

Article | Received 29 June 2026; Revised 29 September 2026; Accepted 9 October 2026; Published 23 October 2026
<https://doi.org/10.55092/let20260015>

Governing adjudication with AI: the normative-empirical gap in Korea's legal judgment prediction research and a sequenced agenda for the judiciary's AI roadmap



Sungsoo Choi^{1,2} and Taeyeon Oh^{1,3,*}

¹ Seoul AI School, aSSIST University, Seoul, Republic of Korea

² Korean Bar Association, Seoul, Republic of Korea

³ Department of Management Information Systems, Dong-A University, Busan, Republic of Korea

* Correspondence author; E-mail: toh1@dau.ac.kr.

Highlights:

- Korea's 2025 Judicial AI Roadmap makes adjudication a public-sector AI case.
- A defining feature of Korea's Legal Judgment Prediction (LJP) research is a normative–empirical disconnect.
- Engineer-jurist collaboration connects judicial AI design with the constitutional standards of culpability, reason-giving, and judicial independence.
- Korean criminal LJP is re-conceived as three stages: guilt determination, applicable provision, and sentence.
- Governed data access and interpretable baselines must precede any claim of deployable judicial AI.

Abstract: In December 2025 the Supreme Court of Korea adopted a staged Judicial AI Roadmap. Sentencing-system improvement and recidivism assessment sit among its declared tasks. A public-sector institution has thus committed to governing adjudication, a liberty-bearing state function, with AI, and the questions studied internationally under the heading of Legal Judgment Prediction (LJP) have become immediately practical for Korea. This article maps Korea's LJP literature in a scoping review. Reading that corpus against the global field, it claims that the field's defining feature is a normative-empirical disconnect. Its normative (legal) and empirical (technical) literatures have developed separately, with little cross-engagement between them, in that no study in either literature is co-authored by an engineer and a jurist, and no normative work evaluates a domestic model. This article makes two contributions from this diagnosis. First, it argues that engineer-jurist collaboration must tie the technical design of Korea's LJP to the culpability principle, reason-giving, and judicial independence, treating these constitutional values as design constraints rather than ex post safeguards, and proposes that governed data access and interpretable baselines with legally legible feature attribution (e.g., SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)) precede deployment. Second, it re-conceives the prediction target of LJP datasets for Korean criminal procedure as three distinct stages, which are guilt determination, applicable provision, and sentence, showing that task definition is itself a legally consequential design decision.



Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

Keywords: legal judgment prediction; AI governance; judicial AI; sentencing prediction; culpability principle; explainable AI; public sector AI; Republic of Korea

1. Introduction

Governments now reach for artificial intelligence as an instrument of governance, not merely as an object of regulation. They deploy it inside the institutions that deliver public functions. The judiciary is the hardest case of all. What a court administers is coercive state power over liberty, not a public service. Once a court begins to govern adjudication with AI, the familiar questions of efficiency and accuracy can no longer be held apart from due process, fairness, and the legitimacy of state punishment.

Korea has now made that move explicit. In December 2025 the Supreme Court's newly established Artificial-Intelligence Committee adopted a "Judicial AI Roadmap and Development Strategy", a staged program running through 2030 under the vision of "realizing judicial justice through human-centered AI" [1]. Several of its tasks bear directly on the substance of proceedings. They include an intelligent legal-information and similar-precedent recommendation service, improvement of the sentencing system, and recidivism assessment. Each is flagged higher-risk, to be introduced only after pilot operation and verification. With this step AI in Korean adjudication crosses a threshold. It passes from a speculative prospect to a declared institutional agenda, and a decade of international work conducted under the heading of Legal Judgment Prediction (LJP), the task of predicting a court's decision from the facts of a case [2], becomes a governance problem for Korea rather than an academic curiosity.

This article asks a public-sector governance question. Is Korea's research base ready to support the judiciary's AI program, and if not, what must be built first? To answer it, the article reads the Korean LJP literature against the global field and reaches a diagnosis with one central claim. The defining feature of Korean LJP scholarship to date is a normative-empirical disconnect. Constitutional and criminal-procedural standards have remained a separate line of inquiry, and they have not been translated into the design choices that empirical modeling requires.

From this diagnosis the article makes two contributions. First, it proposes a sequenced governance-and-research framework in which engineer-jurist collaboration ties technical design to constitutional and criminal-procedure standards, treating those standards as design constraints rather than ex post safeguards, so that governed data access and interpretable baselines precede any claim about deployable judicial AI. Second, it re-conceives the international LJP subtask taxonomy for the Korean criminal-procedural order into three stages, guilt determination, applicable provision, and sentence, and further shows that, within guilt determination, the global "charge prediction" label conflates structurally distinct predictive targets that do not transpose wholesale.

The rest of the article proceeds in five steps. Section 2 sets out the background of LJP and its techniques in brief, the Korean sentencing system and judgment-disclosure law, the constitutional and criminal-procedural standards that bound judicial AI. Section 3 turns to the analysis, reading the global research baseline and the Korean research landscape against that baseline and culminating in the normative-empirical disconnect. Section 4 develops the governance-and-research framework. Section 5 states the implications and the limitations of the analysis. Section 6 concludes.

2. Background

2.1. Legal judgment prediction and its techniques

LJP is the task of predicting the outcome of a court case from a text describing its facts [2]. It is conceived as decision support, not decision-making. The promise is a data-derived reference point for litigants, practitioners, and, on some accounts, courts, offered in the name of greater consistency and improved access to justice [3]. Technically, the work reduces to two layers of representing the text and modeling the outcome. Natural-language processing (NLP) turns unstructured judgment text into machine-readable form. The methods run from simple word-frequency representations, such as bag-of-words and Term Frequency–Inverse Document Frequency (TF-IDF) weighting, to embeddings that place semantically related expressions close together [4]. Machine learning (ML) then learns a mapping from such representations to the outcome, treating charge or article prediction as classification and sentence length as classification or regression. For years the field’s de facto baseline was a classical pairing of TF-IDF features with a Support Vector Machine (SVM), with Random Forests and Ensembles close behind [5].

The frontier has since moved. Deep learning arrived, then pre-trained models of the BERT family, and now large language models (LLMs) [6]. These often improve performance, but they behave as “black boxes” that can absorb and reproduce training-data bias [6]. In a legal setting, this opacity carries procedural significance, since a decision’s justification matters as much as its result. The concern has driven work on explainable-AI (XAI) methods such as SHapley Additive exPlanations (SHAP) [7] and Local Interpretable Model-agnostic Explanations (LIME), which attribute a prediction to its individual input features [8].

2.2. Korean criminal system and judgment-disclosure law

In Korea’s civil-law jurisdiction, criminal proceedings at court level move through three stages, consisting of the determination of guilt, the identification of applicable provisions, and the imposition of a sentence. Sentencing in Korea is unusually structured. The Sentencing Commission, established in 2007, issues guidelines for dozens of major offense categories. Under them a court moves from the statutory penalty range through an adjusted range to the final sentence, weighing a defined set of factors classified as aggravating or mitigating and as special or general. The factors are explicit and categorically defined. They are binned and weighted by category rather than scored numerically. The system therefore lends itself to feature-based machine learning and, as Section 4.3.1 argues, to legally legible explanation. However, the guidelines are advisory in the sense that a court must follow them as a general rule but may depart with stated reasons. As a result, the open-ended statutory factors of Article 51 of Korea’s Criminal Act (the offender’s age, character, intelligence and environment, the relation to the victim, the motive, means and result of the crime, and the circumstances following it) remain the governing frame.

Data availability is a separate constraint. Under Korean judgment-disclosure law, public access to judgments is qualified rather than comprehensive. Although judgments are accessible in principle, Article 59-3 of the Criminal Procedure Act and Article 163-2 of the Civil Procedure Act allow courts to restrict disclosure in specified circumstances. In practice, not all judgments are available online or free of charge.

2.3. *The constitutional and criminal-procedural standards that bound judicial AI*

Among the operative constraints governing adjudication with AI in Korea, three are central.

- The culpability principle. Korea's Constitutional Court has repeatedly affirmed it as derived from the rule of law and from human dignity under Article 10 of the Korean Constitution, holding that “no penalty may be imposed on a person without culpability” [9,10].
- The reason-giving obligation. Articles 39 and 323 of the Criminal Procedure Act require a judgment to specify the criminal facts, the gist of the evidence and the application of statutes, and, where facts grounding aggravation or mitigation have been argued, the court’s judgment on those facts. Explainability, on this view, is a matter of statutory justification, not merely of user trust.
- Judicial independence under Article 103 of the Constitution. An AI recommendation generated or endorsed within the court system could exert de facto pressure on a trial judge’s discretion, a worry separate from bias and transparency.

3. Analysis of the Korean research landscape

3.1. *The Korean field against the global baseline*

3.1.1. The global research baseline

The international field supplies the reference point against which the Korean field is read, and it is summarized here only to that end, drawing principally on one systematic review and two large-scale surveys of the field [11–13], with adjacent surveys used for corroboration. A peer-reviewed systematic review of LJP synthesized thirty-eight primary studies (2014–2023) around datasets, subtasks, methods, evaluation, and challenges [11]. Three features of that baseline matter for the comparison.

First, concentration. The field leans overwhelmingly toward charge prediction, which accounts for approximately 85%–88% of studies, while law-article prediction and sentencing prediction remain marginal, each accounting for less than 10%. The analysis leans on this order-of-magnitude split, corroborated in its broad shape by adjacent surveys [3], not on the precise figures. A dedicated LJP survey sharpens the point by decomposing the field into six subtasks ordered by difficulty (violation, court decision, charge, law article, alleged law article, and prison term), and observing that the available corpora each cover only a subset of them, with prison term prediction supported by a single Chinese dataset [12]. Read against that taxonomy, the concentration is not merely a matter of study counts, because the subtasks closest to sentencing are the hardest and the least resourced. One clarification matters doctrinally. “Charge prediction” is a mixed category. It folds together the European Court of Human Rights (ECtHR) dataset’s binary State-level violation judgments and the multi-class offense identification of Chinese and Indian datasets [5]. The label is a coding artifact, not a doctrinal unity, which is exactly why it cannot be imported into Korean procedure wholesale, as Section 4.2 develops.

Second, geography. The field draws on a small set of court-derived datasets concentrated in a few jurisdictions, namely Europe, China, the United States and India. Those resources are overwhelmingly English- and Chinese-language, so findings may not carry over to other languages or procedural traditions.

Third, method and evaluation. The methodological backbone is classical and interpretable, with TF-IDF and SVM foremost, alongside random forests, ensembles, and embedding methods, and it

furnishes a cumulative baseline against which newer methods are measured. Evaluation is the weaker link. It places considerable emphasis on accuracy, often without explaining why that metric is appropriate, even though accuracy can be misleading on the imbalanced datasets typical of legal data. The review clearly identifies the field's principal obstacles, namely limited quantity and accessibility of data, severe class imbalance, weak cross-jurisdiction generalizability, and the opacity of powerful models. Explainability, it argues, is crucial to the trust on which judicial adoption depends.

A deeper limitation cuts beneath all of these. Of 171 LJP papers surveyed, only about 7% used data available before the decision, while the rest trained on the “facts” sections of final judgments, which are summaries written by the deciding court and tailored to the outcome already reached [13]. Such systems therefore demonstrate not forecasting but the recovery of a decision already made, and the gap is measurable, with one architecture scoring F1 0.92 on final-judgment facts falling to 0.64 on pre-decision facts [13]. High reported accuracy therefore cannot be read as predictive capability, a caution that bears on the Korean studies below.

3.1.2. The Korean corpus

The Korean corpus was assembled through a three-stream bilingual search covering Korean academic databases, grey literature, and internationally indexed work by Korean-affiliated authors, over January 2014 to the search date of 27 August 2026, the search having been run during revision of the manuscript. The search strategy was developed and executed by the first author and independently reviewed by the second author. The criteria were deliberately broadened along publication type, language, and period, because a peer-reviewed journal-only, English-only standard would return too small a set of Korean studies to characterize the field. Screening the 2175 records identified yielded 42 Korean LJP records, of which 22 are normative and 20 empirical, the latter representing 14 distinct studies once duplicate outlets are consolidated. Figure 1 records the identification process across the three parallel search streams. The 93 search strings that produced it are reproduced in full, with their English glosses and the number of records each returned, in Supplementary data S1 (sheet 1_Search_strategy). Work adjacent to legal judgment prediction does not form part of the corpus, because in each case the predicted value is not an adjudicative outcome of a court or a binding tribunal. Such work includes demonstrations of general-purpose LLMs on Korean cases [14], evaluation on legal-examination items [15,16] and de-identification [17]. It also includes information extraction [18], non-binding dispute determinations [19] and data structuring [20].

On this basis, Table 1 sets out the fourteen studies of the Korean empirical corpus by dataset, NLP method, modeling approach and evaluation. Six studies draw on the LBOX OPEN benchmark dataset, six on self-built court corpora and two on non-court tribunal records. Nine of the fourteen predict a criminal adjudicative outcome, three a civil money judgment and two the decisions of non-court administrative tribunals. LLM and GPT systems figure prominently in the recent work. Evaluation is concentrated on accuracy. Explainability is rarely reported. All fourteen are authored by engineering or computing teams, and none lists an author with legal training.

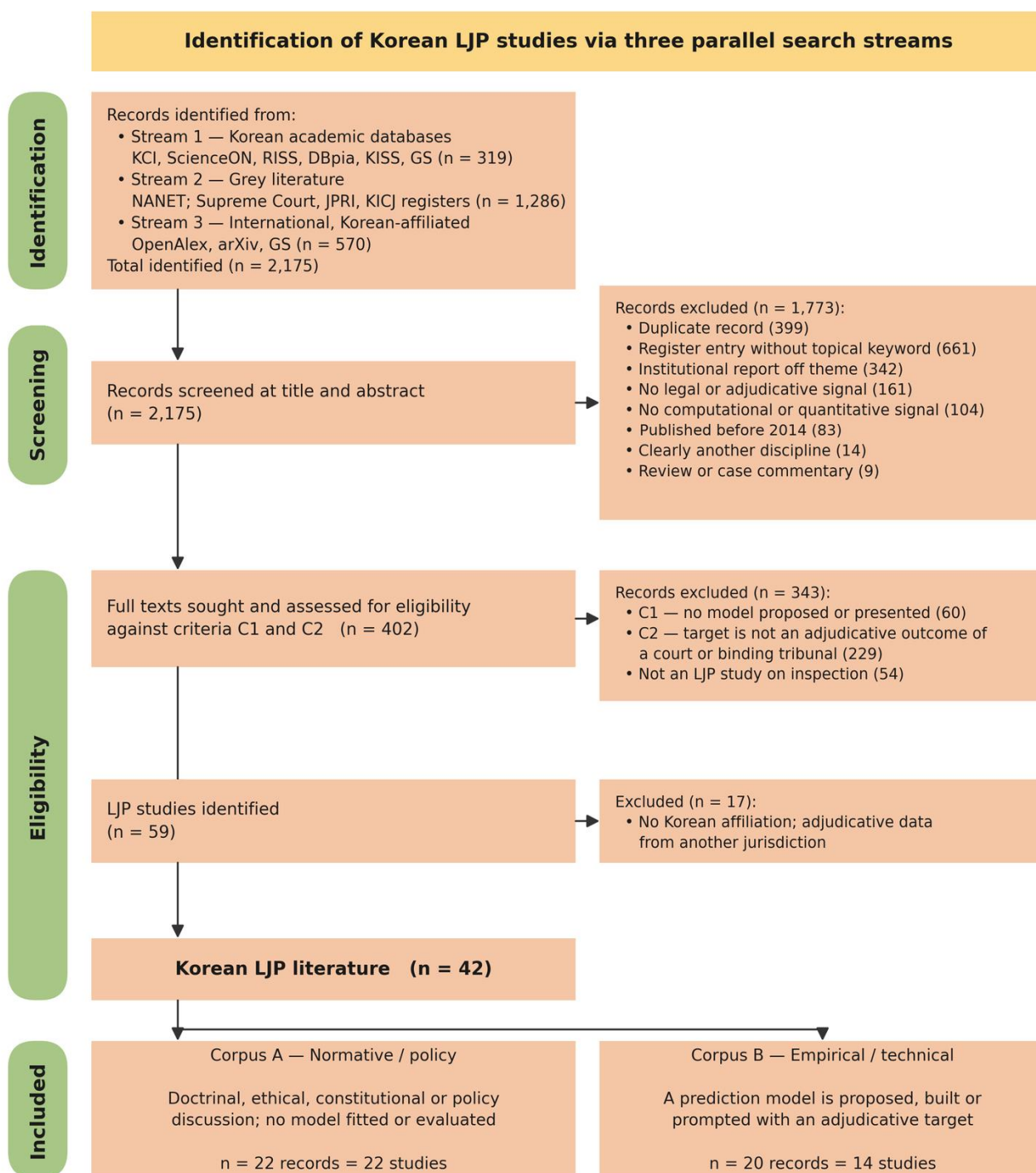


Figure 1. Identification of Korean LJP studies via three parallel search streams (PRISMA-ScR flow).

Table 1. The Korean empirical LJP corpus (N = 14).

Study	Subtask	Dataset	Experimental methodology	Evaluation	Author discipline
Byun <i>et al.</i> [21]	Term-of-penalty	Self-built	None (hand-encoded); Regression, matrix factorization	MAPE	Engineering
Shin [22]	Term-of-penalty	Self-built	None (hand-encoded); Regression, matrix factorization	RMSE, MAE, MAPE, R ²	Engineering
Kim [23]	Term-of-penalty	Self-built	Word2Vec; Multiple regression	R ² , F	Engineering
Ko <i>et al.</i> [24]	Term-of-penalty	Self-built	None (structured variables); Correlation and regression analysis	R ² , accuracy	Engineering
Lee [25]	Charge	Non-court tribunal	FastText; BiLSTM	Accuracy	Engineering
Hwang <i>et al.</i> [26]	Charge; Law article; Term-of-penalty	LBox	Tokenization; Decoder-only Transformer (LCUBE)	Accuracy, F1	Engineering
Park <i>et al.</i> [27]	Charge; Law article; Term-of-penalty	Non-court tribunal	Embeddings; Multi-task LSTM with attention	Accuracy, F1	Engineering
Park & Lee [28]	Charge	Self-built	FastText, n-grams; Random Forest, neural network	Accuracy	Engineering
Lee [29]	Charge	LBox	Embeddings; KPF-BERT, LBox-lcube vs. GPT-3.5/4.0	Accuracy	Engineering
Cho & Han [30]	Term-of-penalty	LBox	Precedent retrieval; ChatGPT with retrieval	Weighted F1	Engineering
Min <i>et al.</i> [31]	Term-of-penalty	LBox	Word2Vec, TF-IDF; CNN	Accuracy	Engineering
Min & Noh [32]	Term-of-penalty	LBox	Embeddings; ALBERT + Bi-LSTM	Accuracy, F1, RMSE	Engineering
Choi [33]	Term-of-penalty	LBox	TF-IDF; BERT-family	F1	Engineering
Jeong [34]	Term-of-penalty	Self-built	BERT encoding; BERT backbones, multimodal	Accuracy, F1, MAE, RMSE	Engineering

Note. Case type of the prediction target. Criminal, nine studies [21,22,26,28–33]; civil, three studies ([23] divorce property claims, [24] and [34] apartment- and residential-building defect compensation); non-court administrative tribunal, two studies ([25] tax tribunal, [27] maritime safety tribunal). For [26] the tasks charted are the benchmark's criminal tasks.

By contrast, Korea's normative literature is somewhat more extensive than its empirical counterpart and is set out separately in Table 2, which records each study's focus and the global subtask to which it relates. These studies broadly accept that AI-based tools will be adopted in the judicial process, while pressing questions of explainability, fairness, the judge's conscience, and the procedural legitimacy of criminal process.

Table 2. The Korean normative LJP literature (N = 22) and its relation to the global subtask scheme.

Study	Related subtask (Dina <i>et al.</i> scheme)	Focus	Contribution	Author discipline
Yang [35]	Charge	Legal-system change	How weak and strong AI would alter the criminal-justice system	Law
Yang [36]	Charge	Replaceability of judges	Whether AI algorithms can substitute for judicial decision-making	Law
Shin [37]	Term-of-penalty; Charge	System development	National R&D final report on an AI judgment-prediction system built on legal and social factors	Engineering
Jeong & Song [38]	Charge	Criminal-justice issues	Principal doctrinal issues raised by the development of AI in criminal justice	Law
Lee [39]	Term-of-penalty	Constitutional review	Constitutional problems of recidivism-prediction algorithms, centered on State v. Loomis	Law
Joo & Jung [40]	Term-of-penalty	Algorithmic bias	Normative design for algorithmic bias in crime prediction and criminal procedure	Law
Seok <i>et al.</i> [41]	Charge	Legal-tech start-ups	Legal issues arising across the service life cycle of judgment-prediction services	Law
Kang <i>et al.</i> [42]	Charge	Legal-tech policy	Legal policy for introducing and responding to legal technology (KICJ institute report)	Law
Jeong [43]	Charge	EU regulation	European Parliament resolution on banning decisions taken by AI in criminal justice	Law
An & Park [44]	Term-of-penalty	Procedural adoption	Principal issues in introducing AI systems into criminal procedure, including sentencing support	Law
Kim [45]	Charge	AI ethics	AI-ethics issues in criminal justice and their criminal-policy implications	Law
Yang [46]	Charge	Litigation finance	Third-party litigation funding and AI-based prediction of legal risk	Law
Lee [47]	Term-of-penalty	Racial bias	Proposal for using AI prediction programs while avoiding racial bias	Law
Park [48]	Term-of-penalty	Sentencing	AI for gathering public sentiment, assessing risk and analyzing sentencing	Law
Joo [49]	Term-of-penalty	Risk assessment	Sentencing judgment using AI risk-assessment programs in criminal procedure	Law
Kim [50]	Term-of-penalty	Pre-sentence investigation	Preliminary study on the use of AI in sentencing investigation	Law
Lee [51]	Term-of-penalty	Legal policy	Legal-policy study for the effective introduction of AI into the judicial system	Law
Lee [52]	Charge	Professional practice	Methodological study on the use of generative AI in the legal profession	Law
Baek [53]	Term-of-penalty	AI judges	Meaning and tasks of the argument for introducing AI judges into criminal trials	Law
Lee [54]	Term-of-penalty	Judicial conscience	AI sentencing and the conscience of the judge	Law
Son [55]	Charge	Fairness	AI fairness and the normative limits of the criminal-justice system: a comparative analysis	Law
Kang [56]	Term-of-penalty	Governance	Institutional and technical policy choices for governing AI-based empirical sentencing analysis; written by a sitting judge	Law

By way of qualification, this corpus is not the whole of Korean scholarship on AI in adjudication. It is the normative subset of the records that reached full eligibility with an identifiable Korean LJP prediction target. Thus, general discussion of AI in the judiciary attaching to no such target was excluded at that stage. Works of that kind, among them [14] and [57–59], appear in the text as background without being counted.

3.2. Subtask coverage

The Korean corpus is next mapped onto the three subtasks of the global scheme, counted three ways (Table 3). The first count covers the subset meeting the inclusion criteria of the global review of Dina *et al.* [11] (N=6), which admit only peer-reviewed articles published between 2014 and 2023, the second covers the full corpus (N=14), and the third applies the six-subtask scheme of Feng *et al.* [12] (N = 14).

Table 3. Distribution of LJP studies by subtask: the global field *versus* the Korean empirical corpus.

Subtask (Dina <i>et al.</i> scheme)	Global share (N = 38)	Korean count, Dina-criteria subset (N = 6)	Korean count, criminal only (N = 9)	Korean count, full corpus (N = 14)	Corresponding subtask (Feng <i>et al.</i> scheme)	Korean count, full corpus (N = 14)
Charge prediction	~87% (33)	4	3	5	Charge	5
Law article prediction	~8% (3)	2	1	2	Law Article	2
Term-of-penalty prediction	~5% (2)	4	7	11	Prison Term	11
– (no counterpart)	–	–	–	–	Court Decision	3
– (no counterpart)	–	–	–	–	Violation	0
– (no counterpart)	–	–	–	–	Alleged Law Article	0

Note. The global shares are the coding decisions of the cited review [11] rather than properties of the field. The criminal-only column restricts the corpus to the nine studies whose prediction target is a criminal adjudicative outcome (see the note to Table 1). Under the six-subtask scheme the criminal-only counts are Charge 3, Law Article 1, Prison Term 7, Court Decision 1, Violation 0 and Alleged Law Article 0. Of the three studies in the Court Decision row of the full corpus, two, [23] and [24], predict civil money judgments. Of the six studies admitted under the Dina-criteria subset, three are criminal [21,26,28], one is civil [24] and two are non-court tribunal decisions [25,27].

The weighting differs from the international one. Charge prediction, which in the Korean frame is guilt determination, accounts for roughly 87% of the global literature but only five of the fourteen Korean studies. Law-article prediction, about 8% globally, appears in only two Korean studies. Term-of-penalty prediction shows the widest divergence. The most neglected category globally, at about 5%, it is where Korean empirical work is most concentrated, with eleven of the fourteen studies. Five of the fourteen do not predict a criminal adjudicative outcome at all, three predicting civil money judgments and two the decisions of non-court administrative tribunals (see the note to Table 1). These five fit imperfectly within the three-subtask scheme, which is why Table 3 also reports the six-subtask classification and a criminal-only count. That mix reflects a field whose prediction-task coverage remains uneven. Term-of-penalty prediction is not only the harder inference but the one most closely bound to judicial discretion.

The comparison holds on both counts, though not identically, and it is the criminal-only count that carries it. Term-of-penalty prediction, about 5% of the global literature, is addressed by seven

of the nine studies whose target is a criminal adjudicative outcome, by eleven of the fourteen in the full corpus, and by four of the six studies admitted under the global review's criteria. The fact of its over-representation therefore depends neither on the broadened criteria nor on the inclusion of civil and administrative work. The reversal of the global ordering is the more sensitive claim. It holds on the criminal-only count, where term-of-penalty prediction at seven of nine clearly displaces charge prediction at three of nine, and it holds on the full corpus; it does not hold in the matched subset, where charge prediction and term-of-penalty prediction are each addressed by four studies. The matched subset should in any event be read with its composition in view, since of its six studies only three are criminal, one predicts a civil money judgment and two the decisions of non-court tribunals.

The central claim is framed for criminal LJP and the Judicial AI Roadmap, so the criminal-only count is the one that bears on it. The broader count is nonetheless reported rather than discarded because the civil and administrative studies exhibit the same methodological pattern, the prediction of a monetary or dispositive quantum from judgment text, and because narrowing the inclusion criteria progressively shrinks the Korean corpus, with restriction to criminal prediction leaving nine studies and further restriction to the criteria of the global review leaving only three.

3.3. *Data infrastructure*

Korea's single most important enabling development is the LBOX OPEN dataset, still the only large-scale, publicly available Korean legal-AI benchmark. It holds roughly 147,000 precedents across multiple tasks, including criminal and civil judgment prediction, with a dedicated Korean legal language model [26]. For Korea it does what the ECtHR and China Judgment Online datasets do internationally.

Two foundational global tasks already inside LBOX OPEN are offense-name classification and statute classification, mapping facts to an offense name and facts to an applicable statute, each released with baseline results [26]. Of the six empirical studies drawing on LBOX OPEN, five address term-of-penalty prediction, while offense classification appears in one and statute prediction only in the benchmark's own baseline (Table 1). Legally restricted access to the full body of decisions explains much of this narrow use, since that body has never been consolidated into open, research-ready prediction datasets at the scale of the published international benchmarks. Other national resources increasingly support legal-language modeling and statutory access, but they do not replace a labeled judgment- or sentence-prediction benchmark.

3.4. *Methods and evaluation*

The Korean methodological profile differs from the global one in an important respect, in that it lacks the cumulative classical baseline against which newer techniques can be compared. The earliest studies fall within the classical, interpretable paradigm of regression, matrix factorization and decision trees on hand-encoded sentencing factors in the two 2018 studies, and regression on morphologically processed or structured variables through 2021 (Table 1). But they were exploratory, and none was taken up as a shared reference point [21]. The studies from 2021 onward begin somewhere else, from deep learning, pre-trained models and LLMs: a dedicated Korean legal language model for the benchmark [26], BERT-family Korean models with legal-syllogism prompting for sentence classification [33], a few-shot ChatGPT setup for sentence-range prediction [30], fine-tuned BERT offense classifiers benchmarked against GPT-3.5 and GPT-4, and an ALBERT-Bi-LSTM sentencing model benchmarked against GPT-4o

and LLaMA (Table 1). Classical features have not disappeared, as TF-IDF reappears in two of the recent studies, in one as an internally reconstructed baseline [33] (Table 1). Yet no later study treats the earlier work as a benchmark, so the field never built the stable, cumulative interpretable baseline an established field relies on.

The Korean studies likewise lack a common basis for evaluation, although the small corpus does not support a stable cross-study metric pattern. Where evaluation is reported, the studies use task-specific but mutually inconsistent metrics, since regression studies report MAPE, RMSE, MAE or R^2 , while classification studies report accuracy or F1, leaving no common basis for comparison (Table 1). Nor does any Korean study report a temporal train/test split, a notable gap, since sentencing norms drift over time and testing should rest on data accessible before the decision was made [13]. Control of label leakage, the exclusion of the parts of a judgment that already encode the outcome, varies sharply across studies. This defect is pervasive in the global field as well [13], where only about 7% of 171 surveyed papers used pre-decision data, and it materially affects how reported figures should be read.

3.5. *The normative-empirical disconnect*

Across both corpora, the fourteen empirical studies and the twenty-two normative works alike, not a single study is co-authored by an engineer and a jurist. Every empirical study is written by an engineering or computing team and none lists an author with legal training, while all but one of the normative works are written by law faculty or practitioners.

The normative literature overlaps substantially with the empirical field but rarely engages its models directly. Twelve of the twenty-two normative studies bear on term-of-penalty prediction, yet none evaluates a domestic model, because their concerns lie elsewhere, in constitutional review, algorithmic bias, risk assessment, judicial conscience and fairness (Table 2). A first strand addresses algorithmic bias and safeguards. It argues that, given the opacity of automated decision-making, securing algorithmic transparency and accountability is needed to reconcile such tools with the rule of law [40]. A second strand treats recidivism-risk prediction as the paradigm case. It dissects the black-box problem of instruments such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), which is a recidivism-risk scoring tool used in US sentencing and parole decisions, and advances explainable AI as the precondition for fairness [57], with a parallel constitutional critique [39]. The work of a sitting judge sorts these criticisms into discrete problems, among them reliance on correlation over causation and the violation of act-based culpability [58]. A third, recent strand turns to generative AI [59], alongside sustained legal scholarship on sentencing and AI [48] and, in 2025–2026, on AI judges, the judge's conscience, AI fairness and the governance of empirical sentencing analysis [56].

This discourse is, however, overwhelmingly comparative and anticipatory. Its central object is the American experience with instruments such as COMPAS, invoked to reason about tools Korea has not built. Its center of gravity is recidivism-risk assessment, not the sentence prediction the domestic empirical work concentrates on. Even where a risk score has entered sentencing, as in *State v. Loomis* [60], the use was upheld only where the score is not determinative and is accompanied by written cautions, and the court itself acknowledged that it could not gauge the severity of the sentence. Court-operated sentencing assistance is nevertheless not unknown. Shanghai's "206 System", for example, uses prior cases and sentencing factors to support sentencing recommendations within

China's smart-converts program [61–63]. Korea has no comparable system in operation, and the December 2025 Judicial AI Roadmap is the preparatory step toward one. The Korean normative literature analyses foreign and hypothetical systems, while the domestic empirical literature builds models without reporting engagement with the criteria that literature sets out. The two literatures have developed separately, with little cross-engagement between them, and that separation is itself a gap in the research base. This is the normative-empirical disconnect that the remainder of this article addresses.

A distinctively Korean regulatory episode illustrates the disconnect. The “Lawtalk” sentence-prediction service, launched in 2020, is an NLP-assisted statistical analytics tool rather than a generative system. It became the object of a prolonged dispute over platform-mediated lawyer advertising, substantially settled by the Constitutional Court's 2022 decision on the Bar Association's advertising rules [64]. The Court struck down the broad restrictions and kept only one, a ban on advertising that markets predictions of court outcomes. Following that decision, the Ministry of Justice overturned the disciplinary measures against the affected lawyers [65]. The net effect is a bounded constraint, under which marketing sentence prediction as attorney advertising remains restricted while its use by courts, prosecutors and firms as internal decision support is not foreclosed.

One recent study begins to bridge the gap, in which a gender-law scholar tested whether ChatGPT and other LLMs apply the Korean Supreme Court's 2018 gender-sensitivity doctrine to real sexual-violence cases [14]. The model invoked the doctrine without consistently internalizing it as a reasoning requirement and reproduced rape-myth reasoning from its training data. As a rare instance of a normative scholar testing a working model on Korean cases, the study points directly to the collaborative model developed in Section 4.5.

4. A governance-and-research framework for judicial AI in Korea

4.1. *The decision-support principle of LJP*

An LJP output may inform a decision on guilt or sentence, but it must never determine one. This is a constitutional requirement. First, judicial independence under Article 103 of the Constitution is protected by denying the output any procedural status the judge must answer to, since a recommendation generated inside the court system would otherwise exert institutional pressure on the discretion Article 103 secures. Second, due process under Article 12(1) of the Constitution is protected by requiring the court itself to author the reasons that carry the decision, so that the defendant can contest grounds a human decision-maker has adopted and can be held to. The output is therefore neither evidence nor a sentencing decision, but an advisory input to a decision the judge alone makes.

4.2. *Build the foundational subtasks*

The foundational tasks must be re-conceived for the Korean criminal-procedural context as a three-stage taxonomy of guilt determination, applicable provision, and sentence. Modeling the three stages as separate tasks keeps the features that predict sentence length out of the prior question of whether the defendant committed the act, so that a sentencing pattern can never become a reason to convict.

Within guilt determination, three distinct targets are available, and they are not interchangeable. The first is the prosecutor's charged offense, the offense named in the indictment. The second is the court's adjudicated offense, which may differ from the first, since the court may convict on a reduced offense within the identity of the charged facts. The third is the guilt, acquittal or dismissal outcome, a disposition rather than an offense label at all.

A model trained on the charged offense learns prosecutorial charging behavior, while one trained on the adjudicated offense learns judicial characterization of the same conduct. A model trained on the disposition learns neither, but rather whether the prosecution succeeded. They have different users, and conflating them produces a system whose output cannot be interpreted. This is what the international corpora obscure. The ECtHR datasets predict a State-level violation disposition, the Chinese datasets predict an offense label, and both are filed under "charge prediction" even though the predicted object is not the same.

Two consequences follow for Korean work. First, any Korean guilt-determination dataset must state which of the three it targets, and its labels must be drawn accordingly, since the indictment and the judgment are distinct documents recording distinct determinations. Second, the divergence between charged and adjudicated offense is not noise to be minimized but a substantive object of study in its own right, because the rate and pattern of that divergence is a measure of how the charging and adjudicating functions relate.

4.3. Interpretable baselines and standardized evaluation

4.3.1. Performance

Korean research should re-establish classical and fine-tuned ML baselines as the standard against which performance is measured, rather than vaulting straight to LLMs without a reference point. On the Korean benchmarks examined here, off-the-shelf LLMs have not been shown to dominate fine-tuned domain models. A few-shot ChatGPT setup performs roughly level with, not ahead of, a fine-tuned BERT baseline on sentence-range prediction [30]. GPT-4 scores at or below half on Korean legal-reasoning items [15]. Domain tuning, not scale, is what narrows that gap, as an eight-billion-parameter open model fine-tuned on Korean legal instruction data and then trained with verifiable-reward reinforcement learning raises accuracy on Korean professional-examination items from 25.0 to 43.3 and on case-law questions from 60.0 to 76.0 [16]. The pattern echoes the international finding that general-purpose LLMs did not uniformly dominate fine-tuned or classical baselines on the harder, integrative steps of guilt determination and sentencing [66–70]. Accordingly, LLM superiority on Korean LJP cannot be assumed and must be measured against a supervised or fine-tuned baseline, though no general performance hierarchy is claimed.

4.3.2. Explainability

SHAP and LIME apply to classical ML models and LLMs alike, since they read only inputs and outputs rather than a model's internals. But their legal value depends on what the inputs are. Attribution, meaning how much each input contributed to the prediction, is legally legible when the inputs are defined categories, such as the aggravating and mitigating sentencing factors of Section 2.2. A SHAP value on

a defined sentencing factor points to a category the record already uses, and where the factor is also a statutory ground of aggravation or mitigation, to one on which Article 323(2) of the Criminal Procedure Act requires the judgment to state its determination once the ground is argued (Section 2.3). Either way the model's basis can be contested in the terms the record already uses. Attribution over defined factors is in this sense more legally legible than token-level saliency over free text, which highlights the words of a judgment rather than the categories on which the sentencing debate turns.

In practice, classical and tree-based models, which already take such features as inputs, admit mature attribution methods [7] that render a prediction's basis legible in legally meaningful terms. That legibility, however, is a property of the feature set rather than of the attribution method, so it survives only if the features are built to track the statutory and guideline sentencing categories. In addition, interpretable models built on defined features have performed as accurately as black-box and commercial tools on recidivism-risk tasks [71], which indicates that interpretability and predictive accuracy are not necessarily in tension.

An objection arises at this point. If the model's features are the guideline factors, and the guidelines already specify how those factors combine, what does the model add beyond the guideline itself? The answer is that it models not the guideline but the discretion the guideline leaves open. As Section 2.2 noted, the Korean guidelines are advisory, they bin and weight factors by category, and they leave the movement from the adjusted range to the final sentence to the court, and Article 51 of the Criminal Act remains the governing frame. A fitted model therefore makes visible what the guideline text cannot, namely which factors actually move sentences and how consistently comparable cases are treated, which makes sentencing practice traceable and reviewable by the Sentencing Commission and on appeal [56].

4.3.3. Standardized evaluation and reproducibility

Standardization requires four elements, each addressed in turn below. First, give the task a common formulation, since sentence-length prediction is naturally regression or ordinal, with standardized bins where classification is used [12]. Second, build in reproducibility through shared dataset partitions on the common benchmark [26]. Third, control label leakage, together with validation on later cases [13], so that reported accuracy reflects prediction rather than the recovery of information already present in the judgment. Finally, on metrics, report several complementary measures fitted to the ordinal, continuous character of sentencing outcomes, such as mean absolute error, distance-weighted agreement, and calibration of predicted ranges, rather than accuracy or undifferentiated macro-F1 [12], which treat a one-month and a ten-year error alike.

4.4. Governed expansion of research-ready data access

The Supreme Court portal generally requires a known case number, offers no bulk download or API, and charges a per-document fee. Constructing a large-scale dataset from official sources is therefore, in practice, severely constrained. The corrective is a governed expansion of research access, not indiscriminate publication. A portion of the judgment corpus should be opened to researchers, through de-identified bulk release, a secure research enclave, or a research-purpose API, so that academic inquiry can run in parallel with the judiciary's program rather than wait on it. This is consistent with the roadmap, which designates extraction and de-identification of personal information a near-term task [1].

The enclave itself needs no new statute. The Court Organization Act already authorizes the judiciary to establish and operate such a facility and to maintain a structured sentencing-factor schema. The disclosure provisions of the Criminal and Civil Procedure Acts open judgments to any person and make de-identification a duty of the court. The Personal Information Protection Act permits pseudonymized personal data to be processed for scientific research without consent, subject to safeguards, audit logging and a re-identification ban. What requires legislative amendment is not the facility but the pipeline into it.

- The Act on Promotion of the Digitalization of Criminal Justice Procedures prohibits any use of information held in the criminal justice information system outside case handling, on pain of criminal penalty, and the courts are covered bodies obliged to store judgments there. A new paragraph should therefore permit processing of pseudonymized criminal justice information for statistical, scientific-research and archival purposes within an approved secure environment, with no export of raw data.
- The delegating paragraphs of the two disclosure provisions leave method and procedure to Supreme Court Rules but provide no bulk channel, and the governing Supreme Court Rule charges per document. Each should therefore carry express authority for bulk, machine-readable supply to approved research institutions, with fees waived or reduced.
- Criminal judgment disclosure is confined to cases that have become final. Disclosure should therefore extend to first-instance judgments before finality, as the Civil Procedure Act was amended in 2020 to do.

4.5. Engineer-jurist collaboration as the connective tissue

The disconnect identified in Section 3.5 has a structural source, in that Korean empirical studies are led predominantly by engineering teams, while practicing legal professionals and normative scholars remain largely absent. Sustained engineer-jurist collaboration is therefore necessary. Collaborations pairing AI engineers with legal-domain experts would do three things at once. First, they would supply the legal expertise that credible Korean LJP modeling requires. Second, they would expose that modeling to the fairness and due-process scrutiny the normative scholarship has articulated. Third, they would give normative scholars a domestic system to test their claims against, rather than the foreign tools they have had to write about instead. The “AI judge” study [14] is an early instance of exactly this pairing. Turning it into sustained engineer-jurist collaboration is the organizational shift this framework calls for, and its timing is doctrinally significant.

The constitutional standards of Section 2.3 must be honored at the design stage. Once the task, labels and features are fixed, no later review can undo those choices. Design-stage collaboration is therefore the last point at which these standards can still be built in. The roadmap should accordingly encourage joint law-and-engineering research and fund it through the same mechanism this field has already relied on, since earlier sentencing-prediction work in Korea was pursued as national research and development on the engineering side [22,37]. Commissioning the roadmap’s higher-risk tasks through that mechanism, with joint legal and engineering leadership a condition of award, would make the collaboration a condition of funding rather than a matter of goodwill. Shared leadership between leaders with complementary profiles has been associated with stronger team performance [72], while work on interdisciplinary collaboration emphasizes trust and clear ex ante expectations about roles and credit [73]. This evidence does not test engineer-jurist co-leadership itself, but it does suggest that sharing substantive authority works.

5. Implications and limitations

5.1. Implications

The Korean case carries a wider lesson. When a state institution turns to AI for a core function, responsible deployment depends less on how mature the technology is or how clear the rules are than on whether the research base is ready. For Korean legal scholarship specifically, the constitutional standards of Section 2.3 (the culpability principle, the reason-giving obligation, and judicial independence) must enter Korean LJP as design constraints on dataset label/features and shared baselines rather than as ex post commentary, which is why the engineer-jurist collaboration is the agenda's organizing condition.

The first is that datasets must be built to the Korean procedural order. The dataset label "charge prediction" does not transpose, because it decomposes into three non-interchangeable targets (the prosecutor's charged offense, the court's adjudicated offense, and the guilt, acquittal or dismissal disposition). Features must likewise be built around the statutory sentencing factors, so that guilt and quantum stay separated as the culpability principle requires and attribution lands on categories the reason-giving obligation already recognizes.

The second is that the field needs a shared reference, meaning not only accuracy measures fitted to the ordinal character of sentencing outcomes, but agreed reporting of attribution methods, standardized partitions, temporal validation and label-leakage control. Only against such a common standard can legal and computational analysis accumulate rather than restart with each study. Explainability carries particular weight here. Applied over defined sentencing factors, SHAP and LIME express a model's basis in the same categories a judgment must address. That supports the reason-giving obligation without substituting for the court's own reasons, and makes judicial discretion traceable and reviewable by the Sentencing Commission and on appeal.

5.2. Limitations

First is an asymmetry between the two sides of the comparison. The Korean corpus was assembled through a structured screening procedure conducted as a scoping review rather than a full systematic review, and it was searched to the present in order to capture a fast-moving literature. The global baseline, by contrast, is drawn from published syntheses that are predominantly English-language in scope and whose coverage ends at their respective publication dates [11–13], and this article did not re-execute their searches forward to the present. The comparison is therefore between a current reading of the Korean field and a somewhat earlier reading of the global one. One qualification applies. Where the argument turns on the current state of the art, namely the comparison between general-purpose LLMs and fine-tuned or classical models, targeted searches were extended to recent international work rather than left to the syntheses.

6. Conclusion

Korea's December 2025 Judicial AI Roadmap makes adjudication a live case of governing a coercive public function with AI. It forces a question that is practical, not speculative: is the research base ready? Reading the Korean LJP field against the international one, this article finds that it is not, and that the reason is structural.

Identifying that normative-empirical disconnect is the article's first contribution. Closing that disconnect requires jurists and engineers to work together on dataset design (labels and features) and on shared baselines (accuracy and explainability) that operationalize constitutional and criminal-procedure standards, including the culpability principle, the reason-giving obligation, and judicial independence. The article's second contribution gives one concrete form to that approach by re-conceiving LJP for Korean criminal procedure as three distinct stages (guilt determination, applicable provision, and sentence) showing that the definition of predictive tasks is itself a legally consequential design decision. This collaboration allows constitutional and procedural standards to operate as design constraints rather than ex post safeguards.

Such a framework would also let the judiciary's program build on what a decade of international research has established, rather than relearn it. For a public sector increasingly inclined to govern with AI, the Korean case carries an important reminder. Constitutional and procedural values are not merely safeguards to be applied after an AI system is built. They are part of the architecture by which a legitimate system must be built.

Data availability statement

This study is a comparative literature review and generated no new datasets by the authors. All materials analyzed are previously published, publicly available sources listed in the references. The LBOX OPEN benchmark discussed is openly available at <https://github.com/lbox-kr/lbox-open>. Two files accompany this article. Supplementary data S1 is the PRISMA-ScR workbook underlying the review. Supplementary data S2 is the completed PRISMA-ScR checklist.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used Claude Opus 5 (Anthropic) solely to improve language clarity and readability. All substantive arguments, analysis, and conclusions are the authors' own. The authors take full responsibility for the integrity, originality, and accuracy of the manuscript.

Acknowledgments

This paper is written with support for research funding from aSSIST University.

Authors' contribution

Conceptualization, S.C. and T.O.; methodology, S.C. and T.O.; investigation, S.C.; writing—original draft preparation, S.C.; writing—review and editing, S.C. and T.O.; review of the AI-engineering content, T.O.; supervision, T.O.; funding acquisition, S.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

- [1] Supreme Court of Korea, Judicial Artificial Intelligence Committee. 7th meeting of the judicial artificial intelligence committee held—‘establishment of judiciary AI roadmap and development strategy’ and ‘building a platform to strengthen the judiciary’s AI fundamentals’ (In Korean). 2025. Available: <https://www.scourt.go.kr/portal/news/NewsViewAction.work?gubun=6&seqnum=2902> (accessed on 27 August 2026).
- [2] Aletras N, Tsarapatsanis D, PreotjiucPietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ Comput. Sci.* 2016, 2:e93.
- [3] Zhong H, Xiao C, Tu C, Zhang T, Liu Z, *et al.* How does NLP benefit legal system: a summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 5–10, 2020, pp. 5218–5230.
- [4] Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 28–August 2, 2019, pp. 4317–4323.
- [5] Zhong H, Guo Z, Tu C, Xiao C, Liu Z, *et al.* Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 31–November 4, 2018, pp. 3540–3549.
- [6] Hou Z, Ye Z, Zeng N, Hao T, Zeng K. Large language models meet legal artificial intelligence: a survey. *arXiv* 2025, arXiv:2509.09969.
- [7] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems 30*, Long Beach, USA, December 4–9, 2017, pp. 4765–4774.
- [8] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy* 2021, 23(1):18.
- [9] Constitutional Court of Korea. Decision of 30 July 2009, 2008Hun-Ga24 (vicarious-penalty provision imposing punishment without culpability violates the culpability principle) (In Korean). 2009. Available: https://isearch.ccourt.go.kr/view.do?idx=00&docId=25428_010200 (accessed on 27 August 2026).
- [10] Constitutional Court of Korea. Decision of 29 October 2009, 2009Hun-Ga6 (vicarious-penalty provision imposing punishment without culpability violates the culpability principle) (In Korean). 2009. Available: https://isearch.ccourt.go.kr/view.do?idx=00&docId=27105_010300 (accessed on 27 August 2026).
- [11] Dina NZ, Ravana SD, Idris N. Legal judgment prediction using natural language processing and machine learning methods: a systematic literature review. *SAGE Open* 2025, 15(2):21582440251329663.
- [12] Feng Y, Li C, Ng V. Legal judgment prediction: a survey of the state of the art. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, Vienna, Austria, July 23–29, 2022, pp. 5461–5469.
- [13] Medvedeva M, McBride P. Legal judgment prediction: if you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, Singapore, December 7, 2023, pp. 73–84.

- [14] Kim JH. Gender sensitivity of an AI judge (In Korean). *Justice* 2026, 213:372–415.
- [15] Park S, Park J, Ahn J. Potential applications and implications of GPT-4 in the legal domain: focusing on the Korean Legal Aptitude Test (LEET) (In Korean). *J. Law Econ. Regul.* 2023, 16(1):7–28.
- [16] Kim S. Vertical smaller reasoning language models in the legal domain (In Korean). Master's Thesis, Sogang University Graduate School, 2025.
- [17] Lim EY, Hyun M. Improving the judgment-disclosure system with generative AI: balancing privacy protection and the right to know (In Korean). *J. Intellect. Prop.* 2025, 20(2):93–118.
- [18] Hwang W, Eom S, Lee H, Park HJ, Seo M. Data-efficient end-to-end information extraction for statistical legal analysis. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, Abu Dhabi, UAE, December 8, 2022, pp. 143–152.
- [19] Kwak AS, Jeong C, Lim JW, Min B. A Korean legal judgment prediction dataset for insurance disputes. *arXiv* 2024, arXiv:2401.14654.
- [20] Park W, Cho E, Chang JH, Kim JC. A foundational study on developing a structural model for AI-based sentencing prediction based on violent crime judgment (In Korean). *J. Internet Comput. Serv.* 2024, 25(1):91–98.
- [21] Byun J, Kim H, Park M, Shin JW. Court decision estimation using machine learning algorithms (In Korean). *J. Korean Criminol. Assoc.* 2018, 12(1):3–17.
- [22] Shin JW, Park MR. Development of an artificial intelligence (AI)-based sentencing prediction system for homicide crimes [Final report, national R&D project 2017M3C1B6053698] (In Korean). 2018. Available: <https://dl.nanet.go.kr/detail/MONO12022000024136> (accessed on 27 August 2026).
- [23] Kim JH. A study on the prediction of property claims using the natural language processing technique of divorce case data (In Korean). Doctoral Thesis, Soongsil University Graduate School, 2020.
- [24] Ko SJ, Lee KT, Kim KH, Kim JH. Prediction of compensation costs in apartment housing defects lawsuits using regression analysis (In Korean). *J. Archit. Inst. Korea* 2021, 37(2):197–204.
- [25] Lee YK, Park KR, Lee HY. Tax judgment analysis and prediction using NLP and BiLSTM (In Korean). *J. Digit. Convergence* 2021, 19(9):181–188.
- [26] Hwang W, Lee D, Cho K, Lee H, Seo M. A multi-task benchmark for Korean legal language understanding and judgement prediction. In *Proceedings of Advances in Neural Information Processing Systems 35: Datasets and Benchmarks Track*, New Orleans, USA, November 28–December 9, 2022, pp. 32537–32551.
- [27] Park HM, Kim JH. Multi-task deep learning model with an attention mechanism for ship accident sentence prediction. *Appl. Sci.* 2022, 12(1):233.
- [28] Park YC, Lee J. Deep learning for predicting Korean court judgments based on judges' cognitive reasoning (In Korean). *J. AI Humanit.* 2023, 13:73–107.
- [29] Lee J. Legal text classification in Korean sexual offense cases: from traditional machine learning to large language models with XAI insights. *Artif. Intell. Law* 2025, advance online publication.
- [30] Cho M, Han J. Term-of-penalty prediction for criminal cases using ChatGPT (In Korean). In *Proceedings of Annual Conference of KIPS*, Pyeongchang, Republic of Korea, May 23–25, 2024, pp. 784–785.
- [31] Min H, Yun J, Noh B. Natural language processing-based judgement prediction system for road traffic accidents: focused on text information for traffic accident situations (In Korean). *J. Korean Soc. Transp.* 2024, 42(4):385–397.

- [32] Min H, Noh B. TRACS-LLM: LLM-based traffic accident criminal sentencing prediction focusing on imprisonment, probation, and fines. *Artif. Intell. Law* 2025, advance online publication.
- [33] Choi EK. Sentencing prediction using classification models with legal syllogism prompting (In Korean). Master's Thesis, Sogang University, 2025.
- [34] Jung CW. Development of an NLP-based judgment prediction model for defect litigation in residential buildings (In Korean). Doctoral Thesis, Hanyang University Graduate School, 2026.
- [35] Yang CS. Artificial intelligence and the change of legal system: in case of criminal justice (In Korean). *Korean J. Leg. Philos.* 2017, 20(2):45–76.
- [36] Yang JM. A study on the possibility of replacing judges with artificial intelligence (In Korean). *Hongik Law Rev.* 2018, 19(1):1–29.
- [37] Shin JW. Development of an artificial intelligence judgment prediction system based on legal and social factors [Final report, national R&D project 2017M3C1B6053698] (In Korean). 2018. Available: <https://dl.nanet.go.kr/detail/MONO12022000024119> (accessed on 27 August 2026).
- [38] Chung YK, Song GB. A study on the development of artificial intelligence and the major issues of criminal justice (In Korean). *Korean Police Stud. Rev.* 2019, 18(2):3–24.
- [39] Lee BG. Constitutional issues of AI's predictive powers and recidivism prediction algorithm: consideration from State v. Loomis (In Korean). *Public Law J.* 2020, 21(2):169–191.
- [40] Joo HK, Jung CY. The problem of algorithmic bias and the use of artificial intelligence in the field of normative decision: focused on crime prediction and criminal justice (In Korean). *Chosun Law J.* 2020, 27(1):115–162.
- [41] Seok J, Yeo I, Lee MY, Lee S, Lee J, *et al.* A study on the legal issues of service life cycle in the start-up: focused on judgement prediction services (In Korean). *Hanyang Law Rev.* 2021, 38(3):177–204.
- [42] Kang SK, Kim MS, Kim BP, Jung YS, Park JW, *et al.* Research on legal policy for introducing and responding to legal tech [Research Report 22-B-16] (In Korean). 2022, pp. 1–317. Available: https://www.kicj.re.kr/board.es?mid=a20201000000&bid=0029&list_no=13360&act=view (accessed on 27 August 2026).
- [43] Jeong SY. The European Parliament resolution on the use of artificial intelligence in criminal justice: on the prohibition of decisions made by artificial intelligence (In Korean). *Korean J. Criminol.* 2022, 34(2):137–172.
- [44] An SW, Park GD. A study on the main points of introduction of artificial intelligence (AI) systems in criminal justice procedures (In Korean). *Korean J. Crim. Inf. Stud.* 2022, 8(2):135–155.
- [45] Kim HJ. Artificial intelligence ethics and criminal policy: focusing on the issue of artificial intelligence ethics in criminal justice (In Korean). *Korean J. Criminol.* 2022, 34(1):79–110.
- [46] Yang JM. A study on third-party litigation funding and legal risk prediction using artificial intelligence (In Korean). *Hongik Law Rev.* 2023, 24(1):61–87.
- [47] Rhee GS. Artificial intelligence prediction program in criminal justice system: focused on its biased algorithm in relation to racial discrimination (In Korean). *Wonkwang Law Rev.* 2023, 39(2):57–73.
- [48] Park HJ. Sentencing and artificial intelligence: focusing on public sentiment analysis, risk assessment, and sentencing analysis (In Korean). *Korean Lawyers Assoc. J.* 2023, 72(4):252–293.
- [49] Joo HK. Sentencing decisions using artificial intelligence risk assessment programs in criminal procedure (In Korean). *Yeungnam Law J.* 2023, 57:35–66.

- [50] Kim J. Use of artificial intelligence in the investigation for sentencing: a preliminary review (In Korean). *Korean J. Criminol.* 2024, 36(1):175–219.
- [51] Lee BR. Legal policy study for effective introduction and application of AI and judicial system (In Korean). *Changuijeongchaek* 2025, 10:108–135.
- [52] Lee KM. A methodological study on the use of generative AI in legal organizations (In Korean). *Leg. Theory Pract. Rev.* 2026, 14(1):177–204.
- [53] Paek M. Implications and challenges of AI judges in criminal justice (In Korean). *Korean J. Comp. Crim. Law* 2026, 28(2):1–52.
- [54] Lee BJ. Artificial intelligence (AI) sentencing and the conscience of the judge (In Korean). *Sogang Law Rev.* 2026, 15(1):191–221.
- [55] Son JY. AI fairness and the normative limits of criminal justice system: a comparative analysis of criminal case law. *Rev. Inst. Econ.* 2026, 20(1):23–58.
- [56] Kang J. Governing AI-based empirical criminal sentencing analysis: institutional and technical policy choices. *SSRN* 2026, SSRN:6718438.
- [57] Yang JM. Recidivism prediction algorithm and issues of explainability and fairness (In Korean). *Contemp. Rev. Crim. Law* 2021, 70:207–240.
- [58] Lee SA. Seven problems at hand of an AI judge: comprehensive analysis of legal and technological arguments over recidivism risk assessment algorithms (In Korean). *JURIS* 2023, 1(64):665–714.
- [59] Kim HJ. Could data lie?: generative AI and bias (In Korean). *IP Data Law* 2023, 3(1):75–105.
- [60] CourtListener. *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017). 2016. Available: <https://www.courtlistener.com/opinion/4237997/state-v-eric-l-loomis/> (accessed on 27 August 2026).
- [61] Wang N, Tian MY. “Intelligent justice”: human-centered considerations in China’s legal AI transformation. *AI Ethics* 2023, 3(2):349–354.
- [62] Zou C. Achievements and prospects of artificial intelligence judicature in China. *Chin. Stud.* 2022, 11(4):197–227.
- [63] Zheng GG. China’s grand design of people’s smart courts. *Asian J. Law Soc.* 2020, 7(3):561–582.
- [64] Constitutional Court of Korea. Decision of 26 May 2022, 2021Hun-Ma619 (constitutionality review of the Rules on Attorney Advertising). 2022. Available: <https://isearch.court.go.kr/view.do?idx=01&docId=68101> (accessed on 27 August 2026).
- [65] Ministry of Justice of Korea. Disciplinary decisions against law-platform-registered lawyers overturned [Press release] (In Korean). 2023. Available: <https://www.moj.go.kr/bbs/moj/182/575576/artclView.do> (accessed on 27 August 2026).
- [66] Martínez E. Re-evaluating GPT-4’s bar exam performance. *Artif. Intell. Law* 2025, 33(3):581–604.
- [67] Vats S, Zope A, De S, Sharma A, Bhattacharya U, *et al.* LLMs—the good, the bad or the indispensable?: a use case on legal statute prediction and legal judgment prediction on Indian court cases. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 6–10, 2023, pp. 12451–12474.
- [68] Shui R, Cao Y, Wang X, Chua TS. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 6–10, 2023, pp. 7337–7348.

- [69] Deng C, Mao K, Zhang Y, Dou Z. Enabling discriminative reasoning in LLMs for legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, USA, November 12–16, 2024, pp. 784–796.
- [70] Watson J, Ribeiro de Faria J, Tomalin M, Magnusson M, Xie H, *et al.* Shortcut learning in legal judgment prediction: empirical evidence from the UK Employment Tribunal. *arXiv* 2026, arXiv:2607.04261.
- [71] Wang C, Han B, Patel B, Rudin C. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *J. Quant. Criminol.* 2023, 39(2):519–581.
- [72] Xu H, Liu M, Bu Y, Sun S, Zhang Y, *et al.* The impact of heterogeneous shared leadership in scientific teams. *Inf. Process. Manag.* 2024, 61(1):103542.
- [73] Bennett LM, Gadlin H. Collaboration and team science: from theory to practice. *J. Investig. Med.* 2012, 60(5):768–775.