지식재산연구
Journal of Intellectual Property

지식재산연구 20권 2호, 2025년 6월, 93-118면
Journal of Intellectual Property Volume 20, Issue 2, June 2025, Pages 93-118.
https://doi.org/10.34122/jip.2025.20.2.93

Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

Generative AI-Based Judicial Ruling Disclosure System Innovation: A Legal and Technological Approach to Balancing Privacy Protection and the Right to Information

Eul Young Lim[1], MinSung Hyun[2]

[1]Master of Engineering; Master of Laws; Manager, KDI School of Public Policy and Management, Republic of Korea
[2]Master of Engineering; Manager, KDI School of Public Policy and Management, Republic of Korea

Corresponding Author: MinSung Hyun (ms_hyun@kdischool.ac.kr)

ABSTRACT

Current Korean court ruling disclosure systems inadequately balance public access and privacy regarding Artificial Intelligence (AI) owing to low disclosure rates and inefficient manual anonymization. This results in hindering AI development as well as the public's right to information. This study proposes and evaluates a generative AI-based automatic anonymization system using fine-tuned Small Language Models (sLLMs, 7-8B scale), designed for secure operations within courts.

Using actual court data, we compared the performance of fine-tuned sLLMs with that of few-shot Large Language Models (LLMs, e.g., GPT-4o). The results demonstrate that fine-tuned sLLMs achieve high accuracy and recall (F1 > 98%), compared to LLMs, proving their feasibility for safe and efficient internal automation without data leakage risks.

Furthermore, we analyzed the legal liability challenges concerning potential AI anonymization errors. Reviewing the applicability of existing doctrines revealed legal uncertainties surrounding AI system failures.

Fine-tuned sLLMs offer a realistic technological solution for harmonizing privacy protection and access rights. This study underscores the need for supportive legal and institutional frameworks to enhance judicial transparency and to advance the legal AI sector.

KEYWORDS

Judicial Ruling Disclosure System, Privacy Protection, Right to Information, Generative Artificial Intelligence, Anonymization, Legal Liability

Open Access

Received: March 06, 2025
Revised: April 08, 2025
Accepted: May 30, 2025
Published: June 30, 2025

Funding: The author received manuscript fees for this article from Korea Institute of Intellectual Property.

Conflict of interest: No potential conflict of interest relevant to this article was reported.

© 2025 Korea Institute of Intellectual Property

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

원저

생성형 인공지능을 활용한 판결문 공개 제도 개선: 개인정보 보호와 알 권리 균형을 위한 법적·기술적 접근

임을영[1], 현민성[2]

[1]KDI국제정책대학원 선임전문원, 공학석사, 법학석사
[2]KDI국제정책대학원 선임전문원, 공학석사

교신저자: 현민성(ms_hyun@kdischool.ac.kr)

차례

1. 서론

2. 판결문 공개와 비식별화의 법적·사회적 맥락
 2.1. 법률에서의 판결문 공개 근거
 2.2. 판결문의 가치와 사회적 역할
 2.3. 판결문 열람·검색 절차 및 한계
 2.4. 판결문 비식별화 방법 및 한계

3. 생성형 AI 기반 판결문 비식별화 시스템 구현 및 성능 평가
 3.1. 생성형 AI의 개념과 판결문 비식별화 적용 가능성
 3.2. 생성형 AI 기반 판결문 비식별화 시스템 설계 및 구현
 3.3. 성능 평가
 3.4. 본 시스템의 법원 활용에 따른 기대효과

4. 본 시스템 활용 시 법적 쟁점
 4.1. 비식별화 오류 시 책임소재: 사람이 실패한 경우
 4.2. 비식별화 오류 시 책임소재: AI가 실패한 경우

5. 결론

국문초록

현행 법원 판결문 공개 제도는 낮은 공개율과 비효율적인 수동 비식별 처리로 인해, 사법 정보 접근성 보장과 개인정보 보호라는 상충하는 가치의 균형 달성이 어렵다. 이는 국민의 알 권리 충족과 법률 연구 및 AI 산업 발전을 저해한다. 본 연구는 이를 해결하고자, 법원에서 안전하게 운영 가능한 생성형 AI 기반 자동 비식별화 시스템의 설계 및 구현 방법론을 제시하고 그 성능을 평가한다.

연구 방법으로 실제 판결문 데이터를 활용하여, 외부 대형 언어 모델(GPT-4o 등)의 Few-shot 방식과 기관 내부 구축에 적합한 소형 언어 모델(7~8B 규모)의 파인튜닝 방식 성능을 비교했다. 실험 결과, 파인튜닝한 소형 언어 모델은 Few-shot 방식의 대형 언어 모델에 준하는 높은 비식별 정확도와 재현율(F1 점수 98% 이상)을 달성하였다. 이는 개인정보 유출 위험 없이 법원 내부에서 안전하고 효율적으로 판결문 비식별화 작업을 자동화할 수 있음을 실증적으로 보여준다.

나아가, 이러한 시스템 도입 시 발생할 수 있는 법적 쟁점, 특히 비식별화 오류 발생 시 책임 소재를 분석했다. 인간의 오류는 물론, AI의 정확도 수준에 따라 기존 과실책임 법리, 국가배상책임, 제조물책임 등의 적용 가능성과 한계를 검토하여, AI 시스템 오류에 대한 법적 불확실성이 존재함을 지적했다.

결론적으로 본 연구는 파인튜닝한 소형 언어 모델이 개인정보 보호와 알 권리 보장을 조화시키는 현실적 기술 대안임을 보여주고, 관련 법·제도적 보완의 필요성을 강조하며, 사법 투명성 제고와 법률 AI 산업 발전에 기여하고자 한다.

주제어

판결문 공개 제도, 개인정보 보호, 알 권리, 생성형 인공지능, 비식별화, 법적 책임

1. 서론

법원 판결문은 사법적 판단의 결과물로서 공적 성격이 강하며, 법적 안정성과 사법기관의 책임성을 강화하는 데 핵심적인 역할을 한다. 판결문 공개는 법 적용 및 해석의 투명성을 제고할 뿐만 아니라, 법학 연구와 교육, 나아가 사회적 신뢰 형성에도 크게 기여할 수 있는 중요한 법적 자료로서의 가치를 지닌다. 특히, 최근 생성형 인공지능[1] 기술의 발전은 판결문 데이터를 새로운 시각에서 활용할 수 있는 가능성을 열어주며, 법학과 기술의 융합 연구에 혁신적인 기회를 제공하고 있다.

그러나 우리나라의 현행 판결문 공개 제도는 구조적 한계를 안고 있다. 민사, 형사, 행정 사건의 공개율은 2019-2022년 4년간 약 45%에 불과하며[2], 판결문 열람 신청 시 발생하는 수수료와 최대 한 달에 이르는 처리 기간은 판결문 접근성을 크게 제한하고 있다.[3] 이러한 문제는 판결문의 공적 가치를 충분히 실현하지 못하게 하는 주요 요인으로 작용한다. 이는 궁극적으로 개인정보 보호와 국민의 알 권리 간의 충돌에서 비롯된다. 사생활 침해의 우려로 인해 사법부는 판결문 공개 이전에 개인정보 비식별화를 필수적으로 진행하고 있으나[4], 이 과정에서 상당한 인적·물적 자원이 소요된다. 결과적으로 공개율 저하와 접근성 악화로 이어지며, 현대 사회에서 개인정보 보호와 정보 접근성 간의 균형을 설정하는 문제는 시급한 사회적 과제로 부각되고 있다.

기존 연구들은 이러한 문제를 다양한 측면에서 분석해 왔다. 김홍화(2017)[5]는 판결문 공개 과정에서 발생할 수 있는 개인정보 침해와 국가배상책임 문제를 논의하며, 비실명화 작업의 한계와 법적 책임의 모호성을 지적했다. 김용주(2024)[6]는 판결문 공개의 제도적 보완 방안을 모색하며, 개인정보 보호와 공개 간의 균형을 맞추기 위한 실질적인 개선책을 제안했다. 최선(2022)[7]은 판결문 공개를 통해 사법부의 책무성을 확보하려는 노력을 분석하며, 공개 확대의 필요성과 한계를 논의했다. 박성미(2021)[8]는 '판결서 인터넷열람 서비스'에서 판결문이 기계판독[9]이 어려운 형태로 제공되어 사법 분야 인공지능 연구에 제약을 초래함을 지적하며, 데이터 활용성을 높이기 위한 공개 방식 개선이 필요함을 강조했다. 임준규(2014)[10]와 전원열(2018)[11]은 온라인상에서의 판결문 공개와 개인정보 보호의 문제를 심층적으로 검토했다.

그러나 기존 연구들은 주로 제도적 개선 방안이나 법적 책임 문제에 초점을 맞추었을 뿐, 기술적 혁신을 통한 문제 해결 가능성에 대해서는 충분히 탐구하지 못했다. 특히, 인공지능 기술

1) 생성형 AI는 딥러닝의 한 분야로, 기존 데이터를 학습한 후 이를 기반으로 새로운 데이터를 '생성'할 수 있는 능력을 가진 모델을 말한다. 예를 들어, 텍스트 생성, 이미지 생성, 음성 합성 등이 생성형 AI의 대표적인 응용 사례에 해당한다.

2) 박수연 외 2인, 「[신년기획][판결문 전면공개를 향하여] ① 국내 판결문, 기본권 침해 우려해 공개 확대 꺼려」, 법률신문, 2024. 1. 1자.

3) 민명기, 「판례를 돈 내고 사는 나라」, 법률신문, 2024. 8. 3자.; 이용경, 「[신년기획][판결문 전면공개를 향하여] ① 기자가 직접 써본 '판결 인터넷 열람 서비스'」, 법률신문, 2024. 1. 1자.

4) 형사소송법 제59조의3, 민사소송법 제163조의2.

5) 김홍화, "판결문 공개시 비실명화작업에 대한 국가배상책임", 「영남법학」, 제44호(2017), 121-140면.

6) 김용주, "재판기록 공개 관련 제도적 보완방안", 「법학연구」, 제65권 제4호(2024), 149-176면.

7) 최선, "판결문 공개제도와 사법부의 책무성", 「헌법학연구」, 제28권 제1호(2022), 421-452면.

8) 박성미, "사법분야 인공지능 발전을 위한 판결문 데이터 개선방안-'판결서 인터넷열람 서비스'를 중심으로-", 「경찰법연구」, 제19권 제3호(2021), 3-36면.

9) 예를들어, 이미지 형태의 스캔 문서나 구조화되지 않은 PDF로 제공되어 자동화된 데이터 처리 및 분석을 어렵게 한다.

10) 임준규, "온라인상에서의 판결문 공개와 개인정보보호의 문제", 「LAW & TECHNOLOGY」, 제10권 제1호(2014), 60-73면.

11) 전원열, "판결 공개에 따른 프라이버시 침해와 민사소송법 제163조의2", 「저스티스」, 통권 제168호(2018), 60-107면.

을 활용한 개인정보 비식별화 시스템의 잠재력과 한계에 대한 연구는 아직 미흡한 실정이다. 이에 본 연구는 현행 판결문 공개 제도의 한계를 분석하고, 생성형 인공지능 기술, 구체적으로는 파인튜닝[12]한 소형 언어 모델(Smaller Large Language Model, 이하 'sLLM'이라고 한다)[13]을 기반으로 한 개인정보 비식별화 시스템의 구현 방법론을 제시하고 그 성능을 실증적으로 검증함으로써 기존 연구의 공백을 메우고자 한다. 본 연구는 다음의 두 가지 핵심 질문을 중심으로 진행된다. 첫째, 현행 판결문 공개 제도의 구조적 문제와 한계는 무엇인가? 둘째, 생성형 인공지능 기술이 개인정보 비식별화의 효율성을 높이고 판결문 공개의 접근성을 향상시키는 데 기여할 수 있는 가능성과 한계는 무엇인가?

본 연구는 이와 같은 질문에 대한 심층적인 논의를 통해 판결문 공개 제도의 개선 가능성을 모색하고, 개인정보 보호와 알 권리 간의 균형점을 찾기 위한 새로운 기술적 접근법과 그 적용 방안을 제시한다. 이를 통해 법적 안정성과 사회적 투명성을 동시에 제고하며, 기술적 혁신을 바탕으로 법학 및 관련 산업 발전에 기여하고자 한다.

2. 판결문 공개와 비식별화의 법적·사회적 맥락

2.1. 법률에서의 판결문 공개 근거

2.1.1. 민사, 형사소송법에서의 판결문 공개조항

민사소송법은 판결문의 공개 범위와 절차를 정함으로써 국민의 알 권리를 보장하고 사법의 투명성을 유지하려는 입법적 의도를 반영하고 있다. 「민사소송법」 제163조의2는 원칙적으로 판결이 선고된 사건(확정되지 않은 사건에 대한 판결서를 포함)의 판결서 공개를 허용하면서도 개인정보 보호를 위해 일정한 제한을 두고 있다. 2015년부터는 확정된 민사소송 판결문의 열람이 가능해졌으며[14], 2023년 1월 1일부터는 2~3심에 계속 중인 민사소송 판결문도 인터넷을 통해 열람할 수 있게 되었다.[15] 또한, 「민사소송규칙」 제76조의2는 법원이 재판업무 수행을 위해 개인정보보호법상 민감정보[16] 및 고유식별정보[17]를 처리할 수 있도록 하면서, 판결문 내 개

12) 파인튜닝은 대형 언어 모델을 특정 작업이나 도메인에 맞게 조정하는 과정으로, 도메인 특화 데이터를 활용해 모델의 성능을 향상한다. 이를 통해 텍스트 번역, 요약, 질의응답 등 다양한 자연어 처리 작업을 지원하며, 적은 양의 데이터로도 특정 산업 분야에 특화된 AI를 효과적으로 구축할 수 있다.

13) 소형 언어 모델(Smaller Large Language Model)은 대형 언어 모델(Large Language Model)의 경량화된 버전으로, 성능을 유지하면서도 리소스 효율성을 극대화한 모델이다. 소형 언어 모델은 수억 개에서 수십억 개의 매개변수를 가지며, 모델의 크기를 줄이면서도 특정 작업에서의 성능을 최적화한다. 소형 언어 모델은 자원이 제한된 환경에서도 실행 가능하며, 에너지 소비를 줄이고 데이터 프라이버시를 보장할 수 있어 엣지 디바이스, 모바일 기기, 로컬 환경에서 활용도가 높다.

14) 구 민사소송법 (2011. 7. 18 법률 제10859호로 개정되기 이전의 것) 제163조의2(확정 판결서의 열람·복사) ① 제162조에도 불구하고 누구든지 판결이 확정된 사건의 판결서(「소액사건심판법」이 적용되는 사건의 판결서와 「상고심절차에 관한 특례법」 제4조 및 이 법 제429조 본문에 따른 판결서는 제외한다)를 인터넷, 그 밖의 전산정보처리시스템을 통한 전자적 방법 등으로 열람 및 복사할 수 있다. 다만, 변론의 공개를 금지한 사건의 판결서로서 대법원규칙으로 정하는 경우에는 열람 및 복사를 전부 또는 일부 제한할 수 있다.

15) 민사소송법 제163조의2(판결서의 열람·복사) ① 제162조에도 불구하고 누구든지 판결이 선고된 사건의 판결서(확정되지 아니한 사건에 대한 판결서를 포함하며, 「소액사건심판법」이 적용되는 사건의 판결서와 「상고심절차에 관한 특례법」 제4조 및 이 법 제429조 본문에 따른 판결서는 제외한다. 이하 이 조에서 같다)를 인터넷, 그 밖의 전산정보처리시스템을 통한 전자적 방법 등으로 열람 및 복사할 수 있다. 다만, 변론의 공개를 금지한 사건의 판결서로서 대법원규칙으로 정하는 경우에는 열람 및 복사를 전부 또는 일부 제한할 수 있다.

16) 「개인정보보호법」 제23조에 따른 민감정보란 사상·신념, 노동조합·정당 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보와 같이 법률에 명시된 정보 또는 그 밖에 정보주체의 사생활을 현저히 침해할 우려가 있어 대통령령으로 정하는 개인정보를 말한다.

17) 「개인정보보호법」 제24조 제1항에 따른 고유식별정보란 법령에 따라 개인을 고유하게 구별하기 위해 부여된 식별정보를 의미한다. 구체적으로 동법 시행령 제19조는 주민등록번호, 여권번호, 운전면허번호, 외국인등록번호를 고유식별정보로 명시하고 있다.

인정보 보호를 위한 조치를 병행하도록 규정하고 있다.[18] 이와 같이, 판결문 공개는 법적 안정성과 판례 연구에 기여하는 한편, 당사자의 개인정보 보호도 함께 고려되고 있음을 보여준다.

반면, 형사판결문 공개는 민사판결문보다 엄격한 제한을 받고 있다. 「형사소송법」 제59조의3은 확정된 형사판결문의 공개를 규정하여 2013년부터 열람이 가능하도록 하였으나, 민사판결문과 달리 확정되지 않은 형사판결문은 열람할 수 없다.[19] 특히, 형사판결문의 경우 공개 제한 사유가 광범위하다. 형사판결문은 민사판결문에서는 고려되지 않는 다양한 이유로 공개가 제한될 수 있다. 「형사소송법」 제59조의3 제1항 단서에 따르면, 심리가 비공개로 진행된 사건, 소년형사사건, 국가안보와 관련된 사건, 그리고 공범관계에 있는 자 등의 증거인멸 또는 도주를 용이하게 하거나 관련 사건의 재판에 중대한 영향을 초래할 우려가 있는 경우는 판결문 공개 대상에서 제외된다. 또한, 소송관계인의 신청이 있는 경우 법원이 공개 여부를 판단하여 열람 및 복사를 제한할 수 있어, 공개 제한 절차가 더욱 강화된다. 이처럼, 형사판결문은 그 공개 대상과 절차 측면에서 민사판결문보다 엄격하게 규제되고 있으며, 이는 형사재판이 가지는 사회적 영향력, 피고인 및 피해자 보호, 국가안보와 공공질서 유지 등 다양한 요소를 반영한 결과로 해석할 수 있다.

2.1.2. 판결문 공개와 개인정보보호법과의 관계

개인정보보호법 제18조 제1항은 개인정보의 목적 외 이용·제공을 제한함을 원칙으로 삼고 있다. 그러나 같은 조 제2항 제8호에서는 예외적으로 “법원의 재판업무 수행을 위하여 필요한 경우”를 허용하고 있다.[20] 이를 근거로 재판업무 수행 과정에서의 개인정보 처리에는 개인정보보호법이 원칙적으로 적용되지 않는다는 해석이 가능하다. 실제로 법원 판결문 작성은 이러한 예외 조항을 근거로 이루어진다.

반면, 판결문을 법원행정처 등에서 전자화하여 대국민 열람 서비스 또는 공개 DB 형태로 제공하는 경우에는 ‘재판업무 수행’이 아닌 법원의 ‘행정사무’로 볼 수 있고[21], 이때에는 개인정보보호법이 적용된다. 이 경우 정보주체의 명시적 동의를 받지 못하는 사례가 대다수를 이루기 때문에, 개인정보를 목적 외로 처리하기 위해서 개인정보보호법의 취지에 따라 별도의 법률적

---

18) 민사소송규칙 제76조의2(민감정보 등의 처리) ① 법원은 재판업무 수행을 위하여 필요한 범위 내에서「개인정보보호법」제23조의 민감정보, 제24조의 고유식별정보, 제24조의2의 주민등록번호 및 그 밖의 개인정보를 처리할 수 있다.
② 법원이 법 제294조 또는 법 제352조에 따라 촉탁을 하는 때에는 필요한 범위 내에서 제1항의 민감정보, 고유식별정보, 주민등록번호 및 그 밖의 개인정보가 포함된 자료의 송부를 요구할 수 있다.
③ 법원사무관등은 소송관계인의 특정을 위한 개인정보를 재판사무시스템을 이용한 전자적인 방법으로 관리한다.
④ 당사자는 법원사무관등에게 서면으로 제3항의 개인정보에 대한 정정을 신청할 수 있다. 그 신청서에는 정정 사유를 소명하는 자료를 붙여야 한다.
⑤ 법원은 재판서가 보존되어 있는 동안 제3항의 개인정보를 보관하여야 한다.

19) 형사소송법 제59조의3(확정 판결서등의 열람·복사) ① 누구든지 판결이 확정된 사건의 판결서 또는 그 등본, 증거목록 또는 그 등본, 그 밖에 검사나 피고인 또는 변호인이 법원에 제출한 서류·물건의 명칭·목록 또는 이에 해당하는 정보(이하 “판결서등”이라 한다)를 보관하는 법원에서 해당 판결서등을 열람 및 복사(인터넷, 그 밖의 전산정보처리시스템을 통한 전자적 방법을 포함한다. 이하 이 조에서 같다)할 수 있다. 다만, 다음 각 호의 어느 하나에 해당하는 경우에는 판결서등의 열람 및 복사를 제한할 수 있다.

20) 개인정보보호법 제18조(개인정보의 목적 외 이용·제공 제한) ① 개인정보처리자는 개인정보를 제15조제1항에 따른 범위를 초과하여 이용하거나 제17조제1항 및 제28조의8제1항에 따른 범위를 초과하여 제3자에게 제공하여서는 아니 된다.
② 제1항에도 불구하고 개인정보처리자는 다음 각 호의 어느 하나에 해당하는 경우에는 정보주체 또는 제3자의 이익을 부당하게 침해할 우려가 있을 때를 제외하고는 개인정보를 목적 외의 용도로 이용하거나 이를 제3자에게 제공할 수 있다. 다만, 제5호부터 제9호까지에 따른 경우는 공공기관의 경우로 한정한다. (중략)
8. 법원의 재판업무 수행을 위하여 필요한 경우.

21) 박우경, “민사소송 및 집행 절차에서의 개인정보 보호에 관한 연구”, 사법정책연구원, 2020, 17면.

근거가 마련되어야 한다. 이에 형사소송법과 민사소송법에서는 이를 위한 별도의 규정을 두고 있다(형사소송법 제59조의3, 민사소송법 제163조의2).[22)]

판결문 공개 시, 개인정보 비실명화 처리는 필수적으로 요구된다. 대법원 규칙인 「형사 판결서 등의 열람 및 복사에 관한 규칙」 제4조와 「민사판결서 열람 및 복사에 관한 규칙」 제4조는 판결문 공개 시 개인정보 비실명화를 의무화하고 있다. 그러나 이러한 조치를 거쳤다 해도, 여전히 해당 정보는 개인정보보호법상 '익명정보(개인정보보호법 제58조의2)'가 아니라 '가명정보(개인정보보호법 제2조 제1호 다목)'에 해당한다. 익명정보란 "시간·비용·기술 등을 합리적으로 고려할 때, 다른 정보를 사용하더라도 더 이상 개인을 알아볼 수 없는 정보"로 정의되지만, 비실명화된 판결문은 원본 판결문과 결합하여 쉽게 복원될 수 있으므로 익명정보로 볼 수 없다. 결국 비실명화된 판결문 역시 가명정보에 해당하므로, 개인정보보호법의 적용 대상이 되며 법원은 개인정보처리자로서 관련 법적 의무를 이행해야 한다.

2.1.3. 판결문 공개와 정보공개법과의 관계

정보공개법은 판결문 공개의 법적 근거를 제공하는 동시에, 특정 정보에 대한 비공개를 허용하는 조항을 포함하고 있어 판결문 공개와 충돌하는 법적 문제를 야기한다. 정보공개법 제2조 제1호는 공공기관이 직무상 작성하거나 취득하여 관리하는 문서를 정보로 정의하고 있으며[23)], 법원도 공공기관의 일종으로서 정보공개 의무를 부담하는 것으로 해석될 수 있다. 이에 따라 법원이 보유한 판결문 역시 정보공개의 대상이 될 수 있으나, 동법 제9조는 예외적으로 비공개가 가능한 정보를 규정하고 있어 법원이 판결문 공개 여부를 결정할 때 중요한 법적 기준이 된다.[24)]

---

22) 이민형, "형사재판절차에서의 개인정보 보호에 관한 연구", 사법정책연구원, 2021, 41-42면.

23) 정보공개법 제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.
1. "정보"란 공공기관이 직무상 작성 또는 취득하여 관리하고 있는 문서(전자문서를 포함한다. 이하 같다) 및 전자매체를 비롯한 모든 형태의 매체 등에 기록된 사항을 말한다.

24) 정보공개법 제9조(비공개 대상 정보) ① 공공기관이 보유·관리하는 정보는 공개 대상이 된다. 다만, 다음 각 호의 어느 하나에 해당하는 정보는 공개하지 아니할 수 있다.
1. 다른 법률 또는 법률에서 위임한 명령(국회규칙·대법원규칙·헌법재판소규칙·중앙선거관리위원회규칙·대통령령 및 조례로 한정한다)에 따라 비밀이나 비공개 사항으로 규정된 정보
2. 국가안전보장·국방·통일·외교관계 등에 관한 사항으로서 공개될 경우 국가의 중대한 이익을 현저히 해칠 우려가 있다고 인정되는 정보
3. 공개될 경우 국민의 생명·신체 및 재산의 보호에 현저한 지장을 초래할 우려가 있다고 인정되는 정보
4. 진행 중인 재판에 관련된 정보와 범죄의 예방, 수사, 공소의 제기 및 유지, 형의 집행, 교정(矯正), 보안처분에 관한 사항으로서 공개될 경우 그 직무수행을 현저히 곤란하게 하거나 형사피고인의 공정한 재판을 받을 권리를 침해한다고 인정할 만한 상당한 이유가 있는 정보
5. 감사·감독·검사·시험·규제·입찰계약·기술개발·인사관리에 관한 사항이나 의사결정 과정 또는 내부검토 과정에 있는 사항 등으로서 공개될 경우 업무의 공정한 수행이나 연구·개발에 현저한 지장을 초래한다고 인정할 만한 상당한 이유가 있는 정보. 다만, 의사결정 과정 또는 내부검토 과정을 이유로 비공개할 경우에는 제13조제5항에 따라 통지를 할 때 의사결정 과정 또는 내부검토 과정의 단계 및 종료 예정일을 함께 안내하여야 하며, 의사결정 과정 및 내부검토 과정이 종료되면 제10조에 따른 청구인에게 이를 통지하여야 한다.
6. 해당 정보에 포함되어 있는 성명·주민등록번호 등 「개인정보보호법」 제2조제1호에 따른 개인정보로서 공개될 경우 사생활의 비밀 또는 자유를 침해할 우려가 있다고 인정되는 정보. 다만, 다음 각 목에 열거한 사항은 제외한다.
가. 법령에서 정하는 바에 따라 열람할 수 있는 정보
나. 공공기관이 공표를 목적으로 작성하거나 취득한 정보로서 사생활의 비밀 또는 자유를 부당하게 침해하지 아니하는 정보
다. 공공기관이 작성하거나 취득한 정보로서 공개하는 것이 공익이나 개인의 권리 구제를 위하여 필요하다고 인정되는 정보
라. 직무를 수행한 공무원의 성명·직위
마. 공개하는 것이 공익을 위하여 필요한 경우로서 법령에 따라 국가 또는 지방자치단체가 업무의 일부를 위탁 또는 위촉한 개인의 성명·직업
7. 법인·단체 또는 개인(이하 "법인등"이라 한다)의 경영상·영업상 비밀에 관한 사항으로서 공개될 경우 법

정보공개법 제9조 제1항 제4호는 진행 중인 재판과 관련된 정보에 대해 공개될 경우 직무수행을 현저히 곤란하게 하거나 형사피고인의 공정한 재판을 받을 권리를 침해할 우려가 있는 경우 비공개할 수 있도록 규정하고 있다. 이는 판결문의 공개가 재판의 공정성 유지와 충돌할 가능성이 있음을 시사하며, 판결문의 공개 여부를 결정할 때 재판의 독립성과 형사절차의 공정성이 고려되어야 함을 의미한다.[25] 또한, 정보공개법 제9조 제1항 제6호 및 제7호는 개인정보 보호 및 영업상 비밀 보호를 이유로 비공개가 가능하도록 규정하고 있다. 이는 판결문 내 개인정보가 포함된 경우, 이를 비공개해야 하는 법적 근거로 작용할 수 있다.

정보공개법과 판결문 공개제도의 관계는 판결문의 성격에 따라 차이가 존재한다. 형사판결문은 형사피고인의 공정한 재판을 받을 권리 보호를 위해 공개가 제한될 가능성이 크며, 정보공개법 제9조 제1항 제4호의 적용을 받을 수 있다.[26] 반면, 민사판결문은 형사판결문과 달리 당사자의 사적 이익이 주요 쟁점이 되므로, 개인정보 보호와 영업상 비밀 보호를 이유로 비공개될 가능성이 상대적으로 높으며, 이에 따라 정보공개법 제9조 제1항 제6호 및 제7호의 적용을 받을 수 있다. 이러한 법적 제한으로 인해 판결문의 공개 범위는 정보공개법의 적용을 받으면서도 개별 판결문의 특성과 법원의 판단에 따라 결정되는 구조를 가진다.

결론적으로, 정보공개법은 판결문 공개의 법적 근거를 제공하면서도, 재판의 공정성 유지 및 개인정보 보호를 이유로 비공개할 수 있도록 규정하고 있어 판결문 공개제도와 상호 긴밀한 관계를 형성하고 있다. 이러한 법적 구조는 판결문 공개의 범위를 결정하는 데 있어 중요한 기준으로 작용하며, 법원이 정보공개법의 적용 범위 내에서 판결문 공개의 한계를 설정하는 과정에서 법적 판단을 요구한다.

2.2. 판결문의 가치와 사회적 역할

판결문은 사법 판단의 최종 산출물로서 그 공개가 사법 투명성과 법적 안정성을 담보하는 핵심적 역할을 수행한다. 재판의 절차와 결과가 국민에게 공개됨으로써 법원의 판단 과정에 대한 객관적 검증이 가능해지고, 사법부의 책무성과 독립성이 강화되는 효과가 있다.[27] 이러한 공개는 법치주의 하에서 국가 권력의 한계를 분명히 하고, 국민의 알 권리를 충족시키며 궁극적으로 사회 전반의 신뢰를 구축하는 데 기여한다.

또한 판결문은 법학 연구와 교육의 기본 자료로서 그 학문적 가치가 높다. 판결문의 구체적 법리 분석과 판례 비교를 통해 법의 해석과 적용에 관한 심도 있는 논의가 가능해지며, 이는 법학자와 실무자 모두에게 중요한 연구 자원으로 활용된다.[28] 나아가 재판 기록에 내포된 법리 및 사실 관계의 상세한 기술은 법률문제에 대한 종합적 이해를 돕고, 법률 서비스의 질적 향상을 위한 기초 자료로서의 역할을 수행한다.

최근 인공지능 및 빅데이터 기술의 발전은 판결문 데이터가 산업적 활용 측면에서도 큰 잠재력을 지니고 있음을 보여준다. 2024년 개최된 오픈데이터포럼 열린세미나(1차)에서, “판결문

인등의 정당한 이익을 현저히 해칠 우려가 있다고 인정되는 정보. 다만, 다음 각 목에 열거한 정보는 제외한다.

가. 사업활동에 의하여 발생하는 위해(危害)로부터 사람의 생명·신체 또는 건강을 보호하기 위하여 공개할 필요가 있는 정보

나. 위법·부당한 사업활동으로부터 국민의 재산 또는 생활을 보호하기 위하여 공개할 필요가 있는 정보.

25) 송오섭, “판결서 열람·검색제도 개선을 위한 몇 가지 고려사항”, 판결문 공개 확대를 위한 국회 토론회 발표자료(2019. 10. 25.), 국회의원 금태섭, 법원행정처, 2019, 28-29면.

26) 송오섭, 위의 글, 28-29면.

27) 김태형, “재판 방송의 가능성과 한계”, 「LAW & TECHNOLOGY」, 제13권 제6호(2017), 48면.

28) 송오섭, 앞의 글, 21면.

데이터야말로 인공지능 시대에 특히 AI 기반 리걸테크 발전에 의미 있는 데이터"라고 평가되며, 양질의 법률 데이터 확보가 기술 혁신의 초석임이 강조되었다.[29] 실제로 판결문을 학습 데이터로 활용한 연구들은, 브라질 연방 법원의 사례에서 인공지능 기반 예측모델이 인간 전문가보다 우수한 성능을 보였음을 입증하고[30], 미국 대법원 판결 예측에서도 고도화된 기계학습 기법이 유의미한 예측력을 나타냄을 보여줌으로써[31] 법률 예측 모델의 정확도 향상과 재판 결과 분석, 나아가 법률 서비스 개선에 실질적인 기여를 할 수 있음을 시사한다.

이와 같이 판결문 공개는 단순한 정보 제공을 넘어 사법 시스템의 감시와 법치주의 확립, 그리고 AI 기반 법률 혁신의 촉매제로서 중요한 사회적 역할을 수행한다. 다만, 개인정보 보호 등 비식별화 문제와의 균형을 이루어야 하는 만큼, 향후 판결문 공개 제도의 개선과 기술적 보완을 통해, 사법의 투명성과 국민 신뢰 제고, 그리고 법률산업 발전에 기여할 수 있는 토대를 마련할 필요가 있다.

2.3. 판결문 열람·검색 절차 및 한계

판결문 공개제도는 기본적으로 인터넷 열람, 판결서 사본 제공 신청, 그리고 판결서 방문열람의 세 가지 절차로 운영된다. 인터넷 열람제도는 법원 홈페이지를 통해 누구나 PDF 파일 형태의 판결문을 열람할 수 있도록 하는 제도로, 당사자 등의 개인정보는 비실명 처리되어 제공된다. 기존에는 2013년 이후 확정된 형사사건 판결문과 2015년 이후 확정된 민사·행정·특허 사건 판결문만이 공개 대상이었으나, 2023년 1월 1일부터 시행되는 개정 민사소송법에 따라 2~3심에 계속 중인 미확정 민사·행정·특허 소송의 판결문도 공개 범위에 포함된다. 다만, 소액사건이나 심리불속행 기각[32] 사건 등은 공개 대상에서 제외된다.[33]

반면, 2013년 1월 1일 이전 형사판결이나 2015년 1월 1일 이전 민사·행정 판결 등, 특정 기간 이전 확정된 판결문이나 상급심 진행 중인 판결문의 경우에는 판결서 사본 제공 신청 제도를 통해 개별 청구 후 법원 담당자가 비실명 처리를 완료한 판결문을 전자우편 등으로 제공하는 절차를 거치게 된다.[34] 한편, 판결서 방문열람제도는 법원 도서관 등 지정된 특별열람실에서 직접 판결문을 열람할 수 있도록 하여, 인터넷 열람제도의 한계를 보완하고자 하나, 이용 대상자 제한[35], 이용 시간 및 좌석 수의 제약 등으로 인해 학술연구나 변론준비 등 대규모 자료 수집에 있어 실질적인 불편함을 초래하는 문제가 있다.[36]

---

29) 오픈데이터포럼, "2024 오픈데이터포럼 열린세미나(1차) 법원 판결문 개방 국내·외 현황 및 향후 정책방향 제언 PART 1 주제발표", YouTube, <https://www.youtube.com/watch?v=ts_7HvdwIJY>, 00:08:29-00:08:35, 작성일: 2024. 7. 25.

30) Elias Jacob de Menezes-Neto & Marco Bruno Miranda Clementino, "Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts", *PLOS ONE*, Vol.17 No.2(2022), e0272287.

31) Daniel Martin Katz et al., "A General Approach for Predicting the Behavior of the Supreme Court of the United States", *PLOS ONE*, Vol. 12 No. 2(2017), e0174698.

32) 대법원이 상고된 사건의 서류를 검토한 결과, 법에서 정한 특정한 상고 이유(예: 중대한 법령 위반, 판례 위반 등)에 해당하지 않는다고 판단하여, 별도의 변론이나 심리 없이 상고를 받아들이지 않고 기각하는 결정이다.

33) 민사소송법 제163조의2 제1항.

34) 이용재, "현행 판결문 공개제도 검토", 판결문 공개 확대를 위한 국회 토론회 발표자료, 2019, 3-4면.

35) 판결문 검색·열람을 위한 특별창구의 설치 및 이용에 관한 내규 제3조(검색·열람 대상자) ① 다음 각 호의 어느 하나에 해당하는 사람은 이 내규에서 정하는 절차에 따라 특별창구에서 판결문을 검색·열람할 수 있다.
1. 검사, 검찰공무원, 변호사, 법무사, 대학교수
2. 국가기관과 지방자치단체 및 그 각 소속 기관의 공무원, 「공공기관의 운영에 관한 법률」 제4조에 따른 공공기관의 임직원으로서 소속 기관장 또는 단체장의 의뢰로 법원도서관장의 승인을 얻은 사람
3. 「신문 등의 진흥에 관한 법률」 또는 「방송법」에 따라 허가·승인을 얻거나 등록을 마친 언론사 소속 기자로서 법원도서관장의 승인을 얻은 사람.

이와 같이 각 절차는 그 나름의 한계를 내포하고 있으며, 특히 인터넷 열람제도의 경우 비실명 처리 작업으로 인한 추가 비용[37]과 시간 지연[38], 제한된 검색 기간[39] 등으로 인해 다년간의 판결문을 한 번에 열람하기 어렵다는 점이 지적된다. 판결서 사본 제공 신청 제도는 이미 비실명 처리가 완료된 판결문임에도 불구하고 개별 신청마다 수수료 부담과 법원 담당자의 확인 절차가 반복되어 효율성이 저해되며[40], 판결서 방문열람제도의 경우 지정된 열람실의 지리적 한계와 엄격한 이용 규정으로 인해 접근성이 떨어진다.[41] 이러한 절차상의 문제들은 개인정보 보호를 위한 비실명 처리 작업의 비효율성에서 기인하는 바가 크며, 판결문 공개 제도의 공적 취지와 국민의 알권리 보장을 실질적으로 달성하는 데 큰 걸림돌로 작용한다.

2.4. 판결문 비식별화 방법 및 한계

현행 판결문 비식별화 제도의 구체적 절차는 법원행정처가 추진 중인 『2025년 판결서 등 비실명화 사업』 제안요청서[42]에 근거하여 살펴볼 수 있다. 판결문의 비식별 처리는 우선 자동 비실명 처리를 수행하는 지능형 기계학습 솔루션으로부터 시작한다. 해당 단계에서는 판결문 내 개인정보 요소가 기계학습 알고리즘에 의해 자동으로 식별·대체되어, 본문에 포함된 개인정보의 일부가 제거된다. 이후 자동 처리된 결과물은 사업자가 수행하는 1차 검수 편집 단계를 거치게 되며, 이 과정에서는 자동 처리 과정에서 발생할 수 있는 오류나 누락 사항을 수정하기 위한 기본적인 인적 검토가 이루어진다. 이어서 동일한 사업자가 2차 검수 편집 확인 절차를 통해 1차 검수 편집 결과의 정확성과 일관성을 재차 검토하며, 이로써 초동 편집 결과의 신뢰성을 보완한다. 1차, 2차 검수는 전수 검수로 진행한다. 추가적으로, 품질 검수 단계에서는 동일한 사업자가 표본 검수를 통해 처리 대상 판결문 중 일부를 선정하여 비실명 처리의 완전성과 품질을 점검하고, 만일 오류가 발견될 경우 전체 모집단에 대한 재작업을 지시하는 방식으로 품질 관리가 이루어진다. 마지막으로, 법원행정처가 직접 수행하는 표본검사를 통해 최종적인 검사 절차가 진행되며, 이를 통하여 비실명 처리 결과의 적정성이 공식적으로 승인된다.

이와 같이 다단계에 걸친 비실명화 절차는 판결문 공개에 앞서 개인정보 보호를 강화하고, 동시에 판결문의 공적 가치를 유지하려는 취지에서 마련되었으나, 그 실행 과정에서는 여러 한계점이 드러난다. 우선, 자동 비실명 처리에 적용되는 지능형 기계학습 솔루션의 완성도가 아직 신뢰할 만한 수준에 미치지 못하는 점은 주요한 문제점으로 지적된다. 실제로 관련 전문가들은 현행 시스템의 자동 처리 정확도가 약 80% 수준에 머물러 있어, 후속 검수 편집 단계에서 상당한 수정 및 보완이 필요하다고 평가하고 있다.[43] 이와 같은 기술적 한계는 비실명 처리의 전반적인 효율성을 저해할 뿐 아니라, 처리 속도의 지연과 비용 부담 증가로도 이어질 가능성이 있다. 또한, 1차 및 2차 검수 편집 과정이 전적으로 인적 자원에 의존하는 점 역시 체계적인 오류 발생과 업무의 비일관성 문제를 내포하고 있다. 더불어, 품질 검수 및 법원행정처의 최종 검사 단계에서 표본검사의 한계로 인해 전체 판결문의 비실명 처리 품질을 완벽하게 보증하기 어려운 현실적인 문제점이 존재한다.

이러한 문제점을 극복하기 위해서는 자동 비식별화 기술의 고도화가 필수적이다. 전문가 인

36) 김용주, 앞의 논문, 163-164면.
37) 민명기, 「판례를 돈 내고 사는 나라」, 법률신문, 2024. 8. 3자.
38) 김용주, 앞의 논문, 165-166면.
39) 김용주, 앞의 논문, 166면.
40) 이용재, 앞의 글, 4면.
41) 김용주, 앞의 논문, 163-164면.
42) 법원행정처, "「2025년 판결서 등 비실명화 사업」 제안요청서", 법원행정처, 2024, 1-40면.
43) 우주성, 「'하급심 판결문' 공개 확대한 法, 국내 'AI 변호사' 출현 앞당기나」, 아주경제, 2023. 2. 22자.

터뷰에 따르면, "현재는 판결문 중 개인정보를 자동으로 비식별화하는 기술 완성도는 아직 80% 수준에 그치고 있다. 향후 95% 이상까지 비식별화가 자동으로 가능해지면 공개될 수 있는 판결문 내용과 범위도 증가할 수 있어 법률 AI 개발도 가속화할 수 있다"[44]고 언급했다.

3. 생성형 AI 기반 판결문 비식별화 시스템 구현 및 성능 평가

3.1. 생성형 AI의 개념과 판결문 비식별화 적용 가능성

인공지능 기술은 초기의 단순한 단어 벡터 표현[45]을 넘어 신경망[46] 기반의 복잡한 언어 모델로 발전해 왔다. 특히 트랜스포머(Transformer)[47] 구조의 도입은 대형 언어 모델의 성능을 획기적으로 끌어올려, 별도의 파인튜닝 없이도 다양한 과업을 수행할 수 있는 이른바 '범용성'을 현실화하고 있다. 최근에는 이러한 대형 언어 모델을 기반으로 한 '파운데이션 모델(foundation model)' 개념이 주목받고 있으며[48], 생성형 AI는 이 파운데이션 모델을 활용해 텍스트·이미지 등 다양한 비정형 데이터를 '생성'하는 단계에까지 이르렀다. 실제로 대형 기술기업과 스타트업을 막론하고 AI 관련 투자 중 상당수가 생성형 AI에 집중되고 있는 점도 이러한 배경을 뒷받침한다.[49]

파운데이션 모델은 방대한 데이터셋을 자가 학습(self-supervision)[50]으로 사전에 학습한 뒤, 상대적으로 적은 양의 추가 데이터와 연산 자원만으로 다양한 용도에 맞춰 적응(fine-tuning)할 수 있다는 점에서 큰 장점을 지닌다. 언어 모델을 완전히 처음부터 학습하는 대신, 이미 구축된 거대한 매개변수(parameter)[51] 중 필요한 부분만 파인튜닝하여 모델 성능을 효율적으로 극대화할 수 있기 때문이다. 이러한 점에서 파운데이션 모델을 어떠한 방식으로 파인튜닝하여 목적별 특화된 결과물을 생성할 것인가는 학계와 산업계에서 중요한 연구 주제

44) 위의 신문기사.

45) 벡터 표현(Vector Representation)은 단어나 문장과 같은 텍스트 데이터를 컴퓨터가 이해하고 처리할 수 있도록 수학적인 형태, 즉 고차원 공간상의 벡터(숫자 배열)로 변환하는 기법을 의미한다. 이 벡터 공간에서는 의미적으로 유사하거나 관련성이 높은 단어들이 서로 가까운 위치에 표현되는 경향이 있다. 예를 들어, '사과'와 '배'는 '자동차'보다 서로 가깝게 위치할 수 있다. 이는 컴퓨터가 단어의 의미나 문맥을 수치적으로 다룰 수 있게 하는 초기 자연어 처리의 핵심 기술 중 하나였다.

46) 신경망(Neural Network), 또는 인공 신경망(Artificial Neural Network, ANN)은 인간 뇌의 신경세포(뉴런) 연결 구조와 정보 처리 방식을 수학적으로 모델링한 계산 시스템이다. 입력층, 은닉층(하나 이상), 출력층으로 구성된 노드(인공 뉴런)들의 연결망을 통해 데이터 속의 복잡한 패턴을 학습하고 인식한다. 현대 인공지능, 특히 딥러닝(Deep Learning)의 핵심 기술로, 본문에서 언급되는 트랜스포머(Transformer) 구조나 대형 언어 모델(LLM) 등 고성능 AI 모델들의 기반을 이룬다.

47) 트랜스포머(Transformer)는 2017년 제안된 신경망 구조로서, 어텐션(attention) 메커니즘을 이용해 입력 전체의 상호 연관성을 효율적으로 파악한다. 이를 통해 이전의 순차적 처리 방식(RNN 등)보다 병렬화가 용이하고, 번역·텍스트 생성 등 다양한 자연어 처리 과업에서 우수한 성능을 보인다.

48) 파운데이션 모델(foundation model)은 대규모의 범용 데이터로 사전 학습되어, 다양한 과업에 폭넓게 적용할 수 있는 인공지능 모델을 의미한다. 예컨대 GPT(Generative Pre-trained Transformer) 계열, BERT(Bidirectional Encoder Representations from Transformer) 등의 모델이 이에 해당하며, 이미 학습된 언어·이미지·음성 이해 능력을 바탕으로 최소한의 추가 학습(파인튜닝)만으로도 여러 작업에서 높은 성능을 보인다.

49) 박찬, 「2024년 생성 AI 투자, 총 79조로 역대 최고 기록」, AI타임스, 2025. 1. 5자.

50) 자가 학습(self-supervision)은 기계학습 모델이 명시적으로 정답이 표시된(labeled) 데이터 없이, 입력 데이터 자체의 구조나 일부 정보를 활용하여 스스로 학습 목표를 설정하고 배우는 방식이다. 예를 들어, 문장에서 다음 단어를 예측하거나, 이미지의 일부를 가리고 원본을 복원하는 등의 작업을 통해 데이터의 내재적 패턴과 특징을 학습한다. 이를 통해 사람의 개입 없이도 방대한 양의 원시 데이터(raw data)로부터 사전 학습(pre-training)을 수행할 수 있다.

51) 매개변수(parameter)는 인공지능 모델이 학습 과정에서 데이터로부터 학습한 지식이나 패턴을 수치적으로 저장하는 내부 변수(가중치 등)를 의미한다. 모델은 이 매개변수들의 값을 조정해 나가면서 특정 작업을 더 잘 수행하도록 최적화된다. 본문에서 언급한 '거대한 매개변수'는 파운데이션 모델이 방대한 데이터를 학습하며 갖게 된 수많은(수억 개에서 수조 개에 이르는) 내부 변수들을 지칭하며, 이는 모델의 복잡성과 성능의 기반이 된다.

로 대두되고 있다.[52] 특히 전체 매개변수 업데이트 없이 '프롬프트(prompt)'[53]나 특정 모듈[54]만을 학습하는 방식을 통해서도 기존과 유사한 수준의 성능을 달성할 수 있음이 확인되면서[55], 법률 분야를 포함한 다양한 영역에서 폭넓게 적용할 수 있는 가능성이 제시되고 있다.

이러한 생성형 AI 기술은 개인정보 보호 문제와도 직결된다. 판결문처럼 비정형 데이터 안에 상당한 민감 정보가 존재할 경우, 개인정보를 식별·가명·익명 처리하는 과정에서 기존 규칙 기반이나 수작업 중심의 비식별화 방식은 여러 한계를 드러내왔다. 이에 주목하여 해외에서는 이미 인공지능 기법을 활용해 대규모 텍스트 내 개인정보를 자동으로 추출·마스킹하려는 연구가 활발히 진행되고 있으며, 예컨대 S Asthana(2025)는 인공지능을 이용하여 비정형 데이터에서 이름, 생년월일, 여권번호 등과 같은 개인정보를 인식하고 마스킹하는 접근법을 제시하였다.[56] 해당 논문에서 생성형 AI를 직접 활용했다는 언급은 없었지만, 비정형 텍스트에서 개인정보를 효율적으로 처리할 수 있음을 시사한다. J Zhou(2025)는 사용자가 대형 언어 모델(Large Language Model, 이하 'LLM'이라 한다) 기반 챗봇을 사용할 때 개인정보를 최소화할 수 있도록, sLLM을 활용한 온디바이스[57] 프라이버시 보호의 가능성을 실험적으로 검증했다.[58] 그러나 소형 언어 모델에 프롬프트 엔지니어링[59]을 통한 비식별화를 수행하는 데 그쳐 비식별 수준이 낮아 개인정보 보호 효과가 제한적이었다는 한계가 있다. 한편 국내에서는 생성형 AI 모델에 대한 파인튜닝을 통해, 데이터 외부 반출 과정에서도 개인정보가 노출되지 않도록 가명처리 절차를 자동화하는 방안이 제안되었다.[60]

이상에서 생성형 AI의 기술적 잠재력과 판결문 비식별화 적용 가능성을 살펴보았다. 선행 연구들은 인공지능을 활용한 비식별화의 가능성을 시사했으나, 개인정보 보호 효과나 실질적인 시스템 구현 측면에서는 보완이 필요한 지점들이 확인되었다. 이러한 논의를 바탕으로, 다음 절에서는 판결문 데이터의 특수성과 개인정보 유출 위험을 고려하여 기관 내부 환경에서 안전하고 효율적으로 운영될 수 있는 생성형 AI 기반 판결문 비식별화 시스템의 구체적인 설계 및 구현 방안을 기술한다.

---

52) Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models", Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, 2021, pp. 1-214.

53) 프롬프트(prompt)는 인공지능 모델, 특히 언어 모델에게 원하는 작업 수행이나 결과물 생성을 지시하거나 유도하기 위해 사용자가 입력하는 텍스트(명령, 질문, 예시 등)이다.

54) 여기서 '특정 모듈'이란 파운데이션 모델의 전체 매개변수(parameter)를 모두 미세조정(fine-tuning)하는 대신, 모델의 일부 구성요소나 모델에 추가된 소규모의 학습 가능한 매개변수 집합만을 학습시켜 특정 작업에 대한 성능을 높이는 기법들을 포괄적으로 지칭한다. 이는 전체 모델을 재학습시키는 것에 비해 계산 자원을 크게 절약하면서도 효과적으로 모델을 특정 목적에 맞게 조정할 수 있게 한다.

55) Tom Brown et al., "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems 33 (NeurIPS 2020), NeurIPS Proceedings, 2020, pp. 40-41.

56) Shubhi Asthana et al., "Adaptive PII Mitigation for Large Language Models", arXiv preprint, <https://arxiv.org/abs/2501.12465>, 작성일: 2025. 1. 21.

57) 온디바이스(On-device)는 인공지능 모델의 연산이나 데이터 처리가 외부 서버나 클라우드를 거치지 않고, 스마트폰, 노트북 등 사용자의 개인 기기 내부에서 직접 이루어지는 방식을 의미한다. 데이터가 기기 외부로 전송되지 않으므로, 사용자의 민감정보나 개인정보 유출 위험을 원천적으로 줄여 프라이버시 보호 수준을 높이는 데 목적을 둔다.

58) Jijie Zhou et al., "Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots", arXiv preprint, <https://arxiv.org/abs/2410.11876>, 작성일: 2024. 10. 10.

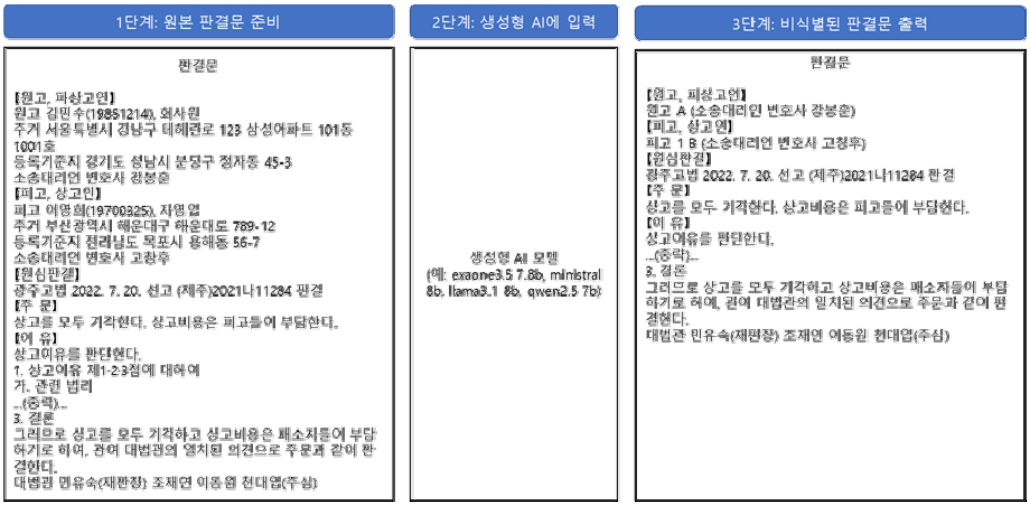
59) 프롬프트 엔지니어링(Prompt Engineering)은 인공지능 모델(특히 LLM)이 사용자의 의도에 맞는 최적의 결과물을 생성하도록 입력하는 지시문, 질문 등의 프롬프트(Prompt)를 효과적으로 설계하고 개선하는 과정을 말한다. 모델의 내부 구조(매개변수)를 직접 수정하는 파인튜닝과 달리, 입력 값 조정을 통해 모델의 반응을 유도하는 기법이다.

60) 임을영, "생성형 AI 활용 단계에서의 개인정보보호법적 쟁점과 기술적 해결방안", 「The Journal of Law & IP」, 제14권 제2호(2024), 187-249면.

3.2. 생성형 AI 기반 판결문 비식별화 시스템 설계 및 구현

판결문 비식별화의 효율성과 정확성을 극대화하기 위해, 본 연구에서는 생성형 AI 모델을 활용한 시스템을 설계하고 구현하였다. 특히, 본 시스템은 외부 LLM 사용으로 인한 개인정보 유출 위험을 차단하고, 기관 내부 환경에서 안전하게 활용할 수 있도록 sLLM을 기반으로 설계하였다.

<그림1 생성형 AI를 활용한 판결문 비식별화 과정>



3.2.1. 시스템 설계

3.2.1.1. 생성형 AI 모델의 역할

본 시스템의 핵심은 생성형 AI 모델이 판결문 내 개인정보를 자동으로 식별하고 마스킹하는 데 있다. 기존의 수동적인 비식별화 작업은 많은 시간과 인력을 요구하며, 오류 발생 가능성 또한 존재한다. 반면, 생성형 AI 모델은 방대한 텍스트 데이터를 학습하여 문맥과 의미를 파악하고, 개인정보에 해당하는 부분을 정확하게 식별하여 자동으로 마스킹 처리를 수행할 수 있다. 이는 비식별화 작업의 효율성을 획기적으로 향상시킬 뿐만 아니라, 휴먼 에러를 최소화하여 비식별화 품질을 높이는 데 기여할 수 있다.

3.2.1.2. 학습 데이터 구성 및 파인튜닝 과정[61]

본 연구는 실제 법원 판결문 데이터를 기반으로 모델 학습 및 성능 검증을 수행하기 위해 다음과 같은 데이터셋을 구축하였다. 우선, 2020년 1월부터 2023년 6월까지 공개된 비식별 판결문 데이터를 수집하였으며, 연구 규모를 고려하여 4,000글자 미만의 판결문 1,748건을 분석 대상으로 선정하였다.

이후, 현재 가장 뛰어난 수준의(State-of-the-Art, 이하 'SOTA'라고 한다) LLM 중 하나인 GPT-4o(2024-08-06 버전)를 활용하여 수집된 비식별 판결문을 기반으로 실제 판결문의 원본 형태를 모사한 합성 데이터[62]를 생성하였다. 이 과정은 모델 학습에 필요한 고품질 데이터를

61) 실험에 사용한 데이터와 모델은 허깅페이스를 통해 공개했다. <https://huggingface.co/ducut91>.
62) 합성 데이터(Synthetic Data)는 실제 세계에서 직접 수집한 원본 데이터가 아니라, 인공지능 모델 등을 이용하여 실제 데이터의 통계적 분포나 특성, 형식 등을 모방하여 인공적으로 생성한 데이터를 의미한다.

확보하기 위한 것으로, 생성된 합성 원본 판결문은 100건의 검증용 데이터셋[63]과 1,648건의 훈련용 데이터셋으로 구분하였다. 특히, 1,648건의 훈련용 데이터셋은 다시 GPT-4o 모델의 Few-shot 방식[64]을 통해 비식별된 판결문을 생성하였다. 그리고 합성 원본 판결문과 비식별된 판결문을 질문/정답 쌍 형태로 구성하여 파인튜닝용 학습 데이터셋으로 활용하였다.

모델 파인튜닝은 7~8B 규모의 sLLM을 대상으로 진행했으며, EXAONE-3.5-7.8B-Instruct, Ministral-8B-Instruct-2410, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct 등 총 4종의 sLLM을 사용하여 실험을 수행하였다. 모델의 파인튜닝은 4 × A100 40GB GPU가 장착된 시스템에서 배치크기[65] 2, 시퀀스 길이[66] 5200, 모델 병렬처리[67] 1, 학습률[68] 5×10^{-5}을 사용하여 3 epoch[69] 동안 훈련했다. 파인튜닝 방식으로는 저순위 적응(Low-Rank Adaptation)[70]을 사용했다.

3.2.1.3. 기관 내부 설치 필요성

판결문은 민감한 개인정보를 포함하고 있으므로, 비식별 시스템은 기관 내부 환경에 설치되어야 한다. 외부 클라우드 기반의 LLM 서비스를 활용하는 경우, 개인정보 유출 문제가 발생할 수 있다.[71] 따라서 본 시스템은 기관 내부 서버에 구축 가능한 sLLM 기반으로 설계하였다. 이는 데이터 보안 및 개인정보보호 측면에서 중요한 의미를 가진다. 기관 내부 설치를 통해 데이터 접근 권한을 통제하고, 외부 네트워크와의 연결을 최소화하여 정보 유출 위험을 원천적으로

63) 평가를 위한 검증용 데이터셋은 판결문 내 정보들을 대법원재판예규인 '판결서 등의 열람 및 복사를 위한 비실명 처리 기준(재일 2014-2)'을 기준에 따라 개인정보인 것 (Positive) 과 개인정보가 아닌 것 (Negative) 으로 구분하여 구성하였다. 여기서 '개인정보인 것'은 이름, 나이, 주소 등 비식별화 대상에 해당하는 정보이며, '개인정보가 아닌 것'은 법관, 검사, 소송대리인이거나 변호인인 변호사 또는 변리사이름, 공공기관 명칭, 시/군/구 단위의 주소 정보 등을 의미한다.

64) Few-shot 방식은 인공지능 모델이 특정 작업을 수행하도록 할 때, 해당 작업과 관련된 대규모 데이터셋으로 모델 전체를 다시 학습시키는 대신, 단지 몇 개(few)의 관련 예시(shot)만을 모델에 입력(주로 프롬프트 형태)으로 제공하여 작업을 수행하게 하는 기법이다. 모델이 사전 학습 단계에서 습득한 방대한 지식과 패턴 이해 능력을 바탕으로, 소수의 예시만 보고도 새로운 작업의 규칙이나 요구사항을 파악하여 처리하도록 유도하는 방식이다. 본문에서는 GPT-4o에 비식별화 작업의 예시를 몇 건 제시하여 전체 훈련용 데이터셋에 대한 비식별화를 수행하도록 한 것을 의미한다.

65) 모델 학습 시 전체 훈련 데이터를 한 번에 처리하기 어려우므로 작은 묶음(배치)으로 나누어 처리하는데, '배치크기'는 이 한 묶음에 들어가는 데이터 샘플의 수를 의미한다. 즉, 모델이 한 단계 학습(파라미터 업데이트)을 진행할 때마다 사용되는 데이터의 양이다. 본문에서는 한 번의 학습 단계마다 2개의 데이터 샘플(판결문 질문/정답 쌍)을 사용했음을 뜻한다.

66) 언어 모델이 한 번에 입력으로 받아 처리할 수 있는 텍스트의 최대 길이를 의미하며, 보통 토큰(token, 텍스트를 의미 단위로 나눈 것)의 개수로 표현된다. 이 길이가 길수록 모델은 더 긴 문맥을 파악할 수 있지만, 계산에 필요한 자원(특히 메모리)이 더 많이 요구된다. 본문에서는 최대 5,200개의 토큰으로 이루어진 텍스트 조각을 모델이 한 번에 처리하도록 설정했음을 의미한다.

67) 인공지능 모델의 크기가 매우 커서 단일 계산 장치(예: GPU)의 메모리에 모두 담을 수 없을 때, 모델 자체를 여러 장치에 분산시켜 동시에 처리하는 기술이다. 이를 통해 매우 큰 모델의 학습 및 추론이 가능해진다. 본문의 '모델 병렬 처리 1'은 모델을 분산시키지 않고 단일 장치 내에서 처리했음을 의미한다.

68) 모델이 학습 과정에서 예측 오류를 바탕으로 내부 파라미터(모델이 학습한 지식)를 얼마나 업데이트할지 그 강도(보폭)를 조절하는 값이다. 학습률이 너무 높으면 최적의 파라미터 값을 지나치기 쉽고, 너무 낮으면 학습 속도가 매우 느려지거나 학습이 제대로 이루어지지 않을 수 있다. 적절한 학습률 설정은 모델 성능에 중요한 영향을 미친다.

69) 전체 훈련(학습) 데이터셋을 처음부터 끝까지 사용하여 모델이 학습을 완료한 횟수를 의미한다. 예를 들어 1 epoch는 전체 데이터를 한 번 모두 학습한 것을 뜻한다. 모델은 보통 여러 번의 에포크를 거치며 데이터의 패턴을 반복적으로 학습하여 성능을 향상시킨다. 본문에서는 전체 훈련 데이터를 총 3번 반복하여 학습에 사용했음을 의미한다.

70) LoRA(Low-Rank Adaptation)는 대규모 신경망의 매개변수를 직접적으로 모두 업데이트하는 대신, 저순위 형태로 표현된 적은 수의 파라미터만을 추가·학습함으로써 모델을 효율적으로 파인튜닝하는 기법이다. 이는 대형 모델에서도 메모리 사용량과 연산량을 줄여, 상대적으로 적은 자원으로도 빠른 학습이 가능하게 한다.

71) 법원 판결문을 비식별하는 과정에서 외부 클라우드 기반의 LLM을 활용하는 경우, 개인정보보호법 제17조(제3자 제공) 및 제28조의8부터 제28조의11(국외 이전) 규정에 따른 요건을 충족하지 못할 위험이 존재한다. 임을영, 앞의 논문, 214-221면.

차단할 수 있다.

3.3. 성능 평가

본 연구에서 구현한 시스템의 성능을 객관적으로 평가하기 위해 정확도(Accuracy)[72], 재현율(Recall)[73], 그리고 F1 점수(F1 Score)[74] 지표를 활용하였다. 성능 평가는 SOTA급 LLM의 Few-shot 방식과 파인튜닝한 sLLM을 비교·분석하는 방식으로 진행하였으며, 사람이 직접 라벨링한 100건의 검증용 데이터셋을 기반으로 평가를 수행하였다.

3.3.1. Few-shot 방식 성능 결과

<표1>은 SOTA급 LLM을 Few-shot 방식을 통해 법원 판결문을 비식별화한 결과를 보여준다. Gemini-2.0-Flash 모델이 99.38%의 가장 높은 정확도를 기록했으며, DeepSeek-V3 모델이 98.81%의 가장 높은 F1 점수를 나타냈다. GPT-4o(2024-08-06) 모델 또한 98.74%의 높은 F1 점수를 기록하며 우수한 성능을 입증하였다. 전반적으로 최신 LLM은 Few-shot 방식만으로도 높은 수준의 판결문 비식별화 성능을 보이는 것을 확인할 수 있다.

<표1 SOTA급 LLM을 활용한 법원 판결문 개인정보 비식별 성능 비교 (Few-shot 방식)>

모델	정확도	재현율	F1 점수
GPT-4o(2024-08-06)	97.82	99.66	98.74
Qwen2.5-Max	96.46	95.83	96.14
DeepSeek-V3	98.73	98.92	98.81
Gemini-2.0-Flash	99.38	95.78	97.55

3.3.2. 파인튜닝 방식 성능 결과

<표2>는 7~8B급 sLLM의 파인튜닝 전후 법원 판결문 비식별화 성능을 보여준다. 파인튜닝 전, Qwen2.5-7B-Instruct 모델이 93.74%의 비교적 높은 정확도를 보였으나, 전반적으로 F1 점수는 70% 내외로 낮은 성능을 나타냈다.[75] 특히 Ministral-8B-Instruct-2410 모델은 재현율과 F1 점수가 매우 낮아 판결문 비식별화에 적합하지 않음을 시사한다.

그러나 sLLM은 파인튜닝을 통해 성능이 현저하게 향상되었음을 확인할 수 있다. 특히 Ministral-8B-Instruct-2410 모델은 파인튜닝 후 98.70%의 매우 높은 F1 점수를 기록했으며, EXAONE-3.5-7.8B-Instruct, Llama-3.1-8B-Instruct 모델 또한 95% 이상의 높은 F1

72) 전체 데이터 중에서 모델이 정답을 맞춘 예측의 비율을 의미한다. 실제 '개인정보'를 '개인정보'라고 예측한 경우와 '개인정보 아님'을 '개인정보 아님'으로 예측한 경우의 합한 비율이다. 모델의 전반적인 예측 성능을 나타낸다. 전체적인 예측 성공률을 보여주지만, '개인정보'와 '개인정보 아님'의 비율이 불균형한 경우에는 이 지표만으로 성능을 온전히 평가하기 어려울 수 있다.

73) 재현율(Recall)은 실제 '개인정보'인 데이터 중에서 모델이 '개인정보'라고 올바르게 예측해낸 비율을 의미한다. 즉, 모델이 찾아야 할 '개인정보'를 얼마나 빠짐없이 찾아냈는지를 측정한다. 민감 정보 탐지나 질병 진단과 같이, 실제 '참'인 경우를 놓쳤을 때(False Negative)의 위험이 큰 경우에 특히 중요하게 고려되는 지표이다.

74) F1 점수(F1 Score)는 '정밀도(Precision)'와 '재현율(Recall)'이라는 두 가지 지표의 조화 평균(harmonic mean)을 계산한 값이다. 정밀도는 모델이 '참(Positive)'이라고 예측한 것 중에서 실제 정답도 '참'인 비율을 의미한다. F1 점수는 정밀도와 재현율을 종합적으로 고려하여 모델의 성능을 평가하며, 정밀도와 재현율이 모두 중요하거나 데이터가 불균형할 때 유용하게 사용된다. 정밀도와 재현율 값이 모두 높을수록 F1 점수도 높아진다.

75) 파인튜닝 전, sLLM의 성능은 Few-shot 방식을 통해 측정했다.

점수를 달성하였다. Qwen2.5-7B-Instruct 모델은 상대적으로 낮은 성능 향상을 보였으나, 파인튜닝 후에도 90% 이상의 F1 점수를 유지하며 안정적인 성능을 나타냈다.

<표2 7~8B급 sLLM의 파인튜닝 전후 법원 판결문 개인정보 비식별 성능 비교>

모델	파인튜닝 전			파인튜닝 후		
	정확도	재현율	F1 점수	정확도	재현율	F1 점수
EXAONE-3.5-7.8B-Instruct	68.26	67.89	68.08	98.59	94.48	96.49
Ministral-8B-Instruct-2410	35.6	4.33	7.72	99.07	98.32	98.70
Qwen2.5-7B-Instruct	93.74	66.49	77.84	94.67	89.39	91.95
Llama-3.1-8B-Instruct	68.26	67.89	68.08	97.19	93.50	95.31

3.3.3. Few-shot 방식 vs 파인튜닝 방식 비교

Few-shot 방식은 SOTA급 LLM을 활용하여 즉각적으로 높은 수준의 비식별화 성능을 얻을 수 있다는 장점이 있지만, 외부 모델 사용으로 인한 보안 및 개인정보 유출 우려가 존재한다. 반면, 파인튜닝 방식은 sLLM을 활용하여 기관 내부 환경에서 안전하게 비식별화 시스템을 구축할 수 있으며, SOTA급 LLM의 Few-shot 방식에 버금가는 높은 성능을 확보할 수 있음을 실험적으로 입증하였다. 특히 Ministral-8B-Instruct-2410 모델의 경우, 파인튜닝을 통해 Few-shot 방식의 최고 성능 모델인 DeepSeek-V3에 근접하는 F1 점수를 달성하며, 파인튜닝 방식의 잠재력을 보여주었다.

3.4. 본 시스템의 법원 활용에 따른 기대효과

본 시스템은 현재 법원에서 판결문 공개 절차에 있어 불가피하게 수반되는 시간적 지연 및 인력 소모의 문제를 획기적으로 개선할 수 있는 잠재력을 지닌다. 현행 법원의 판결문 공개 과정은 1차 전수 검수, 2차 전수 검수, 3차 표본 검수, 그리고 최종 표본검사까지 총 4단계의 복잡한 단계를 거치며, 각 단계마다 상당한 인력과 시간이 투입되어 판결문 공개가 지연되는 주요 원인으로 작용하고 있다. 이러한 상황에서 본 연구에서 구현한 생성형 AI 기반 비식별화 시스템의 도입은 판결문 공개 절차의 효율성을 극대화하고, 궁극적으로 국민의 알 권리 신장에 기여할 수 있는 중요한 전환점을 마련할 수 있을 것으로 기대한다.

본 시스템은 판결문 내 개인정보 식별 및 마스킹 작업을 자동화함으로써, 기존의 수작업 중심 검수 단계를 상당 부분 축소할 수 있다. 특히 1차 및 2차 전수 검수 단계는 본 시스템의 자동화된 비식별 기능으로 대체하거나 그 단계를 현저히 간소화하는 것이 가능하다. 이는 검수 과정에 투입되는 인력과 시간을 대폭 절감함은 물론, 인적 오류 발생 가능성을 최소화하여 비식별 작업의 정확성 또한 향상시킬 수 있는 효과를 가져올 수 있다. 더욱이 시스템의 비식별 정확도가 충분히 높은 수준으로 확보된다면, 3차 표본 검수 및 4차 품질 검사의 범위를 축소하거나, 경우에 따라 단계 자체를 생략하는 방안까지 고려해볼 수 있을 것이다. 이를 통해 판결문 공개 절차를 획기적으로 간소화하고, 공개까지 소요되는 시간을 단축시켜 국민들이 보다 신속하게 사법 정보에 접근할 수 있도록 기여할 수 있을 것이라 기대한다.

4. 본 시스템 활용 시 법적 쟁점

본 시스템이 비식별화에 실패했을 경우 발생할 수 있는 법적 책임을 고찰하고자 한다. 이를

위해 비식별화 오류를 인간이 범한 경우와 AI가 범한 경우로 구분하고, 각각의 책임 귀속과 법적 분쟁 발생 가능성을 살펴본다.

4.1. 비식별화 오류 시 책임소재: 사람이 실패한 경우

판결문 비식별화 작업은 담당 공무원이 개인정보를 식별하고, 이를 다른 문자나 기호로 대체하거나 삭제하는 과정을 통해 이루어진다. 그러나 이 과정에서 담당 공무원이 충분한 주의의무를 다하지 못하면, 개인정보가 유출되어 손해배상책임이 발생할 여지가 있다. 이에, 사람이 직접 진행하는 비식별화 과정에서 오류가 발생했을 때, 그 책임이 누구에게 어떻게 귀속되는지 검토해 본다.

4.1.1. 개인정보보호법상 손해배상청구

개인정보보호법 제39조 제1항은 "정보주체가 개인정보처리자의 법령 위반 행위로 인해 손해를 입은 경우, 해당 개인정보처리자에게 손해배상을 청구할 수 있으며, 이 경우 개인정보처리자는 고의 또는 과실이 없음을 입증하지 않으면 책임을 면할 수 없다"고 명시하고 있다. 즉, 이 규정은 과실책임을 전제로 개인정보처리자의 책임을 인정하는 한편, 입증책임을 개인정보처리자에게 전가함으로써 정보주체의 권익 보호를 극대화하려는 취지를 담고 있다.

4.1.2. 개인정보 자기결정권

헌법 제10조는 "모든 국민은 인간으로서의 존엄과 가치를 가지며, 행복을 추구할 권리를 가진다"라고, 제17조는 "모든 국민은 사생활의 비밀과 자유를 침해받지 아니한다"라고 규정하고 있다. 이는 개인 사생활이 외부로부터 침해되거나 함부로 공개되지 않을 소극적 권리에 더해, 고도로 정보화된 현대사회에서 자기 정보를 자율적으로 통제할 수 있는 적극적 권리도 함께 보장한다.[76] 즉, 헌법 제10조에서 도출되는 일반적 인격권과 헌법 제17조의 사생활의 비밀과 자유가 결합된 '개인정보자기결정권'은 "자신에 관한 정보를 언제, 누구에게, 어느 범위까지 공개하고 이용하게 할지 스스로 결정할 수 있는 권리"이다. 이때 개인정보의 범위는 개인의 신체·신념·사회적 지위·신분 등 개인의 동일성을 식별할 수 있는 일체의 정보로서, 반드시 사적인 영역의 정보에 한정되지 않고 공적 생활에서 형성되거나 이미 공개된 정보까지 포괄한다. 나아가 이러한 개인정보를 대상으로 한 조사·수집·보관·처리·이용 행위는 모두 원칙적으로 개인정보자기결정권에 대한 제한에 해당한다.[77]

4.1.3. 국가배상법상의 책임

판결문 비식별화 과정에서 발생하는 오류로 인해 개인정보가 부적절하게 공개될 경우, 국가배상법 제2조 제1항에 따른 국가의 배상책임이 문제 될 수 있다. 해당 조항은 공무원이 직무를 집행하면서 고의 또는 과실로 법령을 위반하여 타인에게 손해를 입힌 경우 국가나 지방자치단체가 배상책임을 진다고 규정한다. 판례는 여기서 '법령 위반'을 형식적 의미의 법령뿐만 아니라, 인권존중·권력남용금지·신의성실과 같이 공무원으로서 마땅히 지켜야 할 준칙이나 규범을 위반하는 것까지 포함하여 폭넓게 해석하고 있으며[78], 직무상 주의의무 위반 역시 이에 해당할

76) 대법원 1998. 7. 24. 선고 96다42789 판결 참조.
77) 헌법재판소 2005. 7. 21. 선고 2003헌마282,425(병합) 결정 등 참조.
78) 대법원 2015. 8. 27. 선고 2012다204587 판결 등 참조.

수 있다. 따라서 비식별화 업무를 담당하는 법원 공무원이 요구되는 주의의무를 다하지 못하여 개인정보 침해 손해가 발생하였다면, 원칙적으로 국가배상책임이 성립할 가능성이 있다.

그러나 판결서 공개 및 비식별화와 관련하여 민사소송법 제163조의2 제4항 및 형사소송법 제59조의3 제3항은 중요한 예외를 설정하고 있다. 두 조항 모두 "개인정보 보호조치를 한 법원사무관등이나 그 밖의 법원공무원은 고의 또는 중대한 과실로 인한 것이 아니면 제1항에 따른 열람 및 복사와 관련하여 민사상·형사상 책임을 지지 아니한다"고 규정하여, 담당 공무원의 면책 범위를 명시하고 있다. 이러한 면책 규정은 담당 공무원의 단순 과실에 대해서는 개인적으로 책임을 부담하지 않도록 보호하는 의미를 갖는다. 다만, 공무원 개인의 책임을 제한하는 면책 규정이 국가배상법 제2조 제1항에 따른 국가 자체의 배상책임까지 당연히 소멸시키는 것은 아니다. 국가배상책임은 공무원의 '과실'(경과실 포함)만으로도 성립 요건을 충족할 수 있으므로, 담당 공무원에게 경과실만 인정되는 상황에서도 피해자는 여전히 국가를 상대로 배상을 청구할 수 있다.

이러한 면책 조항들은 2011년 7월 18일, 민사 및 형사 판결서에 대한 국민의 접근성을 강화하고 재판의 투명성을 제고하기 위한 법 개정 과정에서 함께 도입되었다.[79] 입법자는 판결서 공개 확대라는 공익적 목적을 달성하는 과정에서 발생할 수 있는 비식별화 오류의 위험성을 인지하고, 이로 인해 담당 공무원이 과도한 책임 부담을 지게 될 경우 판결서 공개 업무 자체가 위축될 수 있음을 우려한 것으로 보인다. 즉, 판결서 공개 활성화와 개인정보 보호라는 두 가치 사이의 균형점을 찾기 위한 정책적 고려가 반영된 결과로 해석할 수 있다. 따라서 비식별화 오류로 인한 손해배상 청구 시, 법원은 담당 공무원에게 고의 또는 '중대한 과실'이 있었는지 여부를 일반적인 국가배상 사건보다 더욱 엄격하게 심리할 것으로 예상된다.

4.1.4. 개인정보 비식별 실패를 다룬 판례

이러한 관련 법리를 전제로, 사람이 담당하는 비식별화 업무가 부실하게 이루어졌을 때의 책임 소재에 관한 판례를 살펴본다.[80] 먼저 살펴볼 판례는 불법행위책임을 인정한 사안이다.[81] 원고 M과 S는, 이혼 소송 과정에서 부모의 불화가 심각해지자, 서울가정법원에 이혼 및 양육자 지정청구가 제기되어 항소심까지 진행된 뒤 확정된 사건의 자녀들이었다. 문제는 이 판결문이 공개되는 과정에서 부모(청구인과 피청구인) 부분만 비실명 처리되었을 뿐, 사건본인에 해당하는 M과 S의 이름과 생년월일은 그대로 노출되어 있었다는 점이었다. 판결문을 검색해 복사해 간 다수의 인터넷 이용자들이 해당 내용을 여러 홈페이지와 블로그 등에 다시 게시함으로써 이들의 개인정보가 사실상 통제 불가능한 상태에 이르렀고, 법원행정처는 뒤늦게 이 사건 판결문을 비실명 처리했으나 이미 광범위한 유출이 이루어진 뒤였다.

법원은 가사사건과 같은 민감한 사안에서 개인정보처리자가 컴퓨터를 통해 개인정보를 취급할 경우, 불특정 다수가 이용하는 인터넷 사이트에 정보를 게시하기 전에 안전성 확보 조치를 충실히 이행할 주의의무가 있다고 보았다. 그러나 법원행정처는 이를 소홀히 하여 원고들의 이름과 생년월일을 그대로 노출했으므로, 직무집행상 과실로 법령을 위반한 잘못이 있다고 판시하였다. 나아가 원고들의 개인정보는 인터넷에 노출된 이상 언제든지 복제되거나 재게시될 위험이 상존하고, 생년월일 등은 다른 개인정보까지 파악하는 연결고리가 될 수 있으므로 개인정

79) 2011년 7월 18일 법률 제10859호로 일부개정되어 2015년 1월 1일 시행된 「민사소송법」 개정이유 및 2011년 7월 18일 법률 제10864호로 일부개정되어 2012년 1월 1일 시행된 「형사소송법」 개정이유.

80) 김홍화의 논문에서 소개한 법원판결문 비식별 실패에 따른 불법행위책임을 인정한 판례와 부정한 판례를 중심으로 비교·분석한다. 김홍화, 앞의 논문, 121-140면.

81) 서울중앙지방법원 2017. 2. 14. 선고 2015가단5072371.

보자기결정권 침해가 충분히 인정된다고 판단하였다. 법원은 이러한 행위가 위법하다고 보고, 국가가 원고들에게 위자료를 지급해야 한다고 보았다. 이는 비식별화 작업의 한계를 인정하면서도, 담당 기관이 주의의무를 다하지 않고 개인정보를 유출해 불특정 다수에게 손쉽게 열람·복제가 가능하게 만들었다면 불법행위책임이 성립한다는 점을 분명히 한 것이다.

반면 불법행위책임을 부정한 판례[82]에서는 유족연금 승계와 관련된 행정소송 판결문이 문제가 되었다. 이 판결문 역시 인터넷에 공개되어 있었고, 판결에서 설시된 가족사는 아버지와 사실혼 관계에 있던 어머니 사이에 관한 민감한 사정이 포함되어 있었다. 판결문 공개 시 대부분의 이름과 인적사항은 A, B, C 등으로 비실명 처리되었으나, 그중 증거거시 내용에서 증인으로 출석한 원고의 이름이 실명으로 나타나 있었다. 원고는 이로 인해 자신의 가족사까지 파악될 수 있다고 주장하며 개인정보자기결정권 침해와 손해배상을 청구하였으나, 법원은 원고의 주장을 받아들이지 않았다. 재판부는 가족사와 관련된 핵심 부분이 모두 비실명 처리되어 있다는 점, 원고의 실명이 오로지 증인 목록에서만 언급되어 있을 뿐 '원고가 가족사 중 특정 인물과 동일하다'는 사실을 일반 독자가 쉽게 연관 지을 수 없다는 점을 근거로 피해자가 특정되지 않는다고 보았다. 결국 원고가 속한 가족사를 구체적으로 파악하기 위해서는 이미 원고의 배경을 자세히 알고 있는 사람이 아니고서는 사실상 이를 연결하기 어렵다는 이유로 불법행위책임을 인정하지 않았다.

이 두 판결은 모두 판결문 공개 과정에서 중요한 개인정보가 노출될 위험이 있다는 공통점을 지니지만, 불법행위책임 인정 여부는 피해자 특정성이 얼마나 명백하게 드러나는가에 따라 크게 달라졌다. 첫 번째 판례에서는 자녀들의 이름과 생년월일이 판결문에 그대로 남아 있어, 일반적인 인터넷 이용자라도 비교적 손쉽게 그 신원을 파악할 수 있었다고 평가되었다. 반면 두 번째 판례에서는 원고의 이름이 일부 노출되었음에도 판결문의 중심 내용인 가족사와 뚜렷이 연결되지 않는다는 이유로 피해자 특정성이 부정되었다. 그러나 김홍화는 두 번째 사안을 두고, 적은 범위라 하더라도 지인이나 제3자를 통해 피해자 특정이 가능하였을 수 있으므로 재판부가 피해자 특정성 판단에서 지나치게 제한적으로 접근했다는 견해를 피력한다.[83] 즉, 지인이거나 제3자라도 추가적인 정보를 갖고 있다면 노출 범위가 의외로 넓어질 수 있는데, 법원이 특정성 인정의 문턱을 지나치게 높이는 태도는 실제 피해를 간과하게 만든다는 것이다.

아울러 이러한 사안에서 국가배상책임이 성립하기 위해서는 공무원의 직무상 과실과 손해 발생 사이에 상당인과관계가 인정되어야 한다. 특히 공무원 개인의 불법행위책임을 인정하려면 중과실 수준이 요구된다는 것이 판례의 지배적 견해인데[84], 실제 비식별화 작업 현장에서 이런 요건을 충족하는 사례는 극히 드물 것으로 예상된다.[85] 다만, 공무원 개인의 중과실이 부정되더라도 국가배상법에 따라 국가가 배상책임을 부담할 수 있고, 이는 판결문 공개 제도 전반에 커다란 부담으로 작용하게 된다.[86]

4.2. 비식별화 오류 시 책임소재: AI가 실패한 경우

다음으로 비식별화 과정에서 오류가 발생하였을 때 AI와 관련된 책임을 어떻게 판단할 것인

82) 서울중앙지방법원 2008. 11. 21. 선고 2008가소36017.

83) 김홍화는 "사안상 충분히 피해자가 특정되었다고 볼만 하고 행정법원 공보판사의 중과실은 없더라도 경과실은 법원정보공개규칙 시행 내규 및 전자우편 등을 통한 판결문 제공에 관한 예규 등을 위반한 과실이 있어 공무원인 공보판사는 손해배상책임을 부담하지 않으나 대한민국은 손해배상 책임을 져야 했다는 견해가 있으며 이에 전적으로 동의한다."고 했다. 김홍화, 앞의 논문, 128면.

84) 대법원 2003. 12. 26. 선고 2003다13307 판결, 대법원 1996. 2. 15. 선고 95다38677 전원합의체 판결.

85) 김홍화, 앞의 논문, 127면.

86) 김홍화, 위의 논문, 127면.

지 살펴볼 필요가 있다. 특히 인공지능 프로그램이 일정 수준 이상으로 발전함에 따라 법적 책임 구조가 어떻게 달라질 수 있는지를, 구체적인 정확도 수준에 따라 나누어 검토해 보도록 한다.

4.2.1. 판결문 비식별 프로그램이 95% 정도의 정확도를 내는 경우

약 95%의 정확도를 지닌 판결문 비식별 프로그램을 가정하면, 이는 인간 작업을 보조하는 수준의, 이른바 '약한 인공지능(weak AI)'에 해당한다. 이러한 AI는 자동화 기능을 제공하나, 결과물의 정확성과 완전성에 대한 최종 검증 책임은 여전히 인간에게 귀속된다. AI가 일정 수준의 자율성을 보인다고 하더라도, 인간의 적극적인 관여와 감독은 필수적이다.

이러한 전제하에 AI의 내재적 한계나 오작동으로 개인정보 침해가 발생한 경우, 현행 법령은 우선적으로 개인정보처리자 또는 이를 운영·감독하는 주체의 과실 여부를 판단한다. 개인정보보호법은 개인정보처리자에게 적정한 보호의무를 부과하며, 정보주체가 손해를 입은 경우 개인정보처리자가 고의 또는 과실이 없음을 입증하지 않는 한 책임을 면할 수 없도록 입증책임을 전환하고 있다(개인정보보호법 제39조 제1항).

이와 같은 법적 구조 아래에서 공무원은 직무수행 시 사용하는 도구의 적정성과 운용 전반에 관해 선량한 관리자로서 주의의무를 부담한다.[87] AI의 보조적 성격과 한계(예: 95% 정확도)를 충분히 인식하고 이에 상응하는 관리적·기술적 검증 절차를 마련·준수하는 것이 핵심적인 주의의무라 할 수 있다.

만약 담당자가 AI 처리결과를 맹신하여 추가 확인 절차를 생략하거나, 5% 내외의 오류 가능성을 충분히 인지하지 못한 채 합리적 수준의 주의의무를 이행하지 않았다면, 과실책임이 인정될 가능성이 높다. 더 나아가 비식별화 성능이 현저히 낮거나 안전성이 검증되지 않은 AI 시스템을 충분한 검토 없이 성급하게 도입·운영한 경우에는, 기술 선택·도입 단계에서부터 주의의무 위반이 인정될 수 있다.[88] 운영 단계에서도 비식별화 결과가 완벽하지 않다는 예견 가능한 위험에 대비해 적절한 수준의 인간 검토 및 감독 절차를 마련하는 것은 필수적이다.[89] 만약 이러한 최소한의 검증 체계조차 갖추지 않거나, 특정 유형의 개인정보(예: 희귀 정보, 문맥 의존적 정보) 식별에 AI가 명백한 취약점을 보인다는 사실이 인지되었음에도 별다른 보완 조치 없이 시스템 운영을 강행하여 개인정보 침해를 야기했다면, 이는 단순한 기술적 한계를 넘어선 중대한 관리·감독상의 과실로 평가될 여지가 크다.

AI가 비식별화 업무의 효율성을 높이는 유용한 도구임은 분명하지만, 최종적인 판단과 책임은 여전히 해당 직무를 수행하는 공무원에게 귀속된다는 점은 변함이 없다. 따라서 AI 결과를 비판적 검토 없이 기계적으로 수용하여 명백하거나 의심할 수 있는 오류를 간과한 경우, 이는 공무원의 기본적 주의의무 해태로서 중대한 과실에 해당할 수 있다. 실제로 인공지능의 민사책임에 관한 기존 연구에서도 "약한 인공지능의 경우라면 그 작동에 인간의 과실이 상대적으로 쉽게 인정될 수도 있다"고 분석하고 있으며[90], 이는 궁극적으로 주의의무 위반이라는 전통적 과실론과 자연스럽게 연결된다.

나아가, 국가기관이 판결문 공개 업무에서 이와 같은 과실을 범한 경우 국가배상책임(국가배

87) 국가공무원법 제56조(성실 의무) 모든 공무원은 법령을 준수하며 성실히 직무를 수행하여야 한다.
88) 개인정보보호위원회, "비정형데이터 가명처리 기준 주요 내용", 개인정보보호위원회, 2024, 3면, 동 자료는 "가명처리 기술의 한계 등을 보완하기 위해 사전 준비단계(연구 및 기술개발 기획)부터 위험성을 충실히 검토하고 적절한 안전조치를 수행"해야 한다고 명시하고 있다.
89) 개인정보보호위원회, 위의 자료, 12면, 동 자료는 "비정형데이터 가명처리의 기술적 한계를 보완하고 잔존위험(Residual Risk)을 낮추기 위해, 처리결과에 대한 자체적인 추가검수 등이 필요"하다고 명시하고 있다.
90) 이수경, "인공지능의 민사책임에 대한 소고", 「민사법학」, 제107호(2024), 228면.

상법 제2조 제1항)이 인정될 수 있으며, 담당 공무원에게 중대한 과실이 인정될 때는 국가 또는 지방자치단체가 구상권(국가배상법 제2조 제2항)을 행사할 수 있다. 과거 판례에서도 수기 비식별 처리 과정에서 누락이 발생한 경우 국가배상책임이 인정된 바 있다.[91] 따라서 AI 활용 시에도 비식별화 한계를 보완하기 위한 관리·감독 의무 위반이 확인된다면, 기존 법리에 따라 동일하게 국가배상책임이 인정될 수 있을 것이다.

4.2.2. 판결문 비식별 프로그램이 99.9% 이상의 정확도를 내는 경우

그렇다면 정확도가 99.9% 이상으로 향상되어 사실상 인간보다 더 높은 비식별 정확도를 지닌 고성능 프로그램 수준에 이른 경우는 어떻게 볼 수 있을까. 이는 운전 기술의 예시에 비유하면 FSD(Full Self-Driving)[92]에 해당하는데, 판결문 비식별을 전적으로 AI가 담당하는 상황이어서 오히려 사람이 개입하는 것 자체가 비효율적이라고 여겨질 정도의 기술 수준이다.

그러나 이러한 고도의 인공지능에서 드물지만 심각한 오류가 발생하여 개인정보 침해가 일어나면, 앞서 언급한 개인정보보호법상의 과실책임 규정을 기계적으로 적용하기에는 여러 한계가 드러난다. 인공지능은 외견상 스스로 합리적 결론을 도출한다고 보이더라도, 그 내부 과정과 근거가 불투명한 '블랙박스(Black Box)' 문제를 내포할 수 있어서, 결과적으로 오류가 발생했을 때 그 근본 원인을 파악하고 특정 주체의 과실이라고 특정하기가 쉽지 않기 때문이다. "인공지능의 판단이 잘못된 경우, 그 결과값에 대한 객관적 잘못 자체는 평가할 수 있어도, 그러한 결론에 이르게 된 과실을 인정하기는 어렵다"는 지적도[93] 이러한 맥락에서 제기된 바 있으며, 이는 개인정보처리자가 '고의 또는 과실이 없음을 입증'하기가 현실적으로도 극히 곤란해질 수 있음을 뜻한다. 즉, 시스템 운영자가 현재 기술 수준에서 합리적으로 요구되는 관리·감독 노력을 다했음에도 불구하고, 시스템의 내재적 복잡성이나 예측 불가능성으로 인해 발생한 오류에 대해 '과실 없음'을 입증하기란 사실상 어려울 수 있다는 것이다.

특히, 생성형 AI의 경우에는 이러한 문제가 더욱 복잡하게 나타난다. 생성형 AI는 그 작동 방식의 본질상 학습 데이터에 내재된 편향이나 불완전성으로 인해 예기치 않은 방식으로 개인정보를 유추하거나 잘못된 형태로 비식별 처리를 수행할 가능성(이른바 '환각(Hallucination)' 현상과 유사한 오류)이 상존한다. 이러한 생성형 AI의 고유한 특성으로 인해 발생하는 오류 가능성을 운영자가 어느 수준까지 예견하고 대비해야 하는지, 그리고 그 예방 조치가 불충분했을 때 이를 '과실'로 평가할 수 있는지 여부는 명확하지 않다. 만약 운영자가 현재 기술 수준에서 합리적으로 요구되는 기술적·관리적 보호조치를 다했다면, 이러한 AI의 본질적 한계에서 비롯된 예측 불가능한 AI 오류에 대해 과실책임을 묻기 어려울 수 있으며, 이는 결국 피해자 구제의 공백으로 이어질 수 있다는 비판이 제기된다.

이처럼 고도화된 AI 시스템의 오류에 대해 전통적인 과실책임 법리만으로는 책임 귀속 및 피해자 구제가 충분하지 않을 수 있다는 우려 속에서[94], 이에 대한 보완책으로 검토되는 것이 제조물책임의 적용 가능성이다. 판결문 비식별 프로그램이 별도의 물리적 장치와 결합하거나 서버 등 유체물 형태로 제공되어, '제조되거나 가공된 동산'으로 평가될 수 있다면(제조물 책임법 제2조 제1호), 그 개발·제조자를 책임 주체로 삼을 여지도 생긴다. 판례에서도 서버용 소프트웨어를 제조물로 본 사례가 있는데, "피고는 MS SQL 서버를 전자서적과 같은 형태로 씨디-롬

91) 서울중앙지방법원 2017. 2. 14. 선고 2015가단5072371.
92) FSD(Full Self-Driving)는 차량이 주행 전 과정을 인공지능으로 처리하여 운전자의 조작 없이도 스스로 주행하는 완전 자율주행 기술을 의미한다.
93) 이수경, 앞의 논문, 236면.
94) 신봉근, "인공지능(AI)의 무과실책임론", 「서강법률논총」, 제14권 제1호(2025), 139면.

(CD-ROM)이나 디스켓 등과 같은 일정한 저장매체에 저장하여 공급하거나, 웹사이트를 통하여 라이선스를 부여하고 프로그램을 다운로드받게 하는 형태로 공급하는데, 전자의 경우 저장장치와 소프트웨어를 일체로서의 유체물로 볼 수 있어 그 소프트웨어 역시 제조물로 볼 수 있고, 후자의 경우 디지털 형태로 공급되는 소프트웨어를 이용하기 위해서는 하드디스크 등과 같은 다른 저장 매체에 저장되어야만 사용할 수 있고 일단 소프트웨어의 공급이 완료된 시점에서 결국 그 소프트웨어가 일정한 저장매체에 담겨져 있는 상태로 되며, MS SQL 서버는 대량으로 제작·공급되는 것이어서 제조물 책임법이 적용되는 제조물에 포함시키는 것이 제조물 책임법의 제정목적에도 부합되므로, MS SQL 서버를 제조물로 봄이 상당하다."고 판시한 바 있다.[95]

다만, 프로그램 자체가 '설계·표현상 결함'을 지니지 않았다면 제조물책임을 인정하기 쉽지 않다는 점도 동 판결에서 확인된다. 결국 판결문 비식별 프로그램이 고도의 자율성을 지니면서도, 그 과정에서 예측 불가능한 오류가 발생하는 경우, 제조자에게 아무런 과실이 없더라도 이러한 오류 자체를 법적인 '결함'으로 평가하여 책임을 인정할 수 있는지가 핵심 쟁점으로 부각된다.

현행 제조물 책임법(제조물 책임법 제3조 제1항)은 제조물의 '공급 당시'에 이미 내재해 있던 결함(제조·설계 또는 표시상의 결함)이 원인이 되어 타인에게 손해가 발생한 경우, 그 결함이 확인되기만 하면 제조업자가 원칙적으로 무과실 책임을 부담하는 구조를 취한다.[96] 그러나 인공지능 프로그램은 최초 공급 이후에도 지속적인 업데이트와 학습을 통해 성능과 작동방식이 동적으로 변동될 수 있으므로, 특정 시점에 발생한 오류가 과연 '공급 당시'부터 존재했던 '결함'에 기인한 것인지 입증하고 특정하는 것이 매우 어려울 수 있다. 이러한 AI의 동적 특성은 전통적 '결함' 개념과 어떻게 접목될 수 있을지가 명확하지 않다는 근본적인 문제를 야기한다.[97] 아울러 판례 수준에서 소프트웨어의 결함 자체를 인정하여 책임을 인정한 예는 아직 확인되지 않아, 이론적으로는 AI의 잠재적 위험성에 상응하는 무과실책임론이 강화되어야 한다는 논의가 가능하더라도, 실제 법원에서 어떠한 판단을 내릴지는 여전히 불투명하다. 이는 결국 AI 기술의 특수성을 고려한 입법적 대응이나 새로운 법 해석론의 필요성을 강하게 시사하는 부분이다.

4.2.3. 소결: 고도화된 AI 시대의 책임 법리 재정립 필요성과 인공지능기본법의 함의

이상에서 검토한 바와 같이, 인공지능 기반 판결문 비식별화 시스템의 오류 발생 시 법적 책임 문제는 해당 AI의 기술적 수준과 자율성 정도에 따라 상이한 법적 과제를 제기한다. 인간의 판단을 보조하는 수준의 '약한 인공지능(weak AI)' 수준에서 비식별화 오류가 발생한 경우에는, 시스템의 한계를 인지하고 적절한 검증 및 감독 의무를 이행했는지 여부를 중심으로 기존 과실책임 체계(개인정보보호법 제39조 등)와 국가배상책임(국가배상법 제2조) 법리를 통해 운영·감독 주체의 책임을 묻는 것이 비교적 용이하며, 어느 정도 해결의 실마리를 찾을 수 있다.

그러나 문제는 4.2.2에서 살펴본 바와 같이, 기술이 고도로 발전하여 인간의 개입이 거의 불필요하거나 비효율적인 수준의 자율성을 갖춘 인공지능 프로그램, 특히 예측 불가능성과 불투명성을 특징으로 하는 생성형 AI 등이 활용되는 경우이다. 이러한 고도화된 AI 시스템에서 발생한 오류에 대해서는 전통적인 과실책임 법리만으로 피해자 구제를 충분히 보장하기 어렵다. 이는 AI의 '블랙박스' 특성상 과실의 존재 및 인과관계를 입증하는 것이 극히 곤란할 뿐만 아니라,

95) 서울중앙지방법원 2006. 11. 3. 선고 2003가합32082 판결.
96) 신봉근, 앞의 논문, 130면.
97) "제조자는 인공지능 모델의 개발(학습용 데이터 세트의 구축, 학습 알고리즘의 선택, 학 습의 실행 등)에 있어서, 인공지능에 의한 출력의 정밀도·확률의 최대화를 목표로 하지만, 인공지능이 어떠한 상황에서도 확실하게 요구대로의 운전조작을 출력하는 상태를 실현하기까지는 이르지 못한다.", 신봉근, 앞의 논문, 133면.

AI의 자율적 판단과 예측 불가능한 작동 방식이 기존 법체계가 상정하는 '예견 가능성' 및 '통제 가능성'이라는 과실책임의 근본 전제와 상충될 수 있기 때문이다. 이러한 법리적 공백을 메우기 위해 제조물책임 법리의 적용 또는 확장 가능성이 논의되지만, 이 역시 AI 소프트웨어의 '제조물성' 인정 여부, 지속적으로 변화하는 AI의 특성상 '공급 시점의 결함'을 특정하기 어렵다는 점, 그리고 소프트웨어 결함에 대한 판례 법리가 아직 충분히 형성되지 않았다는 현실적인 한계에 직면한다. 결국, 이러한 어려움은 고도화된 AI가 야기하는 손해에 대한 책임 귀속을 위해 기존의 과실 책임 원칙을 넘어선 제조물 책임의 확대 적용이나 새로운 형태의 무과실 책임 원리 도입 필요성에 대한 논의로 이어지고 있다.

이러한 책임 법리의 공백과 새로운 책임 원칙 모색의 필요성이 제기되는 상황 속에서, 최근 제정되어 2026년 시행을 앞둔 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 '인공지능기본법')의 등장은 주목할 만하다. 이 법은 AI 오류로 인한 민사상 손해배상 책임의 주체, 요건, 범위 등을 직접적으로 규율하거나 앞서 언급된 무과실 책임의 필요성에 부응하는 새로운 책임 체계를 도입하고 있지는 않다. 따라서 인공지능기본법의 제정만으로 앞서 지적된 과실 책임 및 제조물 책임 적용의 근본적인 난점들이 곧바로 해소된다고 보기는 어렵다.

그럼에도 불구하고, 인공지능기본법은 현행 책임 법리, 특히 과실 책임의 해석과 적용에 중요한 기준점을 제공할 수 있다는 점에서 그 의미를 찾을 수 있다. 특히, 판결문 비식별화 시스템이 해당될 가능성이 높은 '고영향 인공지능'(제2조 제4호 자목 등)에 대해 사업자에게 부과되는 각종 의무들은 중요한 함의를 지닌다. 예컨대, ▲위험관리방안 수립·운영, ▲기술적으로 가능한 범위에서의 설명 방안 수립·시행, ▲이용자 보호 방안 수립·운영, ▲사람에 의한 관리·감독, ▲관련 조치 내용의 문서화·보관 등 고영향 인공지능 사업자의 안전성·신뢰성 확보 조치 의무(제34조 제1항)나, ▲일정 규모 이상 AI 시스템의 안전성 확보 의무(제32조: 위험 식별·평가·완화, 위험관리체계 구축 등)는 AI 시스템 운영·감독 주체에게 기존 과실 책임 법리 하에서 요구되는 '주의의무'의 내용을 구체화하는 법적 근거가 될 수 있다. 즉, 이러한 법정 의무의 준수 여부가 향후 (비록 고도 AI의 본질적인 예측 불가능성으로 인한 책임 귀속의 어려움이 완전히 해소되는 것은 아니더라도) 과실 유무를 판단하는 데 있어 핵심적인 기준으로 작용할 가능성이 높으며, 이를 소홀히 한 경우 과실책임을 인정하는 중요한 근거가 될 수 있다.

또한, 인공지능기본법이 ▲안전성·신뢰성 검·인증 노력 의무(제30조 제3항) 및 국가기관 등의 우선 고려(제30조 제4항), ▲고영향 인공지능 영향평가 노력 의무(제35조 제1항) 및 국가기관 등의 우선 고려(제35조 제2항), ▲투명성 확보 의무(제31조) 등을 규정한 것은, 비록 직접적인 책임 규정은 아니지만, AI 시스템의 개발 및 운영 과정에서의 절차적 정당성과 위험 관리 노력을 강조하는 방향으로 책임 법리 논의를 이끌어갈 수 있음을 시사한다. 이는 책임 판단의 초점을 단순히 결과 발생 여부에서 AI 개발 및 운영 '과정'의 적절성으로 이동시키는 논의와도 맥을 같이 한다.

결국, 인공지능기본법은 현 시점에서 AI 책임 문제에 대한 완전한 해답을 제공하기보다는, AI의 위험을 예방하고 관리하기 위한 '사전적 규제'와 '현행 과실 책임 체계 내에서의 절차적 의무'를 설정하는 데 중점을 두고 있다고 평가할 수 있다. 따라서 AI 오류로 인한 손해 발생 시의 '사후적 책임 분담' 문제, 특히 4.2.2에서 제기된 고도화된 AI의 예측 불가능한 오류로 인해 과실이나 결함 입증이 극히 곤란한 경우의 근본적인 책임 공백 문제는 여전히 민법, 제조물 책임법 등 기존 법률의 해석론적 발전과 향후 추가적인 입법적 보완에 중요한 과제로 남겨져 있다. 고도화된 AI 시스템이 야기하는 책임 문제는 단순히 기존 법리를 기계적으로 적용하거나 일부 수정하는 것만으로는 온전히 해결하기 어려운, 보다 근본적인 도전을 제기하며, AI의 특수성을 전면적으로 고려하여 책임 귀속의 새로운 패러다임을 모색해야 함을 재확인시켜 준다. 따라서

향후 AI 관련 책임 법리 논의는 인공지능기본법에서 제시된 의무들을 현실적인 주의의무 판단 기준으로 삼으면서도, 여전히 남아있는 근본적인 법적 공백을 메우기 위한 노력을 지속해야 한다.

5. 결론

본 연구는 법원 판결문 공개 제도의 현황과 문제점을 심층적으로 분석하고, 개인정보 보호와 국민의 알 권리라는 상충하는 가치를 조화롭게 실현하기 위한 방안으로 생성형 AI 기반의 비식별화 시스템 구축 방법론과 그 성능 검증 결과를 제시하였다. 현행 판결문 공개 제도는 낮은 공개율, 접근성 제한으로 인해 그 공적 가치를 충분히 발휘하지 못하고 있으며, 이는 법적 안정성과 사법 투명성 확보에 걸림돌로 작용하고 있다. 특히, 개인정보 보호를 위한 비식별화 과정은 여전히 인력에 의존하는 수동적인 방식으로 이루어져 효율성과 정확성 면에서 한계를 드러내고 있으며, 이는 판결문 공개 확대의 주요 장애 요인으로 작용하고 있다.

이러한 문제점을 극복하기 위해 본 연구에서는 생성형 AI 기술을 활용한 판결문 비식별화 시스템을 설계 및 구현하고, 실제 법원 판결문 데이터를 기반으로 그 성능을 실증적으로 검증하였다. 실험 결과, 소형 언어 모델을 파인튜닝하여 구축한 시스템은 최고 수준의 거대 언어 모델에 준하는 비식별화 성능을 달성하면서도, 기관 내부망 환경에서 안전하게 운영될 수 있음을 확인하였다. 이는 개인정보 유출 우려 없이 판결문 공개 확대와 데이터 활용 가능성을 동시에 높일 수 있는 효과적인 기술적 접근 방식임을 시사한다.

더 나아가, 본 시스템 도입 시 발생할 수 있는 법적 쟁점, 특히 비식별화 오류로 인한 책임 소재 문제를 심층적으로 논의하였다. 비식별화 실패의 책임은 작업 주체가 인간인지 AI인지, AI의 정확도 수준은 어느 정도인지에 따라 법적 책임의 성격과 범위가 달라질 수 있으며, 특히 고도의 자율성을 가진 AI 시스템의 오류에 대해서는 기존의 과실책임 법리만으로는 충분한 피해 구제가 어려울 수 있다는 점을 지적하였다. 이는 향후 AI 기술의 발전과 함께 비식별화 시스템의 법적 책임 문제를 심도 있게 논의하고, 필요한 법적·제도적 보완책을 마련해야 함을 강조한다.

결론적으로, 본 연구는 생성형 AI 기술이 판결문 공개 제도의 혁신을 위한 핵심적인 도구가 될 수 있음을 보여주며, 기술적 해법과 함께 법적·제도적 개선 방안을 모색함으로써 개인정보 보호와 알 권리의 균형점을 찾고, 사법 시스템의 투명성과 공정성을 제고하는 데 기여할 수 있을 것으로 기대한다. 향후 본 연구에서 설계하고 검증한 시스템을 실제 법원 환경에 적용하고, 지속적인 성능 개선과 법적 쟁점에 대한 심층적인 논의를 통해 판결문 공개 제도가 한 단계 더 발전하는 계기가 되기를 바란다. 또한, 본 연구가 법률 인공지능 분야의 발전을 가속화하고, 나아가 법치주의 확립과 사회적 신뢰 구축에 이바지하는 단초가 되기를 기대한다.

참고문헌

학술지(국내 및 동양)

김용주, “재판기록 공개 관련 제도적 보완방안”, 「법학연구」, 제65권 제4호(2024).

김태형, “재판 방송의 가능성과 한계”, 「LAW & TECHNOLOGY」, 제13권 제6호(2017).

김홍화, “판결문 공개시 비실명화작업에 대한 국가배상책임”, 「영남법학」, 제44호(2017).

박성미, “사법분야 인공지능 발전을 위한 판결문 데이터 개선방안-‘판결서 인터넷열람 서비스’를 중심으로-”, 「경찰법연구」, 제19권 제3호(2021).

신봉근, “인공지능(AI)의 무과실책임론”, 「서강법률논총」, 제14권 제1호(2025).

이수경, “인공지능의 민사책임에 대한 소고”, 「민사법학」, 제107호(2024).

임을영, “생성형 AI 활용 단계에서의 개인정보보호법적 쟁점과 기술적 해결방안”, 「The Journal of Law & IP」, 제14권 제2호(2024).

임준규, “온라인상에서의 판결문 공개와 개인정보보호의 문제”, 「LAW & TECHNOLOGY」, 제10권 제1호(2014).

전원열, “판결 공개에 따른 프라이버시 침해와 민사소송법 제163조의2”, 「저스티스」, 통권 제168호(2018).

최선, “판결문 공개제도와 사법부의 책무성”, 「헌법학연구」, 제28권 제1호(2022).

학술지(서양)

Daniel Martin Katz et al., “A General Approach for Predicting the Behavior of the Supreme Court of the United States”, *PLOS ONE*, Vol.12 No.2(2017).

Elias Jacob de Menezes-Neto & Marco Bruno Miranda Clementino, “Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts”, *PLOS ONE*, Vol.17 No.2(2022).

판례

대법원 1996. 2. 15. 선고 95다38677 전원합의체 판결.

대법원 1998. 7. 24. 선고 96다42789 판결.

대법원 2003. 12. 26. 선고 2003다13307 판결.

대법원 2015. 8. 27. 선고 2012다204587 판결.

서울중앙지방법원 2006. 11. 3. 선고 2003가합32082 판결.

서울중앙지방법원 2008. 11. 21. 선고 2008가소36017 판결.

서울중앙지방법원 2017. 2. 14. 선고 2015가단5072371 판결.

헌법재판소 2005. 7. 21. 선고 2003헌마282,425(병합) 전원재판부 결정.

신문기사

민명기, 「판례를 돈 내고 사는 나라」, 법률신문, 2024. 8. 3자.

박수연 외 2인, 「[신년기획][판결문 전면공개를 향하여] ① 국내 판결문, 기본권 침해 우려해 공개 확대 꺼려」, 법률신문, 2024. 1. 1자

박찬, 「2024년 생성 AI 투자, 총 79조로 역대 최고 기록」, AI타임스, 2025. 1. 5자.

우주성, 「‘하급심 판결문’ 공개 확대한 法, 국내 ‘AI 변호사’ 출현 앞당기나」, 아주경제, 2023. 2. 22자.

이용경, 「[신년기획][판결문 전면공개를 향하여] ① 기자가 직접 써본 ‘판결 인터넷 열람 서비스’」, 법률신문, 2024. 1. 1자.

인터넷 자료

오픈데이터포럼, “2024 오픈데이터포럼 열린세미나(1차) 법원 판결문 개방 국내·외 현황 및 향후 정책방향 제언 PART 1 주제발표”, YouTube, <https://www.youtube.com/watch?v=ts_7HvdwlJY>, 00:08:29-00:08:35, 작성일: 2024. 7. 25.

Jijie Zhou et al., “Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots”, arXiv preprint, <https://arxiv.org/abs/2410.11876>, 작성일: 2024. 10. 10.

Shubhi Asthana et al., “Adaptive PII Mitigation for Large Language Models”, arXiv preprint, <https://arxiv.org/abs/2501.12465>, 작성일: 2025. 1. 21.

연구보고서

박우경, “민사소송 및 집행 절차에서의 개인정보 보호에 관한 연구”, 사법정책연구원, 2020.

이민형, “형사재판절차에서의 개인정보 보호에 관한 연구”, 사법정책연구원, 2021.

기타 자료

개인정보보호위원회, “비정형데이터 가명처리 기준 주요 내용”, 개인정보보호위원회, 2024.

법원행정처, “「2025년 판결서 등 비실명화 사업」 제안요청서”, 법원행정처, 2024.

송오섭, “판결서 열람·검색제도 개선을 위한 몇 가지 고려사항”, 판결문 공개 확대를 위한 국회 토론회 발표자료 (2019. 10. 25.), 국회의원 금태섭, 법원행정처, 2019.

이용재, “현행 판결문 공개제도 검토”, 판결문 공개 확대를 위한 국회 토론회 발표자료, 2019.

Rishi Bommasani et al., “On the Opportunities and Risks of Foundation Models”, Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, 2021.

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