 확인하고 정확하게 이해하기 어려운 실정이다. 규제자와 일부 법률 전문가들은 법령과 규제의 취지, 적용 대상, 집행 절차 등을 숙지하고 있지만, 피규제자는 이를 신속하게 파악하기 어렵기 때문에, 이는 규제자(정부 및 규제 기관)와 피규제자(기업 및 개인) 간의 정보 비대칭성을 더욱 심화시키는 요인으로 작용한다. 한편, 기업이나 개인 사업자는 규제를 정확히 이해하지 못할 경우, 불필요한 법적 리스크를 부담하거나 사업 추진이 지연되는 등 경제적 손실을 입을 가능성이 크다. 반면, 규제자의 입장에서도 정보 비대칭성이 해소되지 않으면 규제 준수를 촉진하기 어렵고, 사후적인 법적 분쟁이나 행정 부담이 증가할 수 있다.

이러한 배경을 고려할 때, AI를 활용한 규제 정보 확인 및 판별은 단순한 법령 해석을 넘어 정보 비대칭 문제를 완화하는 중요한 역할을 할 수 있다. AI 기반 규제 분석 시스템이 구축된다면, 방대한 법령 및 규제 정보를 자동으로 분류·요약하고, 피규제자가 이해하기 쉬운 형태로 제공할 수 있을 것이다. 이는 법률·행정 절차의 효율성을 높이는 동시에, 피규제자의 규제 준수 부담을 줄이고, 규제자와 피규제자 간의 신뢰를 증진하는 효과를 기대할 수 있다.

본 연구에서는 이러한 배경을 바탕으로, 생성형 AI가 규제 판별 과정에서 실질적으로 어떤 역할을 수행할 수 있는지를 분석하고자 한다. 기존 연구들은 주로 AI를 활용한 법률 문서 분석에서 자연어 처리(NLP) 기술의 유용성에 초점을 맞추어 왔으나(Janatian et al., 2023; Noguti, Gonçalves, & Rosa, 2020; 정채연, 2024), AI 모델이 규제 판별 과정에서 실제 법령 정보를 얼마나 정확하게 검색·제시하며, 이를 바탕으로 규제 여부를 효과적으로 판단할 수 있는지를 평가한 연구는 상대적으로 부족한 실정이다.

이에 본 연구는 다음과 같은 연구 질문을 중심으로 생성형 AI의 규제 판별 가능성을 탐색하고자 한다.

1. 생성형 AI 모델(ChatGPT)은 법령 및 규제 조문을 검색할 때, 실제 레퍼런스(공식 법령 문서)와 얼마나 유사한 정보를 제공하는가?
2. 생성형 AI 모델은 주어진 법령 조문을 바탕으로 규제 여부를 얼마나 정확하게 판단

(Scacchi, 2024). 2012년 기준으로만 보더라도, CFR을 모두 읽는 데 약 3년이 걸릴 것으로 추정될 정도로 규제 문서는 방대하고 복잡하다(Scacchi, 2024).

할 수 있는가?

이 연구 질문에 답하기 위해, 본 연구는 대표적인 생성형 AI 모델인 ChatGPT가 규제 판단 과정에서 제공하는 정보의 신뢰성과 활용 가능성을 평가하는 두 가지 주요 분석 과정을 설계하였다. 첫 번째 단계에서는 AI가 검색한 법령 및 조문 정보의 정확성과 규제 판단 결과가 실제와 얼마나 일치하는지를 평가한다. 이를 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-Score, AUROC 등의 성능 지표를 활용한다. 둘째, AI가 특정 법령 및 조문의 내용을 검색할 때, 제공된 정보가 실제 레퍼런스(공식 법령 문서)와 얼마나 유사한지를 검토한다. 이를 위해 ROUGE 및 Jaccard 계수와 같은 자연어 처리(NLP) 기반 유사성 평가 지표를 적용하여, AI가 제공하는 법령 조문이 원본 텍스트와 얼마나 일치하는지를 분석한다.

이러한 분석을 통해 생성형 AI가 법률·행정 분야에서 법령 정보 확인 및 규제 여부 분류 도구로서 실제로 활용 가능한지를 탐색적으로 검토하고, 향후 AI 기반 규제 판별 시스템의 발전 방향을 모색하고자 한다. 본 연구는 생성형 AI의 법률·규제 분석 기능을 초기적으로 평가한 연구로서, AI가 조문 이해와 규제 분류 과정에서 수행할 수 있는 역할과 한계를 실증적으로 살펴보는 데 초점을 둔다.

Ⅱ. 연구 배경

1. 생성형 AI와 법률서비스의 활용

생성형 AI(Generative AI)는 텍스트, 이미지, 음성 등 다양한 형태의 콘텐츠를 자동으로 만들어내는 인공지능을 의미하며, 여기서 '생성(generative)'이라는 개념은 AI가 사용자의 명령 없이도 요구에 맞춰 스스로 학습하고 결과물을 생성할 수 있는 범용 인공지능을 뜻한다(조영임, 2023). 기존의 AI가 주어진 데이터를 분석하고 특정 패턴을 인식하는 데 초점을 맞췄다면, 생성형 AI는 학습된 데이터를 기반으로 새로운 콘텐츠를 창작할 수 있다는 점에서 차별화된다.

생성형 AI는 머신러닝과 딥러닝 기술의 발전과 함께 본격적으로 등장하였으며, 대규모 언어 모델(Large Language Model, LLM), 생성적 적대 신경망(Generative Adversarial Network, GAN), 트랜스포머(Transformer), 모델 변이형 오토인코더(Variational Autoencoder, VAE) 등의 기술을 기반으로 한다.[2)]

초기 생성형 AI 모델은 2000년대 초반부터 발전해 왔으나, 대중의 관심을 본격적으로 받게 된 계기는 2022년 11월 OpenAI가 대규모 언어 모델(LLM)인 GPT-3.5 기반의 상용 대화형 AI, ChatGPT를 출시하면서였다. OpenAI가 개발한 GPT(Generative Pre-trained Transformer)는 트랜스포머(Transformer) 구조와 비지도 학습(unsupervised learning)을 결합한 모델로, 이는 주석이 달리지 않은 방대한 데이터를 활용해 머신러닝 모델을 훈련하는 방식이다(Heaven, 2023). 최초의 GPT 모델인 GPT-1은 2018년에 개발되었으며, 이후 지속적인 성능 개선을 거쳐 GPT-3.5에서 챗봇 형태로 출시되었다. 이어 2023년 3월 발표된 GPT-4에서는 대폭적인 성능 향상이 이루어졌으며, 특히 GPT-3.5 대비 모델 크기가 약 500배 증가했다(양지훈·윤상혁, 2023). 그리고 2023년 11월에는 GPT-4 Turbo가, 2024년 5월에는 GPT-4o가 출시되었으며, 2025년 2월에는 GPT-4.5가 공개되었다. 한편, OpenAI 외에도 다양한 빅테크 기업들이 LLM 기반 생성형 AI를 출시하고 있다. 대표적으로 구글(Google)의 Gemini, 메타(Meta)의 Llama, xAI의 Grok, 바이두(Baidu)의 Ernie, 알리바바(Alibaba)의 Tongyi Qianwen, DeepSeek의 DeepSeek-V3 등이 있다.

생성형 AI의 발전과 함께 다양한 산업 분야에서 이를 활용하려는 연구와 노력이 활발히 이루어지고 있으며, 활용 또한 점점 확대되고 있다. 생성형 AI는 정보통신(IT) 산업뿐만 아니라 제조업, 교육업, 의료업, 금융업 등 다양한 분야에 적용되고 있으며, 법률 분야 또한 예외가 아니다. 법률서비스 분야에서 AI는 사람과의 문답을 통해 판례 및 문헌을 검색하고, 번역과 요약을 수행하며, 법률 문서 초안을 작성하는 등 보조 및 지원의 역할을 이미 수행하고 있다(고정철, 2024).

---

2) 생성형 AI의 기반 기술에 대한 개괄적인 내용은 조영임(2023)에서 확인할 수 있다. 대규모 언어 모델(LLM)의 개념과 발전 과정은 Vaswani et al.(2017)에서 논의되며, 트랜스포머 모델에 대한 주요 연구도 이 논문에서 다루어진다. 생성적 적대 신경망(GAN)의 원리는 Goodfellow et al.(2014)에서 확인할 수 있으며, 변이형 오토인코더(VAE)에 대한 설명은 Kingma & Welling (2013)에서 찾아볼 수 있다.

법률서비스 분야에서 AI 활용이 확대됨에 따라, LLM 기반 생성형 AI의 성과를 평가하는 실증 연구도 최근 진행되고 있다(Choi, Monahan, & Schwarcz, 2024; Martin et al., 2024; Shui et al., 2023). Martin et al.(2024)은 LLM과 전통적인 법률 계약 검토자(주니어 변호사 및 법률 프로세스 아웃소싱 업체)를 정확성, 속도, 비용 효율성 측면에서 비교하였으며, 그 결과 고급 LLM이 법적 이슈를 판단하는 데 있어 인간과 동등하거나 더 높은 정확성을 보였다. 또한, 검토 속도는 인간보다 훨씬 빠르고, 비용 면에서는 전통적인 방법에 비해 99.97%의 절감 효과를 나타냈다.

Choi et al.(2024)는 AI 지원이 법률 분석의 질을 일관되게 향상하지는 않지만, 작업 속도를 크게 증가시키는 효과가 있음을 실험을 통해 확인했다. Choi et al.(2024)는 AI가 인간의 법률 분석에 미치는 영향을 평가하기 위해 무작위 통제 실험을 수행하였으며, 법과대학 학생들을 대상으로 GPT-4의 도움을 받거나 받지 않은 상태에서 현실적인 법률 과제를 수행하도록 무작위로 배정하였다. 이후, 각 과제를 수행하는 데 걸린 시간을 추적하고 블라인드 방식으로 결과를 평가하였다. 실험 결과, GPT-4를 활용한 참가자들은 법률 분석의 질이 소폭 개선되었으나 그 효과는 일관되지 않았으며, 작업 속도는 지속적으로 크게 향상되었다. 특히 법률 분석 능력이 낮은 참가자들이 가장 큰 개선 효과를 나타냈으며, AI 지원은 참가자들의 기본 작업 속도와 관계없이 일정한 시간 절약 효과를 제공했다. 이러한 연구 결과는 법률 업계에서 AI가 생산성과 평등성을 높이는 도구로 활용될 가능성을 시사한다.

다만, 국내 법률 서비스 분야에서 생성형 AI의 활용 성과를 분석한 연구는 아직 제한적이다. 따라서 본 연구는 국내 법률 환경에서 법령 정보 확인과 규제 판별에 있어 생성형 AI의 실질적인 역할과 효과를 평가하는 탐색적 연구로서 의미가 있다.

2. 법령 정보 확인 및 규제 판별

(1) 법령 정보 확인

국민의 권리와 의무를 규율하는 법령 정보에 언제 어디서나 접근할 수 있도록 보장하는 법령 정보 접근권은 민주주의와 법치주의의 핵심을 이루는 기본권으로 인식된다(방동

희, 2021). 이에 따라, 국가는 국민의 알 권리를 보장하기 위해 법령 정보를 제공할 의무를 지며, 이는 「법령정보의 관리 및 제공에 관한 법률」(이하 법령정보법)을 통해 구체화되어 있다. 「법령정보법」 제3조에서는 국가와 지방자치단체가 법령 정보를 신속하고 정확하게 제공할 의무가 있음을 명시하고 있으며, 이를 이행하기 위해 현재 ① 관보를 통한 제공, ② 대한민국 현행법령집 편찬, ③ 국가법령정보센터(www.law.go.kr)를 통한 전자적 제공 방식이 활용되고 있다(방동희, 2021: 246).

이 가운데 국가법령정보센터는 현행법령, 연혁법령, 행정규칙, 판례·해석례 등 대한민국의 모든 법령 정보를 한곳에서 통합 검색할 수 있도록 하여, 법령 정보의 접근성을 높이고 국민의 이해도를 증진하며 법령 활용의 편의성을 강화하는 데 기여하고 있다(방동희, 2021: 254). 또한, 종이 법령집과 달리 전자적 방식으로 제공되는 국가법령정보센터는 법령 시행일에 즉시 정보를 확인할 수 있어 시의성을 확보할 수 있으며, 상하위 법령 간 연계, 법령 체계도, 전자 법령집, 모바일 앱 서비스 등 다양한 부가 서비스를 통해 이용자의 편의성을 더욱 향상하고 있다. 2024년 11월 기준, 국가법령정보센터는 총 467여만 건의 정보를 제공하고 있으며, 일일 평균 82만 명이 방문하는 등 국민이 법령 정보를 확인하는 대표적인 창구로 활용되고 있다(법제처, n.d.-a).

현재 국가법령정보센터의 법령 정보 서비스는 체계적으로 정리된 정형화된 법령 제공에 중점을 두고 있다. 이에 따라 사용자가 특정 키워드를 알고 있거나 법령 구조를 이해하고 있다면 원하는 정보를 쉽게 찾을 수 있지만, 그렇지 않으면 직관적인 정보 검색이 어려울 수 있다. 즉, 생성형 AI와 같은 자유로운 질의 방식이나 문어체 검색을 활용한 복잡한 검색에는 한계가 존재한다. 이에 따라 국민이 특정 법령을 정확히 알지 못하거나 맥락을 기반으로 한 검색을 원할 때 원하는 법령 정보를 효과적으로 찾는 데 어려움을 겪을 수 있다. 따라서 국가법령정보센터가 공신력 있는 법령 정보를 지속적으로 업데이트하며 제공하고 있음에도 불구하고, 법령에 대한 사전적 이해가 부족한 일반 국민은 해당 플랫폼을 직접 활용하기보다는 ChatGPT와 같은 범용 AI 서비스를 통해 법령 정보를 검색할 가능성이 높다. 앞으로 범용 AI 서비스를 통한 법령 정보 검색이 증가할 것으로 예상되는 만큼, 이러한 서비스에서 제공하는 법령 정보가 국가법령정보센터의 공식 정보와 비교해 어느 정도의 정확성을 갖추고 있으며 실질적으로 활용할 수 있는 수준인지 평가할 필요가 있다.

(2) 규제 여부 및 유무 판단

한국은 규제의 생성, 운용, 소멸 전 과정에서 정부가 개입하는 전주기적 규제 관리(regulatory management) 체계를 갖추고 있다. 이 체계에서 형성된 규제를 어떻게 관리하는지도 중요하지만, 그보다 앞서 법령이 규제로서 관리 대상에 포함될지를 정확히 판단하는 것이 핵심이다. 이는 규제로 판별된 법령 조문만이 지속적인 관리 대상이 되며, 비규제 조문은 규제 관리 체계에서 원천적으로 제외되기 때문이다.

현행 규제등록 체계에서는 제·개정되는 모든 법령 조문에 대해 해당 조문이 규제에 해당하는지 판단하는 절차가 필수적으로 요구된다. 구체적으로, 중앙행정기관의 장은 소관 법령을 제·개정할 때, 해당 법령이 규제적 성격을 가지는지를 국무조정실(규제조정실)의 규제심사관에게 사전 검토 의뢰해야 한다(국무조정실, 2024). 규제심사관과의 협의를 거쳐 규제로 확정된 조문은 공식적인 규제로 등록되며, 이후 총괄적인 규제 관리 체계에 편입된다.

법령 조문이 규제에 해당하는지를 판단할 때 일차적인 기준은 해당 조문이 「행정규제기본법」 제2조에서 정의하는 행정규제에 해당하는지다. 「행정규제기본법」 제2조 제1항 제1호에서는 행정규제를 "국가나 지방자치단체가 특정한 행정 목적을 실현하기 위해 국민(국내법을 적용받는 외국인을 포함)의 권리를 제한하거나 의무를 부과하는 것으로서, 법령·조례·규칙 등에 규정된 사항"으로 정의하고 있으며, 구체적인 범위는 시행령 제2조에서 규정하고 있다. 또한, 행정규제 적용이 제외되는 사항은 「행정규제기본법」 제3조에서 정하고 있다. 실무적으로는 국무조정실이 「행정규제 판단기준」을 운영하며, 이를 통해 행정규제의 주체, 객체, 목적, 내용, 형식을 종합적으로 검토한 후, 모든 기준을 충족할 때 행정규제로 판단한다. 또한, 해당 문서에서는 「행정규제기본법 시행령」 제2조 제1항 각호에 대한 해석례를 제공하며, 다양한 사례를 반영해 조건별 판단기준을 상세히 제시하고 있다. 이를 통해 행정규제 여부를 보다 명확하게 판단할 수 있도록 체계적인 해석을 지원한다.

다만, 현행 규제 판단 체계는 담당 공무원의 주관적 판단에 의존하는 한계를 가진다. 사람의 인지에 기반한 규제 판별 방식에서는 전문성 부족, 제한된 시간과 정보, 개인적 편향 등의 요인으로 인해 오류가 발생할 가능성이 높다. 또한, 규제 개념에 대한 명확한

정의가 부재하여, 규제를 잘못 이해함으로써 오판이 발생하는 때도 있다(이혁우, 2009). 이러한 문제로 인해 규제임에도 불구하고 비규제로 잘못 판단되거나, 반대로 비규제임에도 규제로 잘못 판단되는 오류가 발생할 수 있다. 이러한 문제는 규제 판단의 신뢰성과 일관성을 저해하며, 규제 관리 체계의 일관된 운영을 어렵게 만든다. 따라서 보다 객관적이고 체계적인 규제 판별 시스템을 도입하여 규제 판단의 정확성과 일관성을 높일 필요가 있다.

〈표 1〉 주체·객체·목적·내용·형식에 따른 행정규제의 판단기준

구분	포함	제외
주체	중앙행정기관, 지자체, 공무수탁사인	국회, 법원, 헌재, 선관위, 감사원
객체	내·외국(법)인, 법인격 없는 사단·재단	중앙행정기관·지자체 소속공무원 공무수탁사인·소속임직원
목적	특정한 행정목적 실현	일반사회 목적(민·상사)
내용	권리제한, 의무부과	국회·법원·헌재·선관위·감사원 사무, 형사·행형·보안처분/정보·보안, 과징금, 과태료 부과 및 징수, 징·소집, 동원·훈련. 방위산업, 군사시설, 군사기밀
형식	법령, 조례, 조례규칙, 고시 등	지침, 지시, 매뉴얼, 교육자료

출처: 국무조정실(2016: 3)에서 재구성

법령 제·개정 단계에서 특정 조문이 규제에 해당하는지를 판단하는 것이 행정 실무적 관점에서 중요하다면, 피규제자인 기업과 국민의 관점에서는 기존 규제의 내용을 보다 쉽게 확인할 수 있도록 하는 것이 더욱 중요하다. 현재 피규제자가 규제 정보를 확인할 수 있는 다양한 창구가 존재한다. 우선, 앞서 언급한 국가법령정보센터에서는 법령 조문별로 해당 조문이 행정규제에 해당하는지를 확인할 수 있다. 또한, 국무조정실이 운영하는 규제 정보 포털(www.better.go.kr)에서는 중앙행정기관과 지자체별 규제 현황 및 규제 입법 동향 등을 제공하고 있다. 그러나 이러한 정보 제공은 법령과 해당 규제 조문을 단순히 확인하는 수준에 머물러 있으며, 맥락에 따른 검색이나 규제 저촉 여부를 판단하는 데는 한계가 있다. 예를 들어, 특정 사업 모델에 적용되는 규제를 확인하는 용도로는 충분히 활용되기 어렵다.

단순한 정보 제공을 넘어, 정부가 기업에게 보다 적극적으로 규제 유무를 확인해 주는 제도도 운용되고 있다. 대표적인 사례가 실증특례 및 임시허가와 함께 규제샌드박스 제도의 일환으로 시행되고 있는 '규제 신속 확인제도'이다. 규제 신속 확인제도는 신기술을 활용한 사업을 추진하려는 기업 등이 규제 유무가 불분명하다고 판단할 경우, 규제 확인을 요청하면 해당 규제 부처가 30일 이내(최장 60일 이내)에 규제 여부를 공식적으로 확인해 주는 제도이다. 원칙적으로, 기한 내에 규제 부처가 회신하지 않으면 규제가 없는 것으로 간주되며, 이에 따라 해당 서비스(또는 제품)의 시장 출시가 가능해진다.

그러나 현재 규제 신속 확인 절차는 담당 공무원의 전문성과 경험에 상당 부분 의존하고 있으며, 규제 확인 과정 역시 전적으로 사람의 판단으로 이루어지고 있다. 즉, 신청된 서비스(또는 제품)가 특정 부처의 소관 규제에 적용되는지를 검토하는 과정에서 자동화된 분석 방식이나 기술적 솔루션이 전혀 활용되지 않고 있다. 이러한 방식은 '신속 확인'이라는 명칭과 달리 규제 확인 과정에 과도한 시간과 행정적 자원이 소요되는 구조적인 한계를 초래하고 있다. 또한, 담당 공무원의 경험과 직관에 의존한 의사결정은 인간의 인지적 한계로 인해 판단 오류를 발생시킬 가능성이 높다. 특히, 공무원 인사관리 시스템 특성상 담당 업무가 1~2년 주기로 변경됨에 따라, 규제에 대한 전문성과 지식이 부족한 공무원이 업무를 맡는 경우 오판의 위험성이 더욱 커진다.

아울러, 규제 확인의 정확성을 저해하는 요인으로는 분석해야 할 규제의 수가 지속적으로 증가하고 있다는 점, 규제 검토를 위한 정보의 질과 양이 제한적이라는 점, 혁신적인 사업 모델에 대한 공무원의 이해가 부족하다는 점, 그리고 사업 내용과 규제 간의 연계성을 명확히 파악하기 어렵다는 점 등이 있다. 이러한 문제들은 규제 신속 확인 절차의 신뢰성과 효율성을 떨어뜨림으로써, 객관적이고 체계적인 규제 분석 시스템의 도입이 필요함을 시사한다. 본 연구에서는 생성형 AI를 활용해 법령 정보를 분석하고, 이를 기존 행정규제 데이터와 비교하여 규제 여부 판단에서 AI의 적용 가능성을 탐색하고자 한다.

Ⅲ. 연구 방법

1. 연구 설계

본 연구는 생성형 AI 모델(ChatGPT)이 법령 검색 및 규제 판단 과정에서 제공하는 정보의 신뢰성과 활용 가능성을 평가하기 위해 설계되었다. 이를 위해 AI 모델이 제공하는 법령 및 규제 관련 응답이 실제 법령 문서와 얼마나 유사한지, 그리고 규제 여부를 얼마나 정확하게 판단하는지를 분석한다. 평가의 기준이 되는 기준 데이터(Ground Truth)는 공식 법령 정보 및 등록된 행정규제를 기반으로 설정하였으며, 이를 통해 생성형 AI 모델의 법령 검색 및 규제 판단 성능을 객관적으로 검증하고자 한다.

AI 모델의 성능 평가는 네 가지 주요 분석 과정을 통해 이루어진다.

첫째, AI가 제공한 정보가 실제 법령 정보와 얼마나 정확하게 일치하는지를 평가한다. 이를 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-Score, AUROC 등의 분류 성능 평가 지표를 활용하여 AI의 판단 성능을 검증한다.

둘째, 자연어 처리(NLP) 기반 평가를 수행한다. AI가 검색한 법령 조문의 텍스트 유사도를 측정하기 위해 ROUGE, Jaccard 계수 등의 유사성 평가 지표를 활용하여 AI가 제공한 법령 조문과 실제 원본 텍스트 간의 일치도를 분석한다.

셋째, AI가 동일한 질문에 대해 일관된 답변을 제공하는지 확인한다. 이를 위해 AI의 응답을 동일한 입력에 대해 3회 반복 생성하고, 그 결과 간 변동성을 측정하여 AI가 동일한 질문에 대해 안정적인 응답을 제공하는지 검증한다. 또한, AI가 규제 판단의 근거로 제공하는 내용의 일관성을 평가하기 위해 반복 생성된 응답 간 유사성을 비교한다. 이는 규제 판단의 근거가 명확히 존재하지 않는 상황에서도 AI가 자체적으로 일관된 논리를 유지하는지 검토하기 위한 과정이다.

넷째, 법령 및 조문의 속성에 따라 결과의 차이가 있는지를 분석한다. 구체적으로 소관부처, 법령 유형, 실제/가상 조문, 시행 기간, 조문 길이 등의 속성을 고려하여 AI의 응답이 특정 조건에서 변동성을 보이는지 평가한다. 이를 위해 AI 응답 결과가 법령 및 조문 속성별로, 통계적으로 유의미한 차이를 보이는지 검토하며, 필요한 경우 카이제곱 검정(chi-squared test), T-검정(t-test), 일원분산분석(One-Way ANOVA) 등의 방법을

적용하여 그룹 간 차이를 검정한다.

이러한 분석을 통해 AI 모델이 법령 검색 및 규제 판단 과정에서 신뢰할 수 있는 정보를 일관되게 제공하는지 평가하고, 특정 속성에서 발생하는 변동성의 원인을 규명하여 AI 모델의 개선 방향을 제시하고자 한다.

2. 데이터 수집

본 연구에서는 평가의 기준이 되는 법령(조문) 및 규제 정보 데이터를 다음과 같은 절차를 통해 수집하였다.

먼저, 법령 데이터는 2024년 10월 15일 기준, 규제 법률 수가 많은 상위 15개 부처를 대상으로 등록규제를 포함하는 법령을 추출하였다. 부처별로 법률 2개, 시행령 1개, 시행규칙 1개씩 총 60개 법령을 선정하였다. 이 과정에서 국회 또는 대법원 소관 법령, 그리고 2개 이상의 소관 부처를 가지는 법령은 제외하였다. 규제 법령 데이터는 규제정보포털(www.better.go.kr)의 규제 현황 페이지에서 부처별 무작위(random)로 추출하였다. 시행령 및 시행규칙의 경우, 해당 법령에 등록규제가 포함되지 않으면 무작위 추출을 다시 수행하여 규제 조문이 포함된 법령을 확보하였다.

다음으로, 선정된 60개 법령에서 총 150개 조문을 추출하였다. 법률의 경우 각 법령에서 규제 조문, 비규제 조문, 가상 조문을 각각 1개씩 선정하였다. 가상 조문의 경우 실제 존재하는 법령을 기반으로 조문 번호 및 조문 제목을 임의로 생성하였다. 시행령 및 시행규칙의 경우 각 법령에서 규제 조문 1개, 비규제 조문 1개씩 총 2개 조문을 추출하였다. 또한, 실제 존재하는 국내 법률을 기반으로 가상 조문을 생성했을 뿐만 아니라, 국내에 존재하지 않는 법률을 기반으로도 가상 조문을 생성하였다.[3] 이를 위해 법제처 세계법제정보센터(www.world.moleg.go.kr)에서 국내 법체계와 유사한 일본 법률 10개를 선택하고, 각 법률당 1개 조문을 추출하였다. 최종적으로 총 160개 조문을 구성하였다.

마지막으로, 추출된 조문의 기본 정보(소관 부처, 법령 유형, 실제/가상 조문, 시행 기

3) 국내 법률을 기반으로 생성한 가상 조문의 경우, 실제 조문 내용이 존재하지 않는다. 따라서 조문 내용을 기반으로 검색하는 '프롬프트②'에서는 이를 제외하였다. 반면, 해외 법률을 기반으로 생성한 가상 조문은 실제 조문 내용을 포함하고 있으므로 '프롬프트②'에 포함하였다.

간, 규제 여부, 조문 길이)를 확인한 후 변수화하여 분석에 활용하였다. 이러한 기준 데이터의 특성은 다음 〈표 2〉와 같다.

〈표 2〉 기준 데이터(Ground Truth)의 특성

구분		프롬프트① (조문 생성)		프롬프트② (규제 분류)	
		N	%	N	%
소관 부처	1. 기획재정부	10	6.3	8	6.2
	2. 교육부	10	6.3	8	6.2
	3. 과학기술정보통신부	10	6.3	8	6.2
	4. 행정안전부	10	6.3	8	6.2
	5. 문화체육관광부	10	6.3	8	6.2
	6. 농림축산식품부	10	6.3	8	6.2
	7. 산업통상자원부	10	6.3	8	6.2
	8. 보건복지부	10	6.3	8	6.2
	9. 환경부	10	6.3	8	6.2
	10. 고용노동부	10	6.3	8	6.2
	11. 여성가족부	10	6.3	8	6.2
	12. 국토교통부	10	6.3	8	6.2
	13. 해양수산부	10	6.3	8	6.2
	14. 중소벤처기업부	10	6.3	8	6.2
	15. 금융위원회	10	6.3	8	6.2
	16. 정보 없음	10	6.3	10	7.7
법령 유형	1. 법률	100	62.5	70	53.8
	2. 시행령	30	18.8	30	23.1
	3. 시행규칙	30	18.8	30	23.1
실제/가상 조문	1. 실제	120	75.0	120	92.3
	2. 가상	40	25.0	10	7.7
시행 기간 (2024. 10. 31 기준)	1. 3개월 미만	21	13.1	16	12.3
	2. 3 ~ 6개월 미만	34	21.3	30	23.1
	3. 6개월 ~ 1년 미만	37	23.1	27	20.8
	4. 1년 ~ 3년 미만	34	21.3	27	20.8
	5. 3년 이상	24	15.0	20	15.4
	6. 정보 없음	10	6.3	10	7.7

구분		프롬프트① (조문 생성)		프롬프트② (규제 분류)	
		N	%	N	%
규제 여부	1. 행정규제	60	37.5	60	46.2
	2. 비규제	100	62.5	70	53.8
조문 길이	1. 100자 미만	23	14.4	23	17.7
	2. 100 ~ 200자 미만	37	23.1	37	28.5
	3. 200 ~ 300자 미만	27	16.9	27	20.8
	4. 300자 이상	33	20.6	33	25.4
	5. 정보 없음	40	25.0	10	7.7
계		160	100.0	130	100.0

주: 개수는 조문 수 기준임. 프롬프트②에서는 국내 법률에서 생성된 가상 조문은 제외함.
출처: 저자 작성.

생성형 AI를 활용한 법령 및 규제 정보 생성 과정은 다음과 같이 진행되었다.

먼저, 법령 정보 생성(프롬프트①)은 2024년 10월 31일에 수행되었으며, 당시 사용된 ChatGPT 모델은 GPT-4o 및 GPT-4o mini였다.[4] 조문 내용을 생성하기 위해, AI 모델에 법령명, 조문 번호, 조문 제목을 검색 키워드로 입력하여 해당 조문의 내용을 생성하도록 요청하였다. 생성된 데이터의 일관성과 변동성을 평가하기 위해 총 160개 조문을 무작위로 배열하고, 동일한 조건에서 3회 반복하여 생성하였다.[5]

〈표 3〉 프롬프트① (조문 생성) 실행 단계

구분	내용
지시문 및 질문 내용	다음 대한민국 법령 및 조문정보에 관한 구체적인 조문내용을 찾으시오. 현재 시행되고 있는 조문에 대해 최대한 원문에 맞춰 상세하게 안내하시오. 법에 대한 설명은 포함하지 않고 조문 내용만 표출하시오. 이후 다른 명령이 없는한 입력된 법령 및 조문정보에 대해 계속 수행하시오.
조문 정보 입력(예시)	주차장법 시행규칙 7조 (노외주차장의 설치 통보 등)
응답 결과 (예시)	1. 법 제4조제1항에 따라 노외주차장을 설치하려는 자는 별지 제1호서식의 노외주차장설치(변경)신고서에 다음 각 호의 서류를 첨부하여

4) 2회차 실행에서는 경량화된 버전인 ChatGPT-4o mini가 적용되었다.
5) 반복 과정에서는 회차마다 메모리를 초기화한 후, 새로운 프롬프트를 입력하여 일관된 조건을 유지하였다.

	특별자치시장·특별자치도지사·시장·군수 또는 구청장(자치구의 구청장을 말한다. 이하 같다)에게 제출하여야 한다. 1) 노외주차장의 위치도 2) 대지 및 건축물의 평면도 3) 노외주차장의 구조도 4) 차량의 출입 경로와 주차장 진입로의 구조·규격을 표시한 서류 2. 제1항의 신고를 받은 특별자치시장·특별자치도지사·시장·군수 또는 구청장은 법 제4조제1항에 따라 노외주차장의 설치 여부를 확인하여야 한다.

출처: 저자 작성.

다음으로, 행정규제 분류(프롬프트②)는 2024년 11월 6일에 수행되었으며, GPT-4o를 활용하였다. 프롬프트②는 단순한 조문 내용 생성과 달리, 규제 여부를 판단하는 논리적 해석과 개념적 분류 작업이 수반되므로, 이를 보완하기 위해 사전 단계로 6개의 개념 질문을 먼저 제시하였다. 이 질문들은 '규제란 무엇인가?', '행정규제기본법에서 규제를 어떻게 정의하는가?', '규제와 비규제를 구분하는 기준은 무엇인가?' 등으로 구성되어 있으며, AI가 조문 판단에 앞서 일관된 해석 기준과 개념 체계를 형성하도록 유도하는 데 목적이 있다. 이를 통해 규제 판단 결과의 정합성과 신뢰도를 높이고자 하였다. 실제 정보 생성은 일부 가상 조문을 제외한 총 130개 조문을 대상으로 하였으며, 동일한 입력 조건에서 3회 반복 생성을 통해 결과의 일관성과 변동성을 평가하였다. AI 모델에는 법령명, 조문 번호, 조문 제목, 조문 내용을 입력한 후, 해당 조문의 규제 여부, 규제 판단 확신도, 규제/비규제 판단의 근거를 생성하도록 요청하였다.

〈표 4〉 프롬프트② (규제 분류) 실행 단계

구분	내용
사전 개념 질문	1) 규제란 무엇입니까? 2) 한국에서 규제에 관한 기본법은 무엇입니까? 3) 행정규제기본법에서 규제는 어떻게 정의됩니까? 4) 규제를 판별하는 기준은 무엇입니까? 5) 행정규제기본법에서 규제에서 제외하는 기준은 무엇입니까? 6) 행정규제기본법 제3조 제2항에서 말하는 규제의 제외 기준은 무엇입니까? 이러한 기준을 후속 질문에 적용해서 판단하시오.
지시문 및 질문 내용	다음은 대한민국 법령의 조문 내용이다. 해당 조문 내용을 근거로 행정규제 여부를 판별하시오. ①행정규제인 경우 1, 비규제인 경우 0으로 응답하시오. ②행정규제라고

	어느 정도 확신하는지 0%(명확한 비규제) ~ 100%(명확한 행정규제) 사이의 값(정수)을 제시하시오. 명확한 행정규제인 경우 100에 가까운, 명확한 비규제일수록 0에 가까운 숫자를 제시하시오. ③해당 조문이 행정규제 또는 비규제라고 판단한 근거를 제시하시오.
조문 정보 입력(예시)	재해구호법 33조 (재해구호업무의 위탁) 구호기관은 재해구호업무를 효율적으로 수행하기 위하여 대통령령으로 정하는 바에 따라 다음 각 호의 업무를 구호지원기관에 위탁할 수 있다. 1. 이재민등에 대한 급식 제공 2. 구호세트의 제작 및 재해구호물자의 관리·공급 3. 구호물자 보관창고의 설치·운영 및 관리 4. 그 밖에 행정안전부장관이 필요하다고 인정하는 재해구호업무
응답 결과 (예시)	재해구호법 33조 (재해구호업무의 위탁) 규제 여부: 0 규제 확신도: 20% 판단 근거: 구호기관이 재해구호업무를 효율적으로 수행하기 위해 구호지원기관에 업무를 위탁할 수 있는 규정입니다. 이는 구호업무 수행 방식에 대한 내부적 지침에 해당하며 국민의 권리나 의무에 직접적인 제한을 가하지 않으므로 규제가 아닙니다.

출처: 저자 작성.

3. 분석 방법

(1) 분류 성능 평가 및 유사성 평가

생성형 AI가 수행하는 조문 정보 검색 및 규제 판별 결과의 정확성과 신뢰성을 평가하기 위해 본 연구에서는 다양한 성능 평가 지표를 활용하였다. 이를 위해, 자연어 처리(NLP) 및 머신러닝 분야에서 대표적으로 사용되는 유사성 평가 지표와 분류 성능 평가 지표를 적용하였다.

우선, 법령 정보 중 분류적 성격을 지닌 변수(ex. 조문 길이)와 규제 여부 판별에 대해 AI가 올바르게 판단했는지를 평가하기 위해 분류 성능 평가 지표를 활용하였다. 이진 분류에서 분류 성능 평가 지표 산출하는 데 필요한 기본 요소는 다음 〈표 5〉와 같으며, AI 모델의 분류 성능 평가에서는 모델의 예측 오류 사례(ex. 위음성, 위양성)를 줄이는 것이 중요하다.

〈표 5〉 이진 분류에서의 혼동 행렬

		예측 값	
		양성(Positive)	음성(Negative)
실제 값	양성(Positive)	진양성(True Positive, TP)	위음성(False Negative, FN)*
	음성(Negative)	위양성(False Positive, FP)*	진음성(True Negative, TN)

주: * 모델의 예측 오류
출처: Sokolova & Lapalme (2009: 429).

본 연구에서는 분류 성능 평가 지표로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-Score, AUROC(Area Under the Receiver Operating Characteristic Curve)를 활용하였다. 첫째, 정확도는 모델의 전반적인 성능을 나타내는 대표적인 지표로, AI가 전체 예측 중 정답을 맞힌 비율을 의미한다. 둘째, 정밀도는 모델이 긍정(Positive)으로 예측한 값 중 실제로 긍정인 비율을 의미한다. 예를 들어 AI가 규제라고 예측한 사례 중 실제 규제 사례의 비율이라고 할 수 있다. 셋째, 재현율은 실제 긍정인 데이터 중에서 모델이 올바르게 긍정으로 예측한 비율을 의미한다. 예를 들어 실제 규제 사례 중 AI가 정확하게 예측한 비율이다. 넷째, F1-Score는 정밀도와 재현율의 조화 평균으로, 두 지표 간의 균형을 평가한다. 일반적으로 정밀도를 높이면 재현율이 낮아지고, 반대로 재현율을 높이면 정밀도가 낮아지는 경향이 있는데(Manning, Raghavan, & Schütze, 2008), F1-Score는 이러한 상충 관계를 고려하여 AI 모델의 전반적인 성능을 평가하는 지표이다. 정확도, 정밀도, 재현율, F1-Score는 모두 0에서 1 사이의 값을 가지며, 일반적으로 0.8 이상이면 우수한 성능을 의미하고, 0.5 이하이면 낮은 성능을 의미한다(Walker II, n.d.). 마지막으로 AI 모델이 제시하는 규제 여부에 대한 확신도(0 ~ 100%)를 기반으로 AUROC를 산출하였다. AUROC는 ROC 곡선 아래 면적을 의미하며, 분류 모델의 전반적인 성능을 평가하는 지표이다. 여기서 ROC 곡선은 모델의 임계값을 변화시키며, 진짜 양성률(True Positive Rate, TPR)과 거짓 양성률(False Positive Rate, FPR)을 각각 계산하여 그린 곡선을 의미한다(Fawcett, 2006). ROC 분석은 일반적으로 양성과 음성 데이터의 비율이 균등한 경우 적절한 평가 지표로 사용될 수 있다(Flach & Kull, 2015).[6] AUROC는 0과 1 사이의 값을 가지며,

6) 기준 데이터(Ground Truth)에서 규제와 비규제의 비율은 54:46으로, 불균형 비율(Imbalance Ratio, IR)은

1에 가까울수록 모델의 분류 성능이 우수함을 의미한다. 한편, AUROC 값이 0.5인 경우 모델이 무작위로 분류한 것과 동일한 수준의 성능을 나타내며, 0.5 미만인 경우 무작위 분류보다 열등한 성능을 의미한다(Hanley & McNeil, 1982; Jaskowiak, Gosta, & Campello, 2021).

다음으로, 법령 및 조문의 내용 검색 성능과 규제 판별 근거에 대한 AI 모델의 성능을 측정하기 위해 본 연구에서는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)와 Jaccard 계수(Jaccard Coefficient)를 활용하였다. 첫째, ROUGE는 기계 번역 및 요약 모델의 성능 평가에 널리 사용되는 지표로, 원본 문서와 AI가 생성한 텍스트 간의 n-gram 중첩도를 기반으로 정량적 비교를 수행한다(Lin, 2004). 본 연구에서는 ROUGE-1과 ROUGE-2를 활용하여 법령 검색 결과와 원본 법령 문서 간 유사도를 평가하였다.[7] 구체적으로 ROUGE에서는 정밀도, 재현율, F1-Score를 산출하였다. 둘째, Jaccard 계수는 두 개의 문서가 공유하는 단어 집합의 비율을 측정하는 방식으로, 법령 조문과 AI 문서 간 일치도를 분석하는 데 유용하다(Manning et al., 2008). ROUGE와 Jaccard 계수는 모두 0부터 1 사이의 값을 가지며, 값이 클수록 두 텍스트 간의 유사도가 높음을 의미한다. 이상의 성능 평가 지표를 정리하면 다음 〈표 6〉과 같다.

〈표 6〉 분류 성능 및 유사성 평가 지표 요약

구분	평가지표	공식	출처
분류 성능 평가	정확도	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	Sokolova & Lapalme(2009)
	정밀도	$\mathrm{Pr}ecision = \frac{TP}{TP+FP}$	Sokolova & Lapalme(2009)
	재현율	$Recall = \frac{TP}{TP+FN}$	Sokolova & Lapalme(2009)
	F1-Score	$F1 = 2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall}$	Manning et al.(2008)
	AUROC	$AUROC = \sum_{i=1}^{n-1}(FPR_{i+1} - FPR_i) \times \frac{TPR_{i+1}+TPR_i}{2}$	Fewcett(2006)

1.17로 계산되었다. 이는 데이터가 균형에 가까운 상태로 판단되며, 따라서 AUROC를 적용하는 것이 적절하다.

7) ROUGE-1은 단어 단위의 n-gram(uni-gram) 일치율을 측정하며, ROUGE-2는 연속된 두 단어 단위의 n-gram(bi-gram) 일치율을 측정한다. ROUGE-2는 문맥적 유사도를 평가하는 데 유용하며, AI가 생성한 텍스트가 원본 문서의 단어 배열을 얼마나 충실히 따르는지를 분석할 수 있다.

구분	평가지표	공식	출처
유사성 평가	ROUGE 정밀도	$\text{Precision} = \frac{\text{겹치는 } n-gram \text{ 개수}}{\text{후보 요약의 전체 } n-gram \text{ 개}}$	Lin(2004)
	ROUGE 재현율	$Recall = \frac{\text{겹치는 } n-gram \text{ 개수}}{\text{참조 요약의 전체 } n-gram \text{ 개수}}$	Lin(2004)
	ROUGE F1-Score	$F1 = 2 \times \frac{\text{Precision} \times Recall}{\text{Precision} + Recall}$	Lin(2004)
	Jaccard 계수	$Jaccard\ Coefficient = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	Manning et al.(2008)

주: Jaccard 계수에서 A는 원본 텍스트의 단어 집합, B는 생성 텍스트의 단어 집합을 의미함
출처: 저자 작성.

(2) 변동성 분석

동일한 데이터를 사용하여 생성형 AI가 법령 정보 검색 및 규제 판별을 3회 반복 수행한 후, 결과값의 변동성을 평가하였다. 범주형 변수의 경우, 3회 반복 수행한 결과의 일치도를 분석하였으며, 연속형 변수에 대해서는 변동계수(coefficient of variation, CV)를 계산하였다. 일치도는 전체 시행 결과에서 동일한 결과가 나온 비율로 정의하였다.[8]. 일치도의 값은 0과 1 사이에 위치하며, 1에 가까울수록 반복 수행 시 결과의 일관성이 높음을 의미한다. 한편, 변동계수(CV)는 연속형 변수의 변동성을 평가하는 지표로, 표준편차를 평균으로 나눈 값으로 정의된다. 변동계수가 0이면, 모든 값이 동일하여 변동성이 없음을 의미한다. 반면, 0보다 크면 일정한 변동성이 존재하며, 1을 초과하면 변동성이 평균보다 크고 값의 분포가 넓음을 나타낸다. 따라서, 변동계수가 작을수록 결과의 변동성이 낮고, 일관성이 높다고 볼 수 있다. 마지막으로, 반복 측정의 일관성을 평가하기 위해 급내상관계수(Intraclass Correlation Coefficient, ICC)를 계산하였다. ICC 값은 0과 1 사이의 범위를 가지며, 값이 1에 가까울수록 반복 측정값 간의 신뢰도가 높음을 의미한다.[9] 즉, ICC가 높을수록 측정값들이 서로 유사하고, 재현성이 우수함을 나

8) 예를 들어, 3회 반복 수행한 결과가 (1, 0, 1)일 경우, 동일한 결과가 나온 비율을 계산하면, 전체 3회 중 동일한 값(1)이 2회 등장하므로 일치도는 $\frac{2}{3} \approx 0.67$로 계산된다.

9) ICC 값이 0.5 미만이면 낮은 신뢰도(poor reliability)를 나타내며, 0.5에서 0.75 사이는 중간 수준의 신뢰도(moderate reliability), 0.75에서 0.9 사이는 높은 신뢰도(good reliability), 그리고 0.9를 초과하는 경우 매우 높은 신뢰도(excellent reliability)를 의미한

타낸다.

(3) 차이 검정

본 연구에서는 AI가 생성한 데이터의 속성을 분석하고, 기준 데이터의 특성(법령 유형, 시행 기간 등)에 따라 AI의 법령 정보 생성 및 규제 판단 결과에 유의미한 차이가 있는지를 평가하였다. 이를 위해 평가지표 및 분석 변수의 속성을 고려하여 카이제곱 검정(Chi-square test), t-검정(t-test), 일원분산분석(One-Way ANOVA)을 수행하였다.

구체적으로, AI가 생성한 조문 및 규제 정보에 대해 소관 부처, 법령 유형, 시행 기간, 실제/가상 조문, 조문 길이, 규제 여부, 시행 회차에 따른 차이를 분석하였다. 이를 바탕으로, 기준 데이터의 속성에 따라 AI 생성 데이터의 응답이 어떻게 달라지는지를 평가하였다.

- 프롬프트① (조문 생성): 소관 부처, 법령 유형, 시행 기간, 실제/가상 조문, 조문 길이, 시행 회차에 따른 차이 검정
- 프롬프트② (규제 분류): 소관 부처, 법령 유형, 시행 기간, 규제 여부, 조문 길이, 시행 회차에 따른 차이 검정

(4) 통계 분석

정확도, 정밀도, 재현율, F1-Score, AUROC와 같은 분류 성능 평가 지표는 Stata 14.2를 이용하여 계산하였다. 한편, 내용 검색 결과의 유사성을 평가하기 위한 ROUGE 및 Jaccard 계수는 Python 3.13을 사용하여 산출하였다. 마지막으로, 변동성 분석 및 차이 검정 또한 Stata 14.2를 통해 수행하였다.

변수별 평가 방식을 살펴보면, 조문 길이 및 규제 여부는 분류 성능 평가를 수행하였으며, 조문 내용과 규제 판단 근거는 ROUGE 및 Jaccard 계수를 활용하여 유사성을 평가하였다. 변동성 분석은 모든 평가 지표에 적용하였으며, 차이 검정은 주요 변수(소관 부

다(Koo & Li, 2016: 158).

처, 법령 유형, 시행 기간, 실제/가상 조문, 조문 길이, 규제 여부, 시행 회차)에 따라 AI 생성 데이터의 응답 차이를 분석하는 방식으로 진행되었다. 모든 통계 검정은 유의수준 0.05에서 수행되었다($p < 0.05$ 기준).

Ⅳ. 연구 결과

1. 프롬프트① (조문 생성) 분석 결과

(1) 기준 데이터와 생성 데이터의 비교

프롬프트①의 결과로 생성된 데이터와 기준 데이터의 특성을 비교하면 다음 〈표 7〉과 같다. 우선 조문 길이 분포를 비교한 결과 기준 데이터에서 100자 미만의 조문 비율은 14.4%였으나, 생성 데이터에서는 38.1%로 증가하였다. 또한, 100~200자 미만 구간에서도 기준 데이터가 23.1%였던 반면, 생성 데이터에서는 50.8%를 차지하여 짧은 조문이 과도하게 생성되었음을 확인할 수 있다. 반면, 200~300자 미만(16.9% → 6.0%) 및 300자 이상(20.6% → 4.4%) 구간에서는 생성 데이터의 비율이 기준 데이터 대비 감소하였다. 이는 생성된 조문의 길이가 기준 데이터보다 짧은 경향을 보이며, 기준 데이터의 원래 조문 길이 분포를 제대로 반영하지 못했음을 의미한다. 또한, 기준 데이터에서 정보 없음으로 처리된 항목이 25.0%였던 반면, 생성 데이터에서는 0.6%로 거의 사라져 AI가 정보가 없는 경우에도 임의로 조문을 생성하는 경향이 있음을 보여준다. 연속형 변수 분석에서도 생성 데이터의 평균 조문 길이는 137자로 기준 데이터 평균(189자)보다 52자 짧아, AI 모델이 기준 데이터보다 상대적으로 짧은 조문을 생성하는 경향이 확인되었다.

〈표 7〉 프롬프트①의 기준 데이터와 생성 데이터 비교

(a) 범주형 변수

구분		기준 데이터		생성 데이터	
		N	%	N	%
조문 길이	1. 100자 미만	69	14.4	183	38.1
	2. 100 ~ 200자 미만	111	23.1	244	50.8
	3. 200 ~ 300자 미만	81	16.9	29	6.0
	4. 300자 이상	99	20.6	21	4.4
	5. 정보 없음	120	25.0	3	0.6
계		480	100.0	480	100.0

주: 본 데이터는 160개 조문을 3회 반복하여 총 480개로 구성됨.
출처: 저자 작성.

(b) 연속형 변수

구분		N	Mean	S.D.	Min.	Max.
조문 길이	기준 데이터	480	188.66	203.01	0	1,222
	생성 데이터	480	136.66	144.22	0	1,606

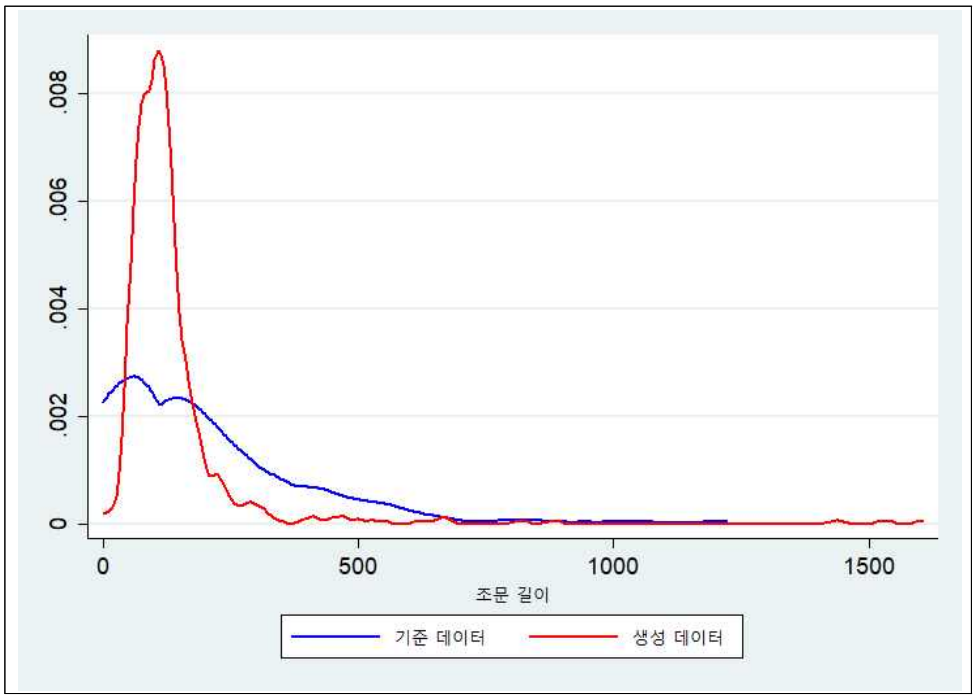
주: 기준 데이터에서 가상 조문은 조문 길이의 값으로 '0'을 입력함.
출처: 저자 작성.

(2) 조문 길이 추정 평가

〈그림 1〉은 기준 데이터와 생성 데이터의 조문 길이에 관한 밀도 분포를 보여준다. 〈그림 1〉에서 확인할 수 있듯이, 기준 데이터는 생성 데이터보다 다양한 조문 길이를 포함하며, 특정 길이에 과도하게 집중되지 않는 경향을 보인다. 반면, 생성 데이터는 100자 미만에서 밀도가 급격히 높아지고 긴 조문의 비율이 현저히 낮아, 짧은 조문이 과도하게 생성되는 편향이 나타났다.

한편, 〈표 8〉에서 AI가 생성한 조문 길이의 변동성을 살펴보면, 변동계수(CV)는 각 조문에 대해 3회 반복된 값의 변동계수를 구한 후, 변동계수의 평균을 산출한 값으로, 0.454로 나타나 반복 시행 간 개별 조문 길이의 변동성이 높은 것으로 확인되었다. 또한, 급내상관계수(ICC)도 0.028로 매우 낮아, 생성된 조문 길이가 기준 데이터의 길이 패턴을 거의 반영하지 못하고 있음을 보여준다.

〈그림 1〉 조문 길이의 밀도 분포(KDE Plot)



출처: 저자 작성.

〈표 8〉 조문 길이 추정 평가

구분	Average	CV	ICC
조문 길이	136.66	0.454	0.028

주1: Average는 3회 반복 실행한 평균값.
주2: CV는 개별 관측치 기준 변동계수의 평균.
출처: 저자 작성.

(3) 조문 내용의 유사성 평가

조문 내용의 유사성을 평가한 결과, 생성된 조문과 기준 조문 간의 전반적인 유사도는 낮은 수준으로 나타났다. 〈표 9〉에 따르면, Unigram(단어 단위) 기준 ROUGE Precision, Recall, F1-Score는 각각 0.362, 0.270, 0.273으로, 생성된 조문 중 기준 조문과 일치하는 단어의 비율은 36.2%에 불과하며, 기준 조문의 핵심 단어를 충분히

반영하지 못한 것으로 분석된다. Jaccard 계수 또한 0.202로 낮은 수치를 보여, 두 조문 간 단어의 교집합이 합집합에 비해 적음을 시사한다. 실행 간 변동성을 나타내는 변동계수(CV)는 0.408~0.466 범위로 데이터 간 변동성이 비교적 큰 수준으로 분석되었으며, 반복 실행 간의 급내상관계수(ICC)는 0.312~0.553 범위로, 낮음에서 중간 수준의 신뢰도를 나타냈다.

Bigram(2-연속 단어 단위) 기반에서는 유사성이 더욱 낮아지는 경향을 보였다. ROUGE Precision, Recall, F1-Score는 각각 0.189, 0.145, 0.146으로, 단어 수준을 넘어 연속된 문맥에서도 기준 조문과의 일치도가 낮은 것으로 분석되었다. Jaccard 계수는 0.102로 Unigram보다 더 낮아, 생성된 조문이 문맥적 연속성을 제대로 반영하지 못하고 있음을 시사한다. 변동계수(CV)는 0.640~0.689로 높은 수준이며, 급내상관계수(ICC)는 0.117~0.293으로 낮아, Bigram 수준에서는 반복 실행 간 결과의 신뢰도와 일관성이 전반적으로 부족한 것으로 나타났다.

〈표 9〉 조문 내용의 유사성 평가

구분		ROUGE Precision	ROUGE Recall	ROUGE F1-Score	Jaccard Coefficient
Unigram 단위	Average	0.362	0.270	0.273	0.202
	CV	0.421	0.452	0.408	0.466
	ICC	0.553	0.541	0.431	0.312
Bigram 단위	Average	0.189	0.145	0.146	0.102
	CV	0.664	0.660	0.640	0.689
	ICC	0.293	0.285	0.207	0.117

주1: Average는 3회 반복 실행한 평균값.
주2: CV는 개별 관측치 기준 변동계수의 평균.
출처: 저자 작성.

(4) 차이 검정 결과

본 연구에서는 기준 데이터의 조문과 생성된 조문 간의 분량 차이(절댓값 기준)가 비교변수에 따라 차이를 보이는지를 분석하기 위해 독립표본 t-검정과 일원분산분석을 수행

하였다(〈표 10〉 참조). 분석 결과, 소관 부처에 따라 조문 분량의 차이가 유의미하게 다르게 나타났다(F = 1.789, p 〈 0.05). 예를 들어, 기획재정부(90.5자)와 교육부(81.3자)는 생성된 조문과 기준 조문 간의 분량 차이가 상대적으로 적었으나, 보건복지부(224.1자), 문화체육관광부(198.0자), 여성가족부(196.3자)에서는 분량 차이가 크게 나타났다. 또한, 참조하는 조문의 길이도 생성된 조문의 분량 차이에 영향을 미치는 중요한 변수임이 확인되었다(F = 88.724, p 〈 0.001). 평균적으로 기준 데이터의 조문 분량이 많을수록 생성된 조문과 참조 조문 간의 분량 차이도 증가하는 경향을 보였다. 특히, 존재하지 않는 가상 조문의 경우 이론적으로 생성된 조문의 분량 차이 기댓값은 0이어야 하지만, AI 모델은 평균 123.08자의 조문을 생성하는 허위 생성 현상을 보였다.

〈표 10〉 비교 변수별 생성 조문과 참조 조문의 분량 차이 분석

비교 변수	검정유형	통계량(t, F)
1. 소관 부처	ANOVA	1.789 *
2. 법령 유형	ANOVA	0.074
3. 시행 기간	ANOVA	1.133
4. 실제/가상 조문	t-검정	1.738
5. 조문 길이	ANOVA	88.724 ***
6. 시행 회차	ANOVA	0.119

주: * p 〈 0.05 , ** p 〈 0.01 , *** p 〈 0.001
출처: 저자 작성.

다음으로 기준 조문과 생성 조문 간의 내용적 유사성이 비교 변수에 따라 어떻게 달라지는지를 분석한 결과는 〈표 11〉과 같다. 분석 결과, 법령 유형, 실제/가상 조문 여부, 조문 길이, 시행 회차에 따라 생성된 조문의 내용 유사도에 유의미한 차이가 존재하는 것으로 나타났다.

먼저, 법령 유형별로 ROUGE F1-Score 및 Jaccard 계수는 통계적으로 유의미한 차이를 보였다. 예를 들어, Unigram 단위 ROUGE F1-Score는 법률 0.234, 시행령 0.322, 시행규칙 0.354로 나타나, 법률의 내용 유사성이 가장 낮은 것으로 확인되었다. 이는 가상 조문이 법률 단위에 포함되어 나타난 결과일 가능성이 있다.

또한, 실제 조문과 가상 조문 간에도 유의미한 차이가 확인되었다. 실제 조문의

Unigram ROUGE F1-Score는 0.344, 가상 조문의 경우 0.060으로 나타났으며, Jaccard 계수 역시 실제 조문 0.256, 가상 조문 0.040으로 확인되었다.[10)]

참조 조문의 길이 또한 생성 조문의 내용 유사도에 영향을 미치는 주요 요인으로 나타났다. 조문 길이가 짧을수록 유사성이 높았으며, Unigram ROUGE F1-Score 기준으로 100자 미만(0.429) 〉 100~200자 미만(0.348) 〉 200~300자 미만(0.325) 〉 300자 이상(0.296)의 순으로 유사도가 감소하는 경향을 보였다. 이러한 결과는 앞서 확인된 AI 모델이 긴 조문에서도 짧은 조문을 형성하는 경향과 연관이 있는 것으로 보인다. 즉, AI 모델이 긴 조문을 충분한 길이로 생성하지 못하고 짧게 요약하는 경향을 보일 경우, 참조 조문의 길이가 길수록 생성 조문과의 유사성이 더욱 감소하는 결과로 이어질 가능성이 크다.

마지막으로, 시행 회차에 따른 유사성 차이도 확인되었다. Unigram ROUGE F1-Score 기준으로 1회차 0.234, 2회차 0.224, 3회차 0.361로 나타나, 실행 회차에 따라 생성 조문의 유사도가 다소 변동하는 경향을 보였다.

〈표 11〉 비교 변수별 조문 내용 생성 성능 평가

비교 변수	검정 유형	평가지표			
		Unigram 단위		Bigram 단위	
		ROUGE F1	Jaccard	ROUGE F1	Jaccard
1. 소관 부처	ANOVA	0.968	1.437	1.308	1.366
2. 법령 유형	ANOVA	12.818 ***	7.054 ***	4.444 *	2.048
3. 시행 기간	ANOVA	1.876	2.076	2.010	1.478
4. 실제/가상 조문	t-검정	13.974 ***	10.451 ***	7.308 ***	5.092 ***
5. 조문 길이	ANOVA	55.887 ***	32.245 ***	17.753 ***	11.063 ***
6. 시행 회차	ANOVA	19.020 ***	30.294 ***	31.595 ***	34.423 ***

주1: 검정 유형이 ANOVA일 경우, 표기된 값은 일원분산분석(One-Way ANOVA)의 F-값임. 검정 유형이 t-검정일 경우, 표기된 값은 독립표본 t-검정의 t-값임.
주2: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
출처: 저자 작성.

10) 가상 조문 간 유사성은 원칙적으로 0이어야 하나, 일부가 해외 법률을 참고해 작성되면서, 해당 법률 조문과 유사성이 계산되었다.

2. 프롬프트② (규제 분류) 분석 결과

(1) 데이터 특성 비교

〈표 12〉는 프롬프트②를 활용해 AI가 분류한 결과와 기준 데이터 간의 규제 여부 분포 차이를 비교한 것이다. 생성 데이터는 원문 조문을 그대로 입력한 후, 프롬프트② 방식으로 생성형 AI에게 규제 여부를 분류하도록 한 결과를 반영한 것이다. 기준 데이터(정답 레이블)에서는 행정규제로 분류된 조문이 180개(46.2%), 비규제 조문이 210개(53.8%)로 비규제 조문의 비중이 더 높았다. 반면, 생성형 AI가 분류한 결과에서는 행정규제가 255개(65.4%), 비규제가 135개(34.6%)로, 행정규제의 비중이 더 높게 나타났다. 이 결과는 동일한 조문에 대해 기준 데이터와 AI 분류 결과 간에 규제 여부 판단의 분포 차이가 존재함을 보여준다.

〈표 12〉 프롬프트②의 기준 데이터와 생성 데이터 비교

구분		기준 데이터		생성 데이터	
		N	%	N	%
규제 여부	1. 행정규제	180	46.2	255	65.4
	2. 비규제	210	53.8	135	34.6
계		390	100.0	390	100.0

주: 본 데이터는 130개 조문을 3회 반복하여 총 390개로 구성됨.
출처: 저자 작성.

(2) 규제 분류 성능 평가

규제 여부 분류에 대한 성능 평가 결과는 〈표 13〉과 같다. 세 차례 반복 실행을 통해 얻은 평균 정확도는 0.669로, 전체 조문 중 약 66.9%에 대해 생성형 AI가 규제 여부를 정확하게 분류한 것으로 나타났다. 정밀도(0.700)와 재현율(0.682) 또한 유사한 수준으로 나타나, AI가 행정규제를 과도하게 예측하거나 놓치는 편향 없이 비교적 안정적인 분류 성능을 나타냈다. 두 지표의 조화 평균인 F1-Score는 0.665로, 규제 분류에 있어

전반적으로 균형 잡힌 성능을 유지하고 있음을 보여준다. 아울러, 모든 지표에 대한 변동계수(CV)가 낮은 수준으로 나타나 반복 실행 간 결과의 일관성이 확보된 것으로 판단된다. 특히 조문별 반복 실행 시 동일한 분류 결과가 나온 비율인 일치도는 평균 0.915로, 매우 높은 수준의 결과 일관성을 나타냈다.

〈표 13〉 규제 여부 분류 성능 평가

구분	Accuracy	Precision	Recall	F1-Score
Average	0.669	0.700	0.682	0.665
All-CV	0.053	0.029	0.045	0.061

주1: Average는 3회 반복 실행한 평균값.
주2: All-CV는 전체 데이터 기준 변동계수.
출처: 저자 작성.

〈표 14〉는 생성형 AI가 조문별 규제 여부에 대해 산출한 확신도의 평균값과 그에 대한 성능 평가 지표를 제시한 것이다. 전체 조문에 대한 평균 확신도는 63.82%로, AI는 조문이 규제일 가능성에 대해 다소 높은 수준의 확신을 보인 것으로 나타났다. 또한 확신도의 변동계수(CV)는 0.254로 나타나, 조문 간 판단의 일관성 측면에서 다소의 분산이 존재하지만, 과도한 수준은 아닌 것으로 판단된다. 더불어, 반복 실행 간 판단의 신뢰도를 나타내는 ICC는 0.699로, 동일 조문에 대한 AI의 규제 판단이 비교적 안정적으로 유지되었음을 보여준다.

〈표 14〉 규제 판단 확신도 성능 평가

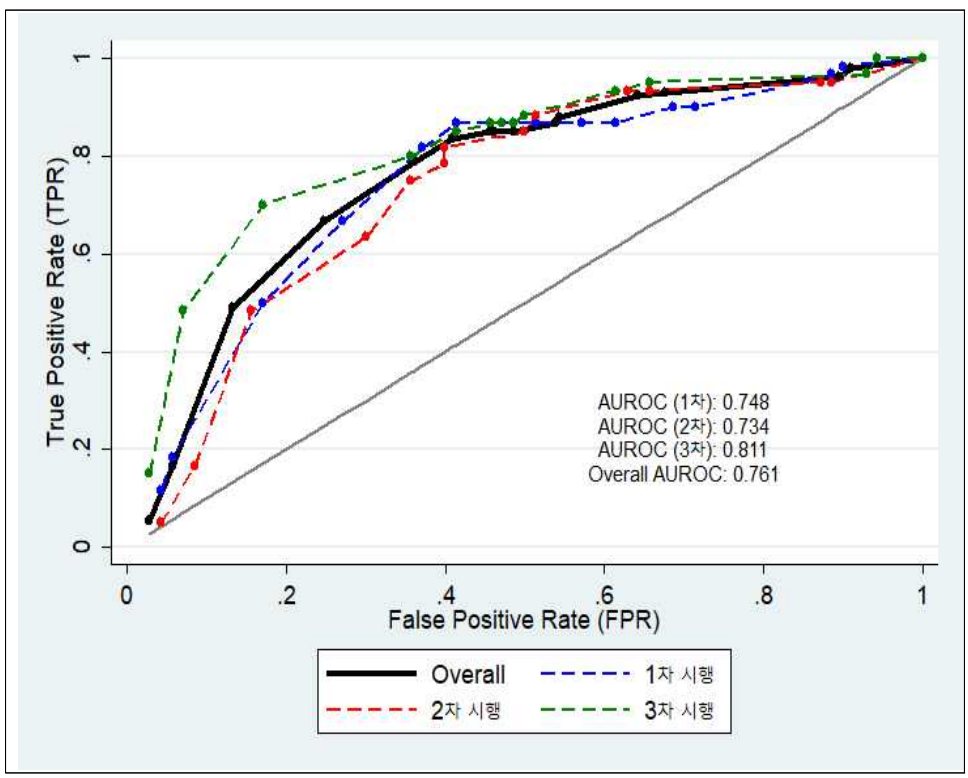
구분	Average	CV	ICC
규제 판단 확신도 (0 ~ 100%)	63.82	0.254	0.699

주1: Average는 3회 반복 실행한 평균값.
주2: CV는 개별 관측치 기준 변동계수의 평균.
출처: 저자 작성.

〈그림 2〉는 생성형 AI가 조문별 규제 여부에 대해 산출한 확신도와, 실제 정답(기준 데이터) 간의 관계를 ROC 곡선 및 AUROC 지표를 통해 평가한 결과를 나타낸다.

AUROC는 분류 임곗값을 변화시키며 계산된 민감도(True Positive Rate)와 위양성률(False Positive Rate)의 관계를 종합적으로 반영하는 지표로, 1에 가까울수록 예측 성능이 우수함을 의미한다. 실행 결과, 1회차 AUROC는 0.748, 2회차는 0.734, 3회차는 0.811로 나타났으며, 전체 평균 AUROC는 0.761로 확인되었다. 이는 ChatGPT가 산출한 확신도 수치가 실제 규제 여부와 일정 수준 이상의 정합성을 갖고 있으며, 확신도 값이 규제 가능성을 예측하는 신호로 활용될 수 있음을 보여준다. 특히 확신도가 높을수록 실제 규제일 가능성이 높아지는 경향은 ROC 곡선의 형태를 통해 확인되었으며, 반복 실행 간 AUROC 값이 안정적으로 유지된 점은 응답의 일관성 측면에서도 의미가 있다.

〈그림 2〉 규제 판단 확신도의 AUROC 평가



출처: 저자 작성.

(3) 규제 판단 근거 응답의 유사성 평가

〈표 15〉는 생성형 AI가 동일한 조문에 대해 세 차례 반복 수행한 규제 여부 판단 응답의 근거 문장 간 유사도를 평가한 결과이다. 본 분석에서는 기준값 없이 생성된 응답 간의 일관성을 파악하기 위해, 각 실행 결과 간(1-2, 1-3, 2-3)의 텍스트 유사도를 계산하고 평균을 산출하였다. 유사도 평가는 ROUGE Precision, Recall, F1-Score 및 Jaccard 계수를 기준으로 수행되었다. 그 결과, Unigram 기준 ROUGE F1-Score는 0.464, Jaccard 계수는 0.335로 나타나, 단어 수준에서는 ChatGPT의 판단 근거가 일정 부분 반복적으로 유사한 표현을 사용하는 경향이 있음을 보여준다. 반면, Bigram 기준에서는 ROUGE F1-Score가 0.196, Jaccard 계수가 0.113으로 낮게 나타나, 문장 구조나 어순 등 구 단위의 표현 방식에서는 반복 응답 간 차이가 나타났다. 이는 생성형 AI의 판단 근거가 핵심 단어는 유지하되, 문장 구성이나 표현 방식에서는 변화를 보이는 경향이 있음을 시사한다.

〈표 15〉 규제/비규제 판단 근거의 유사성 평가

구분	ROUGE Precision	ROUGE Recall	ROUGE F1-Score	Jaccard Coefficient
Unigram	0.464	0.524	0.464	0.335
Bigram	0.196	0.223	0.196	0.113

주: ROUGE, Jaccard Coefficient 점수는 각 실행 결과 간(1-2, 1-3, 2-3) 비교하여 평균을 계산한 값임
출처: 저자 작성.

(4) 차이 검정 결과

〈표 16〉은 기준 데이터와 생성 데이터 간 규제 판단 일치 여부(1=일치, 0=불일치)가 다양한 비교 변수에 따라 차이를 보이는지를 확인하기 위해 실시한 카이제곱 검정 결과를 나타낸다. 분석 결과, '규제 여부' 변수에서 유의미한 차이가 나타났으며(χ^2 = 49.348, $p < .001$), 이는 조문이 규제인지 비규제인지에 따라 생성형 AI의 판단이 기준 데이터와 일치하는 경향에 유의한 차이가 존재함을 의미한다. 반면, 소관 부처, 법령 유

형, 시행 기간, 조문 길이, 시행 회차 등의 변수에서는 통계적으로 유의한 차이가 나타나지 않아, 해당 요소들은 규제 판단 일치 여부에 영향을 주지 않는 것으로 나타났다.

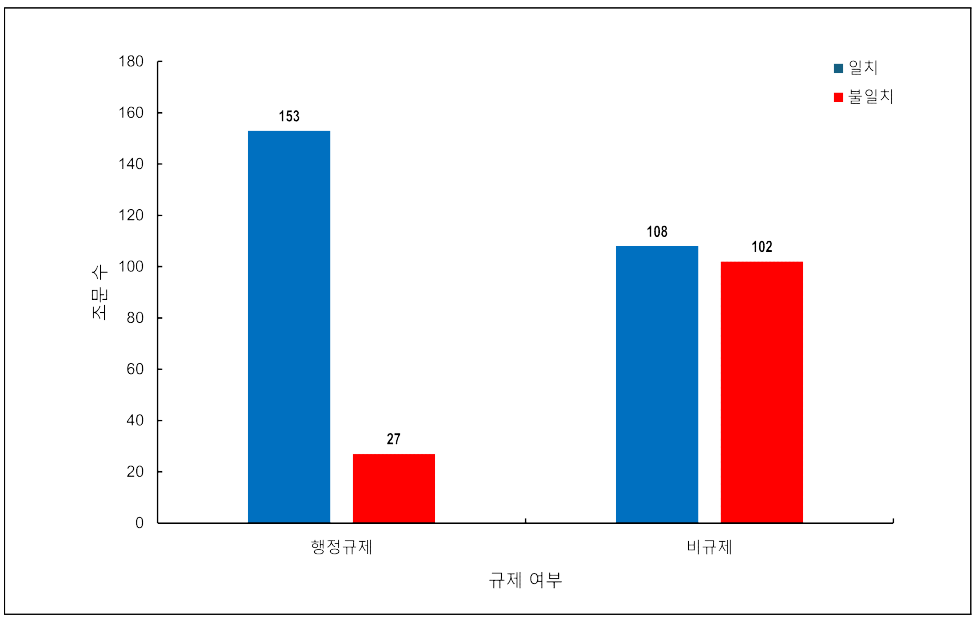
〈표 16〉 규제 판단 일치 여부와 비교 변수 간 카이제곱 검정 결과

비교 변수	X^2	df
1. 소관 부처	18.962	15
2. 법령 유형	2.792	2
3. 시행 기간	4.388	5
4. 규제 여부	49.348 ***	1
5. 조문 길이	7.142	4
6. 시행 회차	1.460	2

주: * p 〈 0.05 , ** p 〈 0.01 , *** p 〈 0.001
출처: 저자 작성.

〈그림 3〉은 기준 데이터의 규제 여부에 따라 생성형 AI의 분류 결과가 일치하거나 불일치한 빈도를 막대그래프로 시각화한 것이다. 비규제 조문의 경우, 전체 210개 중 108건(51.4%)이 일치하고 102건(48.6%)이 불일치하여, 일치와 불일치가 거의 비슷한 수준으로 나타났다. 반면, 행정규제 조문은 총 180개 중 153건(85.0%)이 일치하고 27건(15.0%)이 불일치하여, 상대적으로 높은 일치율을 보였다. 이러한 시각적 분포는 생성형 AI(ChatGPT)가 행정규제 조문에 대해서는 비교적 안정적인 분류 성능을 보이는 데 비해, 비규제 조문에서는 오분류(misclassification)가 보다 빈번하게 발생함을 보여준다.

〈그림 3〉 규제 여부에 따른 ChatGPT 분류 일치/불일치 빈도



출처: 저자 작성.

V. 논의 및 결론

본 연구는 생성형 인공지능(AI), 특히 ChatGPT를 활용하여 법령 조문 정보의 생성 및 규제 여부 판단이 가능한지를 실증적으로 분석하고, 그 신뢰성과 활용 가능성을 평가하는 데 목적을 두었다. 이를 위해 두 가지 주요 연구 질문을 설정하였다. 첫째, 생성형 AI가 제공하는 법령 및 조문 정보가 실제 레퍼런스(공식 법령 문서)와 얼마나 유사한지를 확인함으로써 정보 생성의 신뢰성을 검토하고자 하였다. 둘째, AI가 주어진 조문을 바탕으로 규제 여부를 얼마나 정확하게 판단할 수 있는지를 평가함으로써 규제 분류 도구로서의 실효성을 검토하였다. 이러한 연구 목적에 따라 프롬프트①(조문 생성)과 프롬프트②(규제 분류)의 두 가지 분석 설계를 바탕으로, 분류 정확도, 내용 유사도, 실행 일관성, 비교 변수에 따른 성능 차이 등을 중심으로 평가를 수행하였다.

조문 생성 결과 분석에 따르면, 생성형 AI는 기준 조문에 비해 짧은 길이의 조문을 과도하게 생성하는 경향을 보였으며, 이에 따라 조문 길이의 분포에서 구조적 편향이 나타났다. 예컨대 100자 미만의 조문 비율이 기준 데이터보다 2배 이상 높게 나타난 데 비해, 300자 이상의 조문은 현저히 줄어들어, AI가 장문의 조문을 생성하는 데 구조적 한계를 가지고 있음을 보여주었다. 실제로 생성된 조문의 평균 길이는 기준 조문보다 약 52자 짧았으며, 이는 AI가 법령 문서의 정보 밀도와 서술 구조를 충분히 재현하지 못하고 있음을 시사한다. 또한 정보 결손 상황에서도 AI가 임의로 조문을 생성하는 허위 생성(hallucination) 현상이 확인되었으며, 이는 법령과 같이 정확성이 필수적인 분야에서 생성형 AI의 단독 사용이 위험할 수 있음을 보여준다. 생성된 조문의 내용 유사도 또한 전반적으로 낮은 수준을 나타냈다. ROUGE 및 Jaccard 계수를 활용한 평가에서 Unigram 기준 평균 F1-score는 0.273, Bigram 기준은 0.146으로 나타났으며, 이는 단어 수준을 넘어 문맥적 연속성까지 포함한 조문 구성에서 AI가 기준 조문과의 유사성을 충분히 확보하지 못했음을 의미한다. 더불어 반복 실행 간 변동계수가 높고 급내상관계수(ICC)는 낮아, 생성 결과의 일관성과 재현성 또한 불충분한 것으로 나타났다.

반면, 프롬프트②를 활용한 규제 여부 분류 과제에서는 비교적 우수한 성능과 높은 실행 일관성이 확인되었다. 평균 정확도는 66.9%였으며, 정밀도(0.700), 재현율(0.682), F1-Score(0.665) 등 주요 성능 지표가 균형 있게 나타났다. 특히 동일 조문에 대한 반복 실행 결과의 일치율이 평균 91.5%에 달해, AI의 규제 판단이 일정 수준 이상의 신뢰성과 일관성을 유지하고 있음을 보여주었다. 다만, 세부적으로 살펴보면 규제 조문에 대한 분류 정확도는 85.0%로 비교적 높은 수준을 보인 반면, 비규제 조문은 정확도가 51.4%에 그쳐 뚜렷한 차이를 나타냈다. 즉, 비규제 조문에 대해서는 규제로 잘못 분류되는 사례가 다수 나타나, 향후 비규제 판단의 정밀도를 높이기 위한 보완이 필요함을 시사한다. 한편, AI가 산출한 규제 확신도는 실제 규제 여부와의 정합성이 높았고, AUROC 값이 평균 0.761로 나타나, 확신도가 보조적인 판단 지표로 활용될 가능성도 확인되었다.

생성 결과에 영향을 미치는 외부 요인을 확인하기 위해 수행한 차이 검정 분석 결과, 조문 생성 과제에서는 다양한 변수가 AI 성능에 유의미한 영향을 미쳤지만, 규제 분류 과제에서는 상대적으로 독립적인 판단 경향이 나타났다. 조문 생성에서는 소관 부처에 따라 생성 조문과 기준 조문 간의 길이 차이에 통계적으로 유의한 차이가 있었으며, 참조

조문이 길수록 생성 조문과의 분량 차이도 벌어지는 경향이 나타났다. 또한, 실제 조문과 가상 조문 간 유사성에도 유의한 차이가 존재하였고, 존재하지 않는 조문에 대해서도 평균 123자의 내용을 생성하는 허위 생성 현상이 나타났다. 이는 생성형 AI가 불완전하거나 없는 정보에 대해서도 실제처럼 보이는 결과를 제시하는 경향이 있으며, 정보 왜곡의 가능성을 보여준다. 반면, 규제 분류 과제에서는 '규제 여부' 변수만이 분류 결과의 정확도에 유의한 영향을 미쳤고, 소관 부처, 법령 유형, 시행 회차 등의 요소는 통계적으로 유의미한 차이를 보이지 않았다. 이는 AI가 규제 판단 시 외형적 요인보다는 조문 내 규제 표현 자체에 반응하여 판단을 수행하고 있음을 의미한다.

본 연구는 생성형 AI의 법령 및 규제 분석 기능에 대한 탐색적 접근으로서 의의를 지니나, 다음과 같은 한계점을 갖는다. 첫째, 분석 대상이 특정 모델(ChatGPT)에 한정되어 있어, 다양한 생성형 AI 모델 간 성능 차이나 알고리즘적 특성을 비교하는 연구는 수행되지 않았다. 둘째, 비규제 조문이 규제로 과대 분류되는 현상의 구체적 원인은 규명되지 않았다. 생성형 AI가 규제적 표현에 민감하게 반응했을 가능성은 있으나, 규제 조문과 비규제 조문 간의 언어적·구조적 차이에 대한 정성적 분석이 이루어지지 않았기 때문에, 오분류의 원인을 체계적으로 설명하는 데에는 한계가 있다. 셋째, 일부 분석에서는 기준 데이터의 정확성과 타당성에 대한 검토가 필요하다. 예를 들어, 행정규제 등록 여부를 기준으로 분류할 경우, 실제로 등록된 규제 항목이 규제가 아님에도 규제로 분류되거나, 규제임에도 규제가 아닌 것으로 분류되는 양방향 오류가 존재한다. 또한, 규제 판단의 근거 문장에 대해서는 명확한 정답 데이터가 존재하지 않아, 생성형 AI의 응답을 평가할 기준 자체가 불완전하다는 한계가 있다. 향후 연구에서는 전문가 판단을 포함한 기준 정립과 정성적 검증 절차가 병행될 필요가 있다. 넷째, 본 연구에 활용된 ChatGPT는 법령 정보나 규제 여부에 대해 '추론' 방식으로 응답하였으나, 향후 국가법령정보센터나 규제정보포털 등과 연계되어 사실 기반(fact-based) 정보를 참조할 수 있게 될 경우, 동일한 과제에 대한 AI의 응답 방식과 성능은 달라질 수 있다. 이 경우, 현재와 같은 추론 중심의 분석은 적용 범위에 일정한 한계를 가질 수 있다. 따라서 본 연구의 결과는 생성형 AI의 추론 능력을 중심으로 한 탐색적 분석의 일환으로 이해되어야 한다.

본 연구의 결과를 종합하면, 생성형 AI는 법령 정보를 생성하는 과제보다는 규제 여부를 분류하는 과제에서 더 높은 안정성과 활용 가능성을 보였다. 조문 생성의 경우, 내용

누락, 허위 생성, 실행 간 불일치 등 여러 문제가 복합적으로 존재하여, 생성형 AI의 결과를 신뢰하기는 아직 어렵다고 판단된다. 반면, 규제 분류는 비교적 우수한 정량적 성능과 실행 일관성을 확보하고 있으며, 특히 행정규제 판단에서의 높은 정확도와 확신도 기반 정합성은 실무에서 AI를 규제 검토 보조 도구로 활용할 수 있음을 시사한다.

그러나 비규제 조문에 대한 분류 정확도는 상대적으로 낮게 나타났고, 이는 AI가 규제적 표현에 과도하게 반응하거나 조문 구조의 미묘한 차이를 충분히 해석하지 못하는 데서 비롯된 것으로 보인다. 따라서 실제 활용 시에는 편향 가능성을 고려한 보완 체계를 마련할 필요가 있으며, 이에는 프롬프트 설계(prompt engineering)의 고도화나 과제 특성에 적합한 모델 선택 전략과 같은 사용자 중심의 대응이 포함될 수 있다. 이러한 대응 방안에 대한 실증적 검토는 생성형 AI의 실무 활용 조건을 구체화하고, 정책적 함의를 도출하는 데 기여할 수 있다.

한편, 최근 규제 정보를 더욱 효율적으로 탐색하고 제공하기 위한 도메인 특화형 AI 시스템 개발도 활발히 이루어지고 있다. 예를 들어, 한국행정연구원의 '레그네비게이터', 창업진흥원의 '규제파인더' 등은 규제 분야에 특화된 정보 제공을 목표로 한 사례에 해당한다. 그중 레그네비게이터는 법령 정보를 기반으로 응답을 생성하는 검색증강생성(Retrieval Augmented Generation, RAG) 방식을 채택하고 있어, 본 연구에서 확인된 범용 AI의 허위 정보 생성(hallucination) 문제를 방지할 수 있는 잠재적 대안으로 검토될 수 있다(황하, 2024). 이와 같은 RAG 기반 특화형 모델은, 특히 사실 기반 정보의 정확성 측면에서 추론 방식에 의존하는 범용 AI보다 우수한 성능을 보일 것으로 예상된다.

다만, 이러한 특화형 시스템은 아직 개발 초기 단계에 있으며, ChatGPT와 같은 범용 AI 서비스와의 성능 차이를 실증적으로 검토한 연구는 아직 부족한 실정이다. 앞으로는 특화형 모델이 실제로 어느 정도의 성능 향상을 제공하는지, 또한 범용 모델에 비해 어떤 강점과 한계를 지니는지를 객관적으로 비교·평가할 필요가 있다. 이를 통해 두 모델이 각각 어떤 유형의 과제에 적합한지를 구체화하고, 역할 분담이나 상호보완적 활용 방안에 대한 논의로 확장할 수 있을 것이다. 예컨대, 성능이 확인된 범용 AI 모델에 RAG 구조를 결합하거나, 목적에 맞게 파인튜닝(fine-tuning)을 적용함으로써 개발 비용을 절감하고 사용자 편의성을 높이는 동시에, 두 접근 방식 간의 상호보완적 발전도 가능할 것이다.

마지막으로 본 연구에서도 확인되었듯이, 정보 결손 상황에서 AI가 임의로 내용을 생성하는 허위 정보 생성(hallucination) 현상은 규제 판단의 신뢰도를 저해하고, 잘못된 분류가 규제 담당자의 책임 문제로 이어질 수 있는 우려를 낳는다. 더욱이 AI 판단의 신뢰도가 낮으면 결과 검토 및 확인 절차가 추가되어 실무자의 업무 부담을 가중하므로, 생성 결과의 신뢰성 확보는 실효성 있는 AI 활용의 전제조건이라 할 수 있다. 따라서 AI 생성 정보의 정확성을 높이기 위한 기술적 개선 노력과 더불어, 이를 실무에 효과적으로 적용하기 위해서는 생성 결과의 신뢰도를 어떻게 정의하고 어느 수준까지 수용할 것인지에 대한 사회적 논의와 기준 설정이 병행되어야 한다. 특히 신뢰도 기준 설정 시에는 절대적인 정확성을 지향하기보다, 사회적으로 수용할 수 있는 신뢰 수준에 대한 현실적 검토가 필요하다. 절대적 신뢰도를 전제로 AI 활용을 미루기보다는, 현재 수준에서 실무자가 수용할 수 있는 적정 신뢰도 하에서의 활용 방안을 모색할 필요가 있다. 이와 함께, AI가 어느 수준까지 판단을 담당하고 사람이 어떤 영역에서 검토와 책임을 부담할 것인지에 대한 명확한 역할 분담 기준 또한 마련되어야 한다. 따라서 생성형 AI의 실무 적용 가능성은 기술적 성능 자체뿐만 아니라, 활용 방식, 신뢰도 수용 기준, 역할 분담 체계를 포괄하는 복합적인 논의 속에서 종합적으로 검토되어야 한다.

참고문헌

고정철. (2024). 생성형 AI와 법률서비스 , 국내외 동향 및 시사점. 이슈와 논점, 제2310호. 국회입법조사처.

국무조정실. (2016). 행정규제 판단기준.

국무조정실. (2024). 규제영향분석서 작성지침.

법제처. (n.d.-a). 국가법령정보센터 소개. https://www.law.go.kr/LSW/lawState.do (검색일: 2025. 2. 26)

법제처. (n.d.-b). 법령통계 홈페이지. https://www.moleg.go.kr/esusr/mpbStaSts/stastsList.es?mid=a10109040100&srch_csf_cd=120001 (검색일: 2025. 2. 24).

양지훈·윤상혁. (2023). ChatGPT를 넘어 생성형(Generative) AI 시대로: 미디어·콘텐츠 생성형 AI 서비스 사례와 경쟁력 확보 방안. Media Issue & Trend Domestic Report, 55: 1-9.

이혁우. (2009). 규제의 개념에 관한 소고. 행정논총, 47(3): 335-358.

정채연. (2024). 생성형 AI를 활용한 법률서비스의 쟁점과 과제. 법학연구, 35(3): 401-443.

조영임. (2023). 초거대 AI와 생성형 인공지능. ICT Standard Weekly, 제1145호.

황하. (2024). 레그네비게이터, 원클릭 규제검색 서비스. 미래정책 FOCUS, 여름호. 경제·인문사회연구회. https://www.nrc.re.kr/board.es?mid=a30400000000&bid=0046&act=view&list_no=178490 (검색일: 2025. 4. 30)

Choi, J. H., Monahan, A. B., & Schwarcz, D. (2024). Lawyering in the Age of Artificial Intelligence. Minnesota Law Review, 109: 147-209. https://minnesotalawreview.org/article/lawyering-in-the-age-of-artificial-intelligence/

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8): 861-874.

Flach, P., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right.

In Advances in neural information processing systems 28 (NIPS 2015). Curran Associates, Inc.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. Advances in Neural Information Processing Systems (NeurIPS), 27. https://arxiv.org/abs/1406.2661

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1): 29-36.

Heaven, W. D. (2023, February 16). ChatGPT is everywhere. Here's where it came from. MIT Technology Review. https://www.technologyreview.com/2023/02/16/1068156/chatgpt-is-everywhere-but-where-did-it-come-from/

Janatian, N., Nikpoor, N., Derakhshani, S., & Rahmati, M. (2023). Legal expert systems using large language models: Capabilities, limitations, and future directions. arXiv preprint arXiv:2311.04911.

Jaskowiak, P. A., Costa, I. G., & Campello, R. J. G. B. (2021). The area under the ROC curve as a measure of clustering quality. arXiv. https://arxiv.org/abs/2009.02400

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1312.6114

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2): 155-163.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74-81). Barcelona, Spain: Association for Computational Linguistics.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information

Retrieval. Cambridge University Press.

Martin, L., Whitehouse, N., Yiu, S., Catterson, L., & Perera, R. (2024). Better Call GPT: Comparing Large Language Models Against Lawyers. Onit AI Center of Excellence. Retrieved from https://arxiv.org/abs/2401.16212

Noguti, A., Gonçalves, D. M. B., & Rosa, J. L. G. (2020). Automatic classification of public prosecution service petitions using Natural Language Processing techniques. arXiv preprint arXiv:2010.12533.

Scacchi, M. (2024, December 20). The CFR: A 190,000-page monument to executive overreach. Pacific Legal Foundation. Retrieved February 24, 2025, from https://pacificlegal.org/the-cfr-a-190000-page-monument-to-executive-overreach/

Shui, R., Cao, Y., Wang, X., & Chua, T.-S. (2023). A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. arXiv:2310.11761.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4): 427-437.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30. https://arxiv.org/abs/1706.03762

Walker II, S. M. (n.d.). F-Score: What are accuracy, precision, recall, and F1 score? Retrieved March 5, 2025, from https://klu.ai/glossary/accuracy-precision-recall-f1

Journal of Regulation Studies Vol. 34 No. 1

Generative AI in Legal Information Retrieval and Regulatory Classification: An Exploratory Study Using ChatGPT

Park, Jung-Won

This study explores the applicability of generative artificial intelligence (Generative AI), particularly ChatGPT, to legal information retrieval and regulatory classification. Two tasks were designed: one involved generating legal provisions based on input information such as statute name, article number, and title; the other involved classifying whether a given provision is regulatory or non-regulatory. The study evaluated classification accuracy, content similarity, execution consistency, and performance variation across conditions. Results showed that in the provision generation task, ChatGPT often produced inaccurate, inconsistent, or hallucinated content, with low similarity to reference texts. In contrast, the classification task achieved a mean accuracy of 66.9%, with high consistency and reliability, suggesting potential utility as a supportive tool for identifying regulatory provisions. However, the frequent misclassification of non-regulatory provisions remains a key limitation. As an initial empirical assessment of generative AI in legal and regulatory contexts, this study provides a foundation for future research comparing domain-specific AI models and exploring their use in

regulatory support tasks.

Keywords: Artificial Intelligence, ChatGPT, Legal Information, Regulatory Classification