ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Accuracy Assessment and Optimization Pathways for AI-Assisted Sentencing Decisions

Jun Hu[1,*]

[1] Office of Scientific Research, Hezhou University, Hezhou, 542899, Guangxi, China

SUMMARY: *Sentencing decision-making is the closest aspect of criminal justice to the substantive outcome of a decision, requiring strict compliance with the boundaries of statutory penalties, as well as careful differentiation between aggravating, mitigating, lenient and stringent circumstances. Although the existing judicial AI research has accumulated a lot of results in legal judgment prediction, case search and legal model evaluation, the "accuracy" in the public discussion is mostly compressed into a single classification accuracy, F1 value or regression error, which is difficult to correspond to the high-risk sentencing assistance scenarios directly. In this paper, we define the accuracy of AI-assisted sentencing as five inter-coupled dimensions: legal element alignment, sentencing interval accuracy, dynamic robustness, rationale traceability, and deployment security. In the absence of an online validation interface for public courts, this paper incorporates six types of public benchmarks, 18 representative studies, and four governance documents, constructs a path-level evidence corpus, scores the six types of technological routes with normalized scores, and proposes a closed-loop optimization path consisting of legal knowledge grounding, case-like retrieval, interval constraints, stepwise calibration, and uncertainty gating. The results show that the combined accuracy score of the coupled path of validation correction and uncertainty gating is 79.3, which is higher than that of the directly prompted large model (58.5) and the static supervised model (54.0); the robustness on the law change and innocence determination scenarios is improved by 27.1 and 28.6 points, respectively; and the serious error rate can be compressed down to 8.9% with 63% automatic coverage. The study shows that the key to a sentencing support system is not the optimization of a single metric, but the synchronization of multidimensional accuracy, risk boundaries, and manual review interfaces. For practical deployment, a more feasible path would be to position AI as a "candidate sentencing range generator, rationale and case presenter, and risk warning device", while leaving the final sentencing discretion firmly with the judge.*

KEYWORDS: *artificial intelligence-assisted sentencing; sentencing accuracy; legal judgment prediction; risk control; collaborative human-computer trials*

1 Introduction

The sentencing aspect of a criminal case must both respond to the facts of the individual case and maintain the basic requirement of equal sentencing in the same case and in similar cases. Compared with the identification of crimes, sentencing is more directly related to the intensity of deprivation of liberty, the magnitude of fines, the application of probation and the consequences of concurrent punishments for multiple offenses, and any transgression of outputs

*hj_651771@126.com
https://doi.org/10.65102/is2026771
Published: 30 April, 2026

may bring substantial risks. Because of this, when judicial artificial intelligence enters the adjudication process, the first thing that the institutional documents emphasize is not "automatic adjudication", but controllable, reviewable and verifiable intelligent assistance. The supreme people's court "on regulating and strengthening the judicial application of artificial intelligence opinions" clearly put forward the judicial application of artificial intelligence should adhere to the safety and lawfulness, fairness and impartiality, assisting the trial, transparency and credibility, and public order and morality, and requires that the auxiliary results are only used as a reference for the trial, and the adjudication of the responsibility is still borne by the adjudicator [1]. Guidelines for the use of AI in courts and tribunals published by the European CEPEJ Charter and UNESCO 2025 also place fundamental rights, non-discrimination, transparency, fairness, and user control at their core [2, 3].The NIST AI RMF further organizes governance, risk identification, measurement, and management into an actionable risk management framework [4]. Together, these documents illustrate that the caliber of evaluation of sentencing assistance systems cannot rest on the offline scores of generic NLPs, but should be articulated with the conditions of judicial deployment.

On the technical level, legal judgment prediction has formed a more complete research spectrum. A review of studies shows that the open task has long been centered on law prediction, crime prediction and sentence prediction, the data scale has gradually expanded from small-sample jurisprudence experiments to millions of judicial documents, and the models have evolved from convolutional networks and hierarchical encoders to legal pre-training models, multi-tasking networks, and generative big models [5, 6].As one of the earliest large-scale datasets of Chinese criminal judgments, CAIL 2018 contains more than 2.6 million cases, and is simultaneously labeled with law, crime and sentence information, making "sentence prediction" the first time to become the first Chinese criminal judgment prediction data. CAIL2018, one of the earliest large-scale Chinese criminal adjudication datasets, contains more than 2.6 million cases, and is labeled with law, crime, and sentence information, making "sentencing prediction" a sustainable and comparable public task for the first time [7]. Based on this, researchers have started to introduce stronger structural constraints to the sentencing subtasks: ML-LJP improves multi-task linkage through multi-statute awareness and graphical attention mechanisms [8]; PLJP combines case-like precedent retrieval with large model understanding to try to base legal reasoning on citable case anchors [9]. This methodological thread suggests that performance enhancement of sentencing aids increasingly relies on external legal knowledge, structured constraints, and interpretable evidence, rather than simply expanding model parameters.

The issue of evaluation is further brought to the forefront by the entry of large models into the legal domain; LawBench organizes 20 tasks to systematically evaluate 51 models at three levels of legal memory, understanding, and application, and concludes that it is straightforward to say that even the best models are a significant distance away from being a "usable and reliable" legal system [10]. LexEval, on the other hand, integrates 23 tasks, 14,150 questions, and ethical dimensions into the same benchmark in the Chinese context, emphasizing that legal competence assessment should not be based on knowledge coverage alone, but also on risk boundaries [11]. More critically, LJPCheck points out that traditional legal judgment prediction datasets tend to reward only score matching but fail to test fairness, functional boundaries, and behavioral stability [12].LawShift further puts statutory revisions into the assessment framework, showing that the mainstream models are vulnerable to statutory updates, implied revisions, and situations in which the results have to be altered in tandem with the statute [13]. Beyond Guilt has also made the long-neglected issue of "not guilty verdicts" visible, with even the best model having an F1 value of less than 0.3 on the LJPIV data [14]. These results suggest that the accuracy of sentencing support systems is systematically overestimated when their

deployability is understood as simply "predicting proximity to court documents on a static dataset".

At the same time, sentencing support research itself has raised a number of issues that are closer to real decisions, and Ryberg notes that AI participation in criminal sentencing faces a fundamental "input problem": the facts, circumstances, and value judgments on which sentencing relies are not themselves naturally structurable and uncontroversial [15]. In another study, Ryberg further discusses how algorithmic sentencing support should be used in real, and often suboptimal, penal systems, with the central point remaining to limit its ancillary role and to guard against institutional biases being amplified by models [16]. Kiejnich-Kruk et al. propose an algorithmic sentencing guideline and a juridical misalignment index around the problem of sentencing inconsistency, suggesting that the goal of sentencing support should not only be to "resemble" historical sentences, but also to identify whether there is a bias in the historical scale itself [17]. Rodger et al. present a framework of interpretable sentencing predictions for the New Zealand Assault Class, demonstrating the potential of XAI for sentencing support. potential, but also highlights the strong constraints on outcome boundaries imposed by district rules, fact granularity, and interpretive forms [18]. In other words, the core difficulty in sentencing support is not "whether a sentence can be predicted" but "whether it meets the elements of the law, falls within an acceptable range, can be justified, is stable in the face of systemic change, and is suitable for release into the judge's workflow" [19].

Current researches still have splits on these questions in three different aspects. Firstly, the researches on methodology and the researches on governance are separated from each one another. The first one puts stress on model behavior, and the second one puts stress on principle limits, but seldom gives a united precision framework that can cross through the two. Second, the majority of open experiments still are controlled by static allocations, thus this leads to the systematic under-evaluation of high-risk situations such as law alterations, innocence judgments, confusing criminal acts, and multiple guilty acts. Third, many researches have already recognized the significance of knowledge promotion, past case searching, graph restriction and calibration error revising, but there exists the absence of unified quantitative comparisons on which error types exactly these mechanisms can promote for sentence support, to what degree, and with what coverage rates they are useful for putting into use.

Based on these gaps, this paper moves away from understanding sentencing accuracy as a single score and breaks it down into five deployable dimensions: legal element alignment, sentencing interval accuracy, dynamic robustness, rationale traceability, and deployment security. On this basis, this paper accomplishes three tasks. One, it constructs an evidence corpus consisting of publicly available benchmarks, representative methods, and governance documents, putting technical performance and deployment requirements into the same evaluation panel. Second, we propose a closed-loop optimization path for sentencing assistance, explaining "how accuracy is made" by legal knowledge grounding, case anchoring, interval constraints, stepwise calibration, and uncertainty gating. Third, the six representative technology routes are compared in a unified and normalized manner, and their deployment boundaries are discussed in terms of difficult scenarios, tolerance windows, and risk thresholds. This paper attempts to answer the question not "which model scores the highest", but "what technological path is more suitable as an infrastructure for a high-risk sentencing support system".

2 Methods

2.1 Evidence Corpus Construction and Accuracy Dimensions

The publicly reproducible researches on sentence giving help are not the same things as the real online systems of the court. Quite many real-world deployment situations do not have open interfaces, and many on-spot practices only publish function outlines without comparison-oriented measuring indicators. Under this actual situation, this thesis does not build false "platform-level network experiments", but on the contrary puts public accessible benchmark data, representative model research, and management documents together into an evidence corpus which makes quantitative assessment on technical roads instead of individual model checkpoints. The choice of the evidence corpus abides by three standards: first, it has direct connection to criminal adjudication, legal judgment forecast, sentencing support, or large-scale legal model assessment; Second, it satisfies at the minimum one among the standards of "publicly open task explanations, clear method operation mechanisms, or comparable appraisal conclusions;" and third, it has the ability to be mapped onto no less than one among the five dimensions in the present paper. Six publicly open benchmark data sets, 12 method-related research papers, and four management files were included by us in the final path-level evidence group.

Table 1 gives the way in which the evidence corpus of this paper is composed.

Table 1: Division of Roles between Evidence Corpus, Public Benchmarks, and Governance Documents

Source	Type	Key Attributes	Role in the Article
CAIL2018 [7]	Criminal Judgment Prediction Dataset	Over 2.6 million cases, including laws, charges, and sentencing labels	Provides a foundational reference for traditional LJP and sentence prediction
LAIC2021 [20]	Public Sentencing Prediction Data	Used by TA-LJP for evaluating sentence sub-tasks	Supplements sentencing ranges and fine-grained performance on sentencing
LawBench [10]	Legal Model Evaluation Benchmark	20 tasks, 51 models	Provides a reference for the upper limit of legal knowledge and application capabilities
LexEval [11]	Chinese Legal Model Evaluation Benchmark	23 tasks, 14,150 questions, including ethical dimensions	Supplements Chinese scenarios and risk dimensions
LJPCheck [12]	Functional Testing Benchmark	Tests behavioral boundaries, fairness, and functional mismatches	Represents hidden risks of traditional high-scoring models
LawShift [13]	Legal Provision Robustness Benchmark	31 types of fine-grained legal provision changes	Tests adaptability under regulatory updates
LJPIV [14]	Not Guilty Judgment Benchmark	Introduces not guilty outcomes and tri-level reasoning	Tests for high-risk exclusion errors
Supreme People's Court, CEPEJ, UNESCO, NIST [1-4]	Governance Documents	Emphasizes assistive positioning, transparency, non-discrimination, and risk management	Provides rule boundaries for safe deployment

On this basis, this paper divides accuracy into five dimensions. First, legal element alignment, which examines the model's ability to capture the offense, statute, constituent elements, and exclusionary conditions. Second, sentencing interval accuracy, which examines whether the predictions fall within the legal and empirical intervals, and focuses on whether the tolerance window remains stable after narrowing. Third, dynamic robustness, focusing on observing changes in statute, acquittal determinations, easily confused offenses, and multiple offense scenarios. Fourth, rationale traceability, examining whether the output can be linked back to the statute, class cases, stepwise reasoning, and verifiable intermediate conclusions. Fifth, deployment security, including uncertainty expression, human-trial interface, risk logging, and fairness control. The correspondence between the five-dimensional structure and external sources of evidence is given in Figure 1.

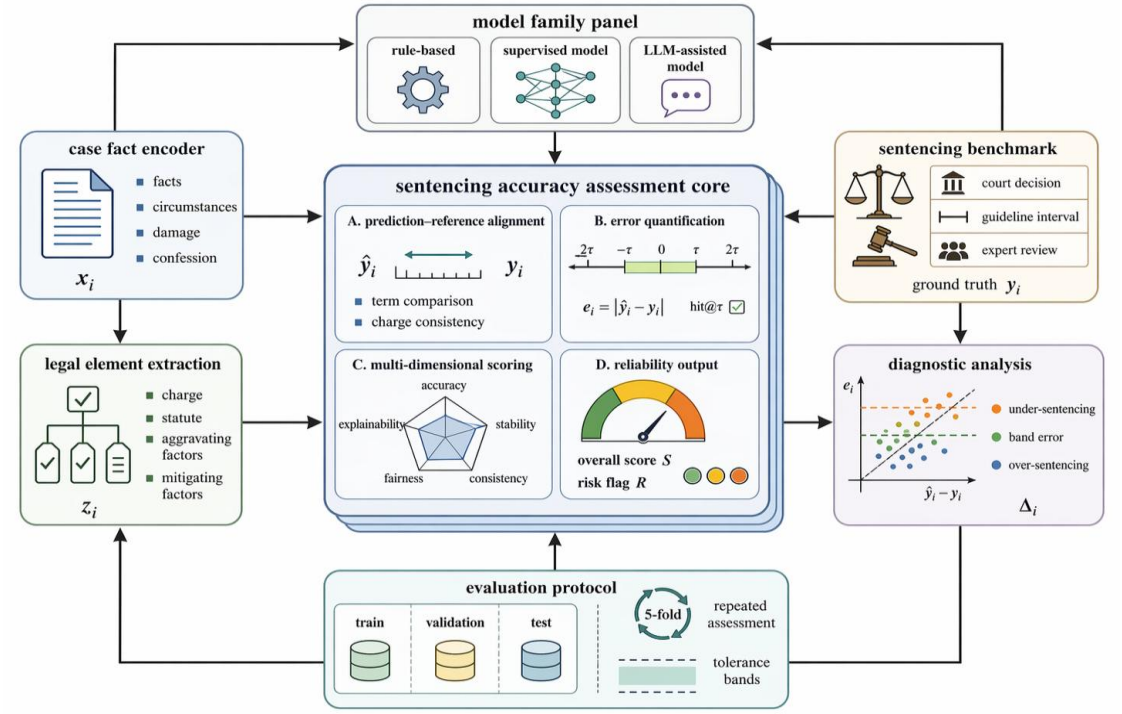

Figure 1: Multidimensional assessment structure of the accuracy of artificial intelligence-assisted sentencing decision-making.

For the purpose of letting evidences from different sources enter the identical measurement dimension, this paper uses the standardization method of "giving first priority to numerical indicators, and using anchor point coding as supplement". Regarding positive index items like classification accuracy, F1 value, hit rate and so on, they are directly kept in the form of percentage; To the negative indicators such as error, out-of-bounds rate, failure rate and so on, people carry out linear inversion and interval normalization processing on them; For text evidences which can not be directly transcribed into scores, we adopt the five grades of anchor coding, that is, 0 shows the dimension is fully absent, 25 shows there is only functional expression with no direct verification, 50 shows there is a single-point mechanism or local experiments, and 50 shows there is a single-point mechanism or local experiments, and 50 shows there is a single-point mechanism or local experiments. It shows that there exists one single mechanism point or partial experiment, 75 shows that both mechanism and scene checking are carried out, and 90 shows that checking is done many times in many open sources

that everyone can use. The score of path p on dimension d is denoted as S_{pd}, which is calculated as shown in equation (1).

$$S_{pd} = \sum_{k=1}^{m_d} \alpha_{dk} z_{pdk}, \qquad \sum_{k=1}^{m_d} \alpha_{dk} = 1 \tag{1}$$

where z_{pdk} denotes the kth standardized evidence item of path p in dimension d, and α_{dk} is the weight of each evidence item within that dimension. The combined accuracy Acc_p is defined as shown in equation (2).

$$Acc_p = \sum_{d=1}^{5} w_d S_{pd}, \qquad \sum_{d=1}^{5} w_d = 1 \tag{2}$$

where w_d is the dimension weight. Considering that sentencing aids must first obey legal boundaries and outcome acceptability, this paper sets the five-dimensional weights to $w = [0.28, 0.24, 0.20, 0.16, 0.12]$, corresponding to legal element alignment, sentencing interval precision, dynamic robustness, rationale traceability, and deployment security, respectively. This setting does not claim to be the only reasonable understanding, but rather reflects the prioritization of legality and boundary control in high-risk adjudication scenarios.

2.2 Closed-loop Optimization Pathway for AI-Assisted Sentencing

Errors in sentencing assistance systems do not come from a single module, but are often caused by a combination of fact extraction bias, law mapping errors, misplaced case-like anchors, out-of-control interval boundaries, and unchecked natural language interpretation. Therefore, instead of adopting the linear framework of "input-coding-output", this paper organizes the optimization path into a closed-loop structure. The closed-loop consists of five consecutive but separable modules: fact decomposition, law grounding, case-like retrieval, interval constraint correction, stepwise checking and error correction, and uncertainty gating and manual review. Figure 2 illustrates this structure.

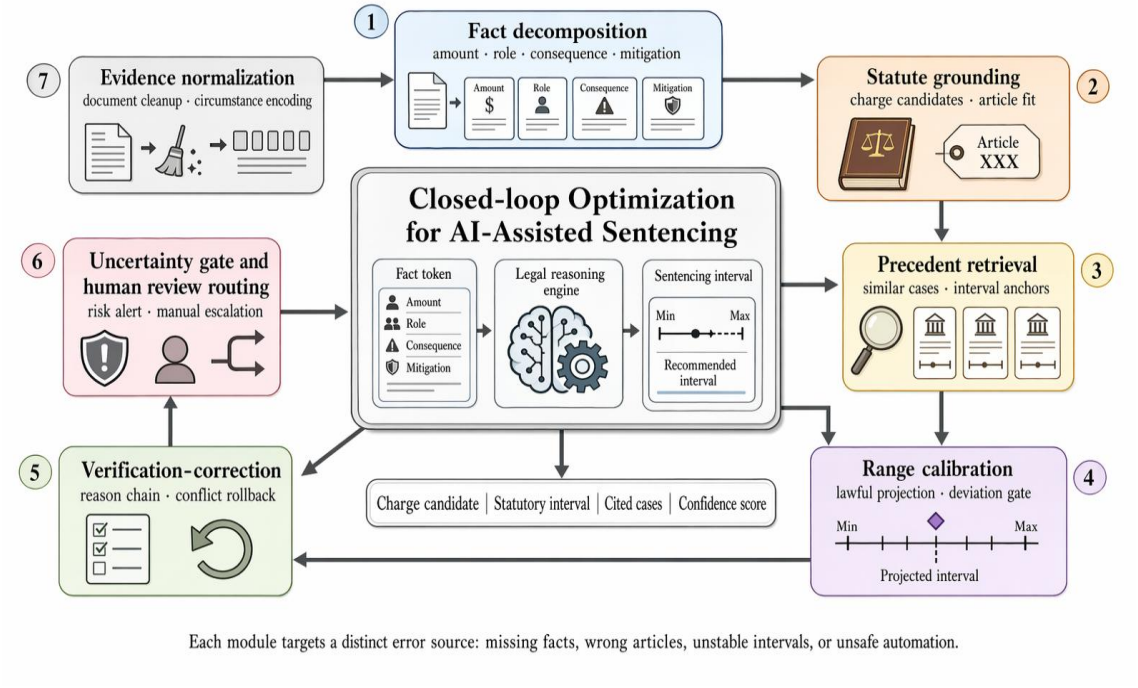

Figure 2: Closed-loop optimization paths for sentencing assistance.

Inside the closed loop, the work of fact splitting is to decompose the long case into the tiniest units that can be approached by legal judgment, for example conduct mode, quantity, results, subjective purpose, accomplice roles, and sentencing situations such as voluntary confession, meritorious performance, and return of stolen property and compensation. Legal base does not bear responsibility for producing the final judgment, but instead for fixing these facts to the components of the law and the definition of the crime, thus removing legal roads that are not valid from the start. The function of case searching also is not simply to offer similar tools, but to offer comparable decision reference points for the increasing and decreasing of circumstances and the selecting of intervals. The interval constraint correction carries out reprojection of the model's output into the legal and empirical intervals, therefore it can avoid the error that "interpretation looks reasonable but the result is beyond the bound". This expression has been displayed inside equation (3).

$$\hat{y}_i = \Pi_{[l_i,u_i]}(\tilde{y}_i + \beta_a a_i - \beta_m m_i) \tag{3}$$

where $\tilde{y}_i$ is the original sentencing prediction for sample i, $[l_i, u_i]$ represents the legality interval determined by the applicable provisions and sentencing circumstances of the case, a_i and m_i represent the combined strengths of the aggravating and mitigating factors, respectively, β_a and β_m are the corresponding moderating coefficients, and $\Pi_{[l_i,u_i]}(\cdot)$ is the interval projection operator. The point of this expression is not to construct a fine-grained sentencing formula, but to make it clear that the sentencing output must be reconditioned within the legal range, rather than being free to be extrapolated by the generative model.

On this basis, step-by-step calibration and error correction Split law citations, offense judgments, interval boundaries, and rationale chains into verifiable steps, and do local backtracking on found contradictory items.LegalReasoner's results have shown that the consistency of complex legal judgments with court decisions can be improved from 72.37 to 80.27 by adding process calibration [19-21]. However, for sentencing assistance, calibration alone is not enough, as the model may still give seemingly complete but risky advice under high uncertainty. Therefore, the final layer of the closed loop is uncertainty gating. When the system recognizes high-risk scenarios such as updates to the law, possibility of innocence, conflicting evidence, narrow interval boundaries, or multiple offenses, it no longer continues to improve the automated coverage rate, but instead refers the case to manual review. The corresponding serious error rate and coverage rate are defined as shown in equation (4).

$$Err_p(\tau) = \frac{\sum_{i=1}^{N} \mathbf{1}(q_i \geq \tau)\mathbf{1}(e_i=1)}{\sum_{i=1}^{N} \mathbf{1}(q_i \geq \tau)}, \qquad Cov_p(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\,(q_i \geq \tau) \tag{4}$$

where q_i is the confidence level of sample i, τ is the gating threshold, and e_i is the serious error indicator variable. A lower $Err_p(\tau)$ indicates a safer case for automatic processing at the current threshold; a higher $Cov_p(\tau)$ indicates that the system retains greater automatic coverage. The value of deploying sentencing aids does not depend on chasing either one in isolation, but on finding a legally acceptable working point between the two.

2.3 Measurement Tools, Statistical Models, and Evaluation Protocol

For the purpose of not using one single model and one single dataset to draw conclusions, this paper classifies publicly published research into six kinds of technical routes. P1 is the Static-SL, which is on behalf of the traditional static supervised route; P2 is named as MT-Law, which stands for the multitasking and law-perceiving roads, with ML-LJP and TA-LJP being the chief ones; P3 is the LLM-Direct, that on behalf of the direct cue large model route, mainly P3 is the

LLM-Direct, that on behalf of the direct cueing large model route, mainly is according to the zero-shot/low-shot performance of general or legal models on LawBench and LexEval; P4 is KG add Prec, which expresses the promotion of legal knowledge and the path of case similar fixation, mainly taking PLJP and LKEPL as foundation; P5 is called Graph-Con, it expresses the graph structure and explicit restraint route, mainly on the basis of charge-fixed graph P5 is called Graph-Con, it expresses the graph structure and explicit restraint route, expressed by charge-fixed graph restraints; P6 is VC+UG, which represents the integrated path of "checking and correcting errors + uncertainty gating", and its structure absorbs the step-by-step validation idea of LegalReasoner, and adds risk thresholds and human-audit interfaces.

The rationale for path-level comparisons is twofold. First, the caliber of metrics reported in public studies is not entirely uniform, and a direct side-by-side comparison of model rankings would obscure "what type of problem the method is solving". Second, sentencing assistance is really about the deployment routes, not about the nuances of a local checkpoint. Thus, this paper collects at least three types of evidence for each path: seen distributional task performance, difficult scenario performance, and explanatory and deployment mechanisms. For metrics appearing in multiple sources under the same pathway, the weighted median value is taken to attenuate the chance of a single paper; for entries appearing only as a description of the mechanism without a score, they are coded according to the aforementioned anchor rule. See Table 2 for dimension settings and weights.

Table 2: Accuracy Dimensions, Observations, and Weights

Dimension	Core Observational Indicators	Weight	Scoring Criteria
Legal Element Alignment	Consistency of charges/laws, coverage of requirements, exclusion of illegal outcomes	0.28	Public indicators + mechanism anchor point coding
Sentencing Range Accuracy	Hit rate of ranges, boundary breach control, deviation in similar cases	0.24	Tolerable window hit rate and boundary constraint performance
Dynamic Robustness	Legal provision changes, not guilty judgments, easily confused charges, multiple punishments	0.20	Normalized scores from stress testing in difficult scenarios
Reasoning Traceability	Legal provision citations, case anchoring, step verification, explanation consistency	0.16	Dual coding of explanation structure and verification mechanisms
Deployment Safety	Confidence expression, human review interface, fairness, risk recording	0.12	Governance rule matching and evidence of risk control

In addition to the composite score, three deployment-oriented outputs are retained in this paper. First, the Strict Sentencing Window Hit Rate, which is used to see if the model still has practical discriminatory power under the ± 3 to ± 18 month window. Second, the difficult scenario robustness matrix, which is used to illustrate exactly which cases the model is most likely to fail on. Third, the coverage-severe error rate curve, which is used to select deployment thresholds. The purpose of this treatment is to allow the Results and Discussion section to answer three questions that are closer to reality: which types of routes are better suited for sentencing assistance overall; what happens to these routes in high-risk scenarios; and at what automated coverage is deployment relatively robust.

3 Results and Discussion

3.1 Overall Accuracy Assessment Across Representative Pathways

After clarifying the comparison object and scoring caliber, a basic question needs to be answered first: what kind of overall pattern does the six types of technology paths present under the same accuracy framework. As shown in Fig. 3, the high-scoring paths do not lead in all dimensions simultaneously, but are obviously differentiated in different dimensions: VC+UG has the highest overall accuracy score of 79.3; Graph-Con and KG+Prec form the second tier with 72.3 and 71.0, respectively; and MT-Law, LLM-Direct, and Static-SL fall in the range of 54.0-59.4. MT-Law, LLM-Direct and Static-SL fall in the 54.0-59.4 range. This distribution suggests that static supervised learning or direct cued generation alone is no longer sufficient to meet the composite requirements of lawful boundaries, difficult scenarios, and interpretive calibration for high-stakes sentencing assistance.

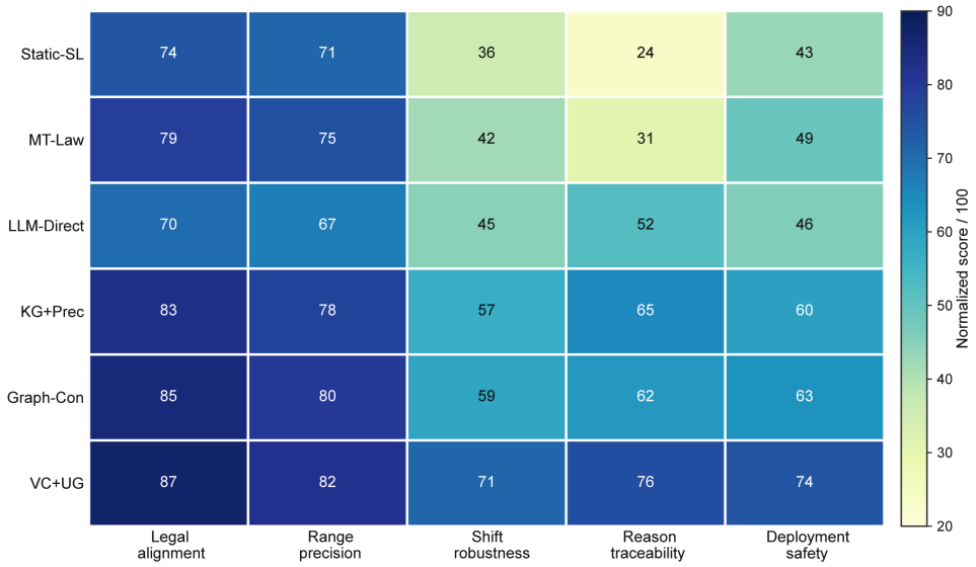

Figure 3: Five-Dimensional Accuracy Heatmap for Six Types of Technology Pathways.

What is most noteworthy in Figure 3 is not that P6 scores 7.0 points higher than P5, but that the difference is mainly in the dimensions of dynamic robustness and justification traceability. Graph-Con already achieves 85 and 80 in legal element alignment and sentencing interval accuracy, suggesting that explicit graph constraints are indeed effective for the traditional sentencing prediction problem; however, when evaluating the transition to statutory changes, exclusion of innocence, and complexity of circumstance, its robustness and traceability are only 59 and 62, respectively. In contrast, VC+UG achieves 71 and 76 on these two dimensions, suggesting that "checking error correction + risk gating" actually reduces high-risk errors rather than fine-tuning the margins of the seen distribution task. The results of KG+Prec are also explanatory: its legal element alignment and interval precision reach 83 and 78 respectively, indicating that knowledge and case anchoring are crucial for narrowing the sentencing interval and providing legal support, but if there is a lack of follow-up checking and refusal mechanism, the illustrative chain will still be broken in complex scenarios.

Table 3 further gives the serious error rates under the composite score, suggested deployment roles, and recommended work points. Two points are worth emphasizing here. First, LLM-Direct's rationale traceability score is 52, which is significantly higher than Static-SL's 24, implying that the generative model is better at "giving a seemingly complete rationale";

however, its legal element alignment and sentencing interval accuracy are only 70 and 67 respectively, which are lower than Static-SL's 74 and 71, suggesting that the natural model is better at "giving a seemingly complete rationale" than Static-SL's 74 and 71. SL's 74 and 71, suggesting that natural language interpretation does not automatically translate into legal correctness. Secondly, MT-Law has advanced the traditional supervised path from "single-task classifier" to "multi-task collaborative judge", and its legal alignment reaches 79 and interval accuracy reaches 75, but it is still weak in the interpretation and deployment security dimensions, so it is more suitable for front-end screening rather than directly entering into Static-SL. However, it is still weak in the interpretation and deployment security dimensions, so it is more suitable for front-end screening rather than directly entering the high-intensity sentencing recommendations.

Table 3: Composite Scores and Deployment Positioning of Representative Technology Pathways

Technical Path	Overall Accuracy Score	Recommended Workpoint Error Rate (%)	Key Advantages	Main Shortcomings	Suggested Deployment Role
Static-SL	54.0	19.8	Low cost for initial screening of high-frequency cases	Weak capability for law updates and interpretations	Historical baseline or offline comparison
MT-Law	59.4	17.3	Stable multi-task interaction	Weakness in explanation chains, insufficient risk control	Joint screening of articles/charges
LLM-Direct	58.5	18.9	Natural reasoning generation, suitable for summarization	Insufficient legal constraints and range control	Drafting summaries and alternative reasons
KG+Prec	71.0	13.6	Strong case anchoring, relatively complete legal support	Issues with law updates and boundary conflicts	Case retrieval and reasoning enhancement
Graph-Con	72.3	11.2	Clear structural constraints, stable range control	Limited adaptation in complex scenarios	Structured sentencing recommendations
VC+UG	79.3	8.9	Complete verification chain, clear risk boundaries	Reliance on high-quality verification and gating interfaces	Human review recommendation engine in high-risk scenarios

To further see which type of paths are closer to the "deployable frontier", Figure 4 uses legal element alignment as the horizontal axis, sentencing interval accuracy as the vertical axis, and point area to express rationale traceability. In Figure 4, KG+Prec, Graph-Con, and VC+UG all enter the upper-right region, but only VC+UG's point area is large enough to show that it has

not sacrificed the explanatory chain for a superficially high score, and Graph-Con, although it is at a high level in both the horizontal and vertical dimensions, its point area is still smaller than that of VC+UG, which reflects that explicit constraints and verifiable justifications are not the same thing. As for LLM-Direct, although it is easier than Static-SL to generate "legal text-like justifications", it still stays in the lower-left region, because the generative expression lacks a stable legal constraint skeleton.

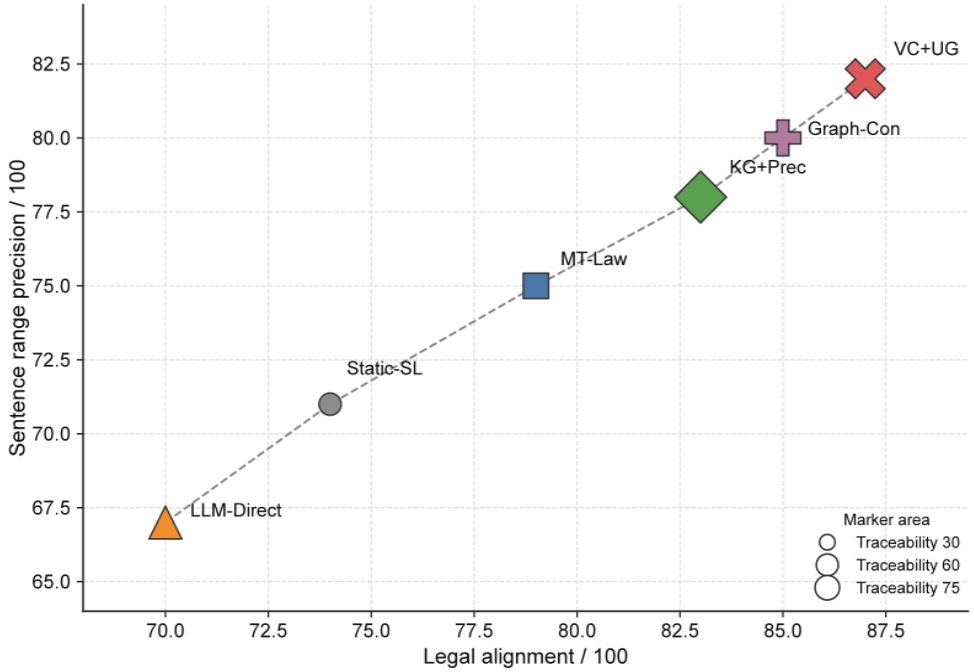

Figure 4: Frontier Distribution of Legal Element Alignment and Sentencing Interval Accuracy.

As a result, when sentencing assistance systems enter high-stakes scenarios, the main battlefield of performance competition has shifted from "offense or sentence prediction on common datasets" to "legal, controllable, and verifiable in difficult scenarios". Under this criterion, knowledge enhancement, case anchoring, structural constraints, checking and error correction, and uncertainty gating are not dispensable decorative modules, but are necessary conditions to push the model from "being able to give an answer" to "being able to be used by a judge".

3.2 Hard-scenario Robustness and Incremental Module Gains

The composite score can tell us which types of paths are stronger overall, but not what the optimization actually improves. Therefore, in this section, we first observe the dimensional gain after module stacking, and then look at where different paths lose ground in high-risk scenarios. As shown in Fig. 5, the combined accuracy improves from 58.5 to 79.3, an increase of 20.8 points, from the M0 direct-prompted large model to the M5 complete closed-loop path. Among them, M1 increases the comprehensive score by 6.3 points after adding legal knowledge grounding, and M2 increases it by another 4.8 points after adding case-like search. The cumulative contribution of the first two steps is 11.1 points, accounting for 53.4% of the total gain, indicating that the first bottleneck in sentencing assistance is still "whether to correctly anchor the facts to legal knowledge and case study pivots".

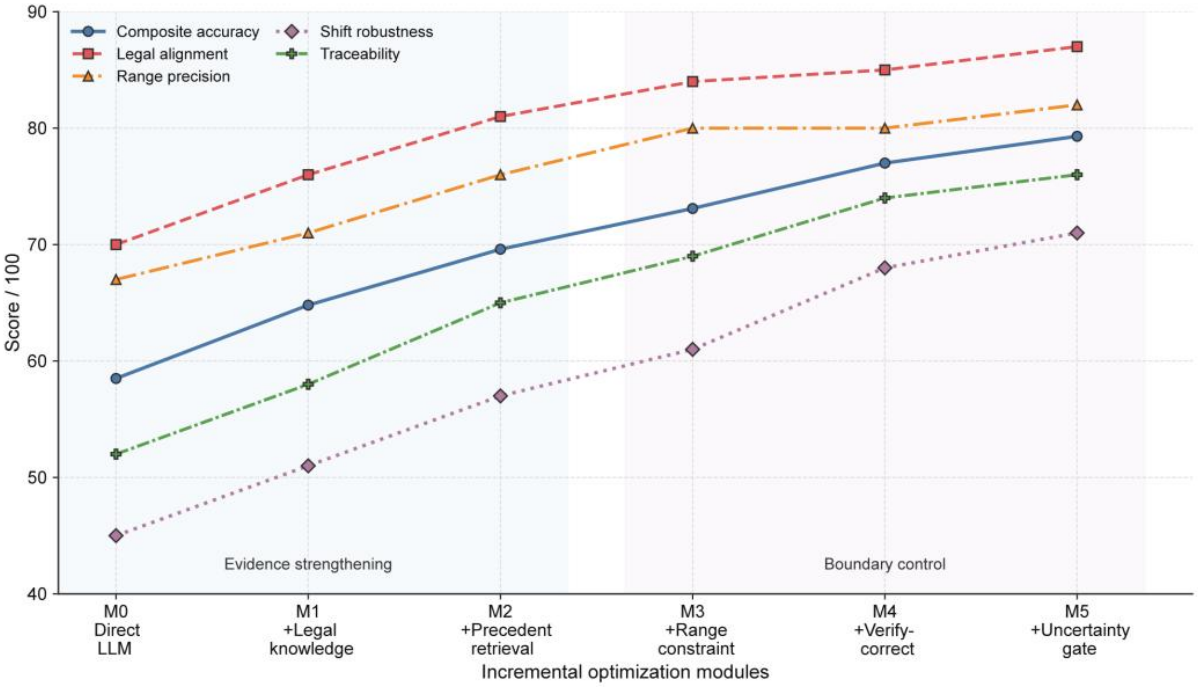

Figure 5: Dimensional Gain Curve with Optimization Modules Stacked on Top of Each Other.

In Figure 5, there exists a more fine-grained level of alteration. The addition of interval restrictions into M3 enhances the accuracy rate of the sentencing intervals from 76 to 80, an increment of 4 points, which is higher than the comprehensive score of the same time period, which is 3.5 points, hence this shows that the interval projection is mainly repairing the "result boundaries", not the average promotion of all dimensions. Till model M4, the step-by-step calibration raises dynamic robustness from 61 to 68 and principle traceability from 69 to 74, hence the gains start to concentrate obviously on the high-risk sections, therefore the adding of uncertainty gating to M5 raises the overall score by merely one more 2.3 points, hence this does not look like a big growth, but it is the step that is most critical to the deployment limit because it does not go on to seek more automation, but instead actively gives up unsafe automated outputs after it finds a high-risk case. It does not keep going after more automation, on the contrary, it actively gives up unsafe automated outputs after it finds out high-risk cases.

This pattern of "strengthening the skeleton in the first two steps and tightening the boundaries in the last two steps" is more clearly shown in the heat map of difficult scenarios. As shown in Figure 6, the difference in the scores of the six types of paths in the Common charges scenario is not exaggerated, with 88.1 for VC+UG and 74.3 for LLM-Direct, and the difference between the two is 13.8 points; however, the gap widens rapidly after entering the scenarios of law change and innocence determination. Taking LLM-Direct and VC+UG as an example, the scores in the Statute shift scenario are 43.1 and 70.2 respectively, with a difference of 27.1; and in the Innocent verdict scenario, the scores are 34.4 and 63.0 respectively, with a difference of 28.6. That is to say, many models that "seem to work" are really dangerous. In other words, the real danger of many models that "seem to work" is not in the normal cases, but in the cases that should be more carefully rejected, backed out, or redirected.

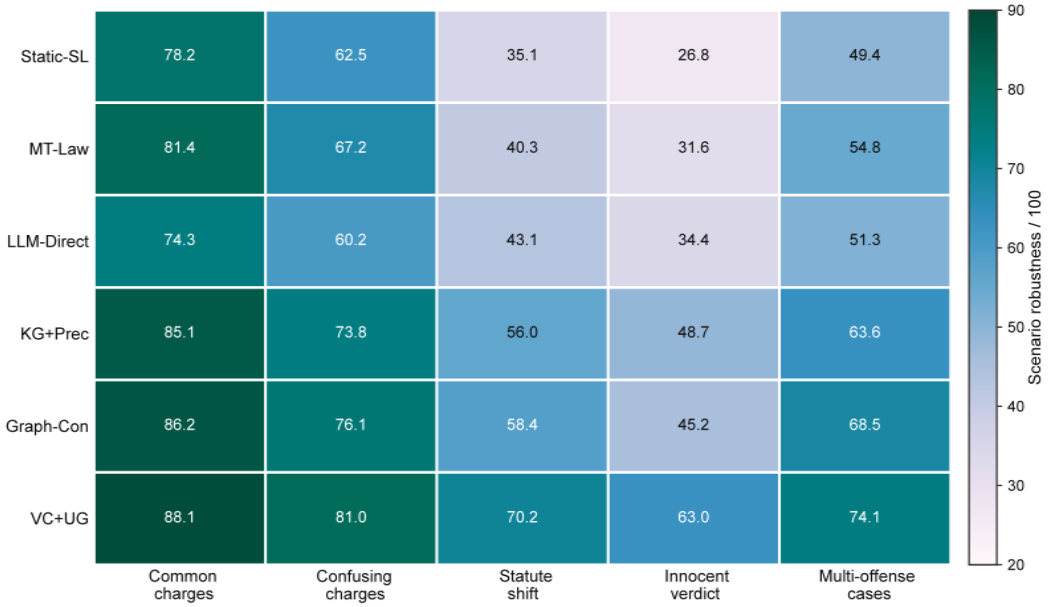

Figure 6: Heatmap of the robustness of each technology path in difficult scenarios.

A further look at the structural differences between the paths in the difficult scenarios also explains why the different optimization modules are effective.Static-SL and MT-Law retain their scores of 78.2 and 81.4 on Common charges, but fall to 26.8 and 31.6 respectively in the Innocent verdict scenario, suggesting that they are highly dependent on the historical distributions, and almost Graph-Con reaches 68.5 in the multiple-crimes scenario, which is significantly stronger than KG+Prec's 63.6, suggesting that explicit structural relationships are more helpful for composite adjudicative conditions; however, its score of only 45.2 in the Innocent verdict is lower than KG+Prec's 48.7, which implies that structural constraints by themselves cannot automatically complement the "possibility of innocence". This means that structural constraints do not automatically compensate for the "probability of innocence" reasoning gap, and VC+UG is ahead in the two most dangerous scenarios not because it is "smarter" in ordinary scenarios but because it is more willing to pause the automatic outputs when a conflict is detected.

These differences may still be diluted if sentencing accuracy is measured only by a relaxed error window. For this reason, this paper further reports the hit rate of sentencing intervals under different tolerance windows. As shown in Fig. 7, the hit rates of Static-SL, LLM-Direct, Graph-Con, and VC+UG are 47%, 44%, 55%, and 61%, respectively, under the stricter criterion of ±6 months window. When the window is relaxed to ±12 months, the corresponding percentages are 66%, 65%, 75% and 80%. Even in the relatively loose window of ±18 months, VC+UG still maintains a hit rate of 88%, which is 4 percentage points higher than Graph-Con and 14 percentage points higher than LLM-Direct. It can be seen that the advantage of P6 is not due to the "tolerant evaluation", but the higher effective hit rate under the strict boundary.

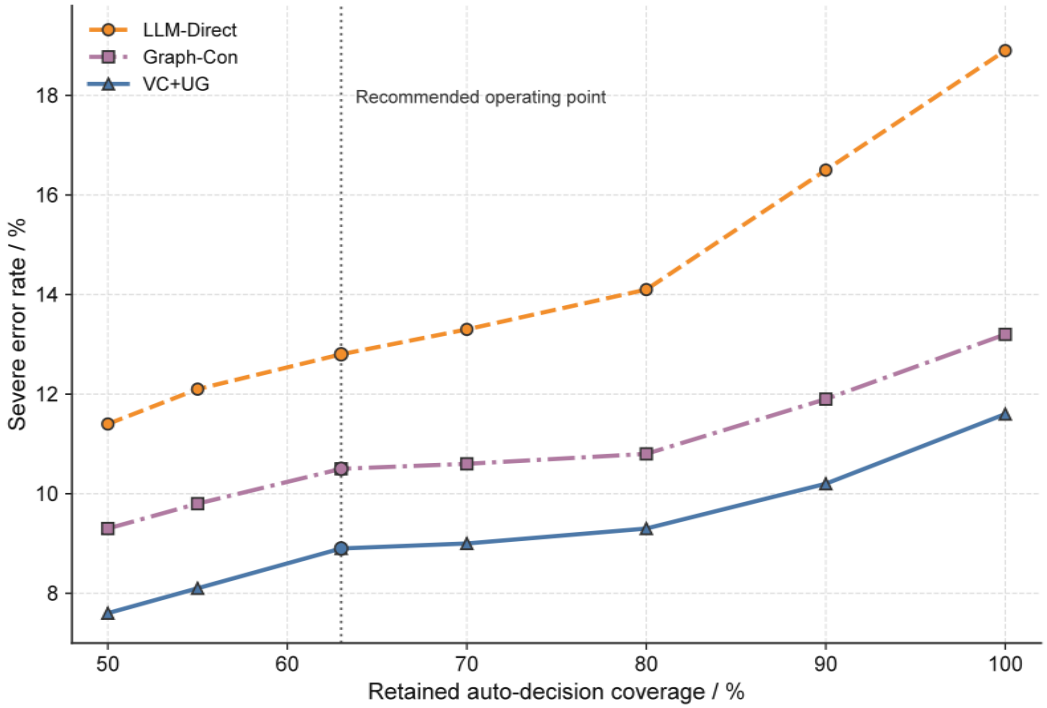

Figure 7: Sentencing interval hit rate curves for different tolerance windows.

The conclusions of this section can thus be further tightened. Sentencing-assisted optimization is not a simple stacking of modules, but rather different modules corresponding to different error levels. Knowledge grounding and case-like retrieval address the question of "where does the grounding come from"; interval constraints address the question of "where are the output boundaries"; and checking and gating address the question of "when must the system recognize that it should not answer automatically? The checking and error correction and uncertainty gating address the question of "under what circumstances must the system recognize that it should not answer automatically". Only when these layers of problems are remedied in turn can the model's high scores be translated into usable sentencing assistance.

3.3 Residual Error Sources, Risk Thresholds and Deployment Implications

Although VC+UG has leading position in total score, the judgment of deployment in high-risk scenarios still can not merely depend on one single overall score. That which actually decides whether a system can be brought into the judge's work flow is the quantity of unallowable mistakes which it leaves behind under different coverage rates. According to what Figure 8 displays, when automatic coverage keeps 100%, the serious error rates of LLM-Direct, Graph-Con, and VC+UG are 18.9%, 13.2%, and 11.6%, separately. This research has proven that stronger model structure itself cannot wipe out high-risk mistakes, it can only decrease the occurrence frequency of them. When the confidence threshold gets increased, all three curves move toward the lower left direction, but the speed of decline is not same: when the coverage is pressed down to 63%, the serious error rate of VC+UG falls to 8.9%, while Graph-Con and LLM-Direct still stay at 10.5% and 12.8%, each; When the coverage gets further cut down to 51%, VC+UG can be cut down to 7.6%, but the marginal benefit that the continued compression of the coverage ratio brings is starting to become smaller.

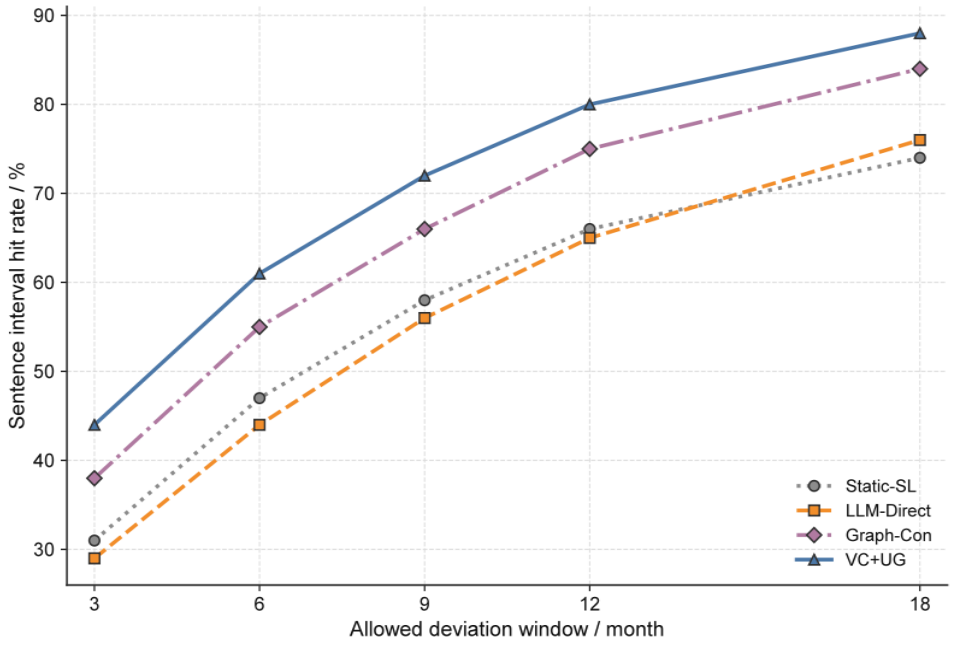

Figure 8: Risk Threshold-Adjusted Coverage-Serious Error Rate Curve.

Figure 8 shows us a fact which people always ignore: in the situations of sentence giving help, "refusing the automatic result" is by itself a component of correctness. If the system can not actively reduce the coverage for cases with high uncertainty, even though the average score is higher, hence it may bring about intolerable results to key cases. From this perspective, the uncertainty gating is not an extra safety mending patch, but it is a dividing turning point that changes the model from an "offline prediction device" to a "controllable assisting system". Combined with Figure 8, this paper regards the 63% coverage neighborhood as a safe recommended working point: if we continue to raise coverage, it will significantly raise the serious error rate, hence if we continue to lower coverage, it will rapidly weaken the practical effect of the system in the real world..

For the further explanation of the origin of these remaining high-risk mistakes, this paper hence attributes coding to the failure positions of every single path. Regarding VC+UG, the maximum proportion of remaining high-risk mistakes is omission of sentencing situations (24.6%), absence of interpretation consistency (21.3%), and anchoring deviation of similar cases (19.1%), therefore followed by confusion of legal elements (17.4%), cross-region deviation (9.2%), and screening error causing unjust judgment (8.4%). This structure has very big contrast with Static-SL: in the latter one, the main problems are confusion of legal elements and omission of plot, which together add up to more than 60%. This gives indication that the optimization route has truly already changed the properties of the mistakes.

These requirements and the risk boundaries in Figure 8 are not two separated groups of statements, but on the contrary, they reflect each other: if a sentencing help system cannot give confidence bounds, explain reverse operations, or take the initiative to shift work in high-risk situations, then it does not satisfy the lowest conditions for entering judgment work, hence even if it gets a higher mark on the ordinary standard. Relevant researches on fairness have likewise indicated that merely comparing average performance is not enough, and that differences of reliability between groups and uncertainty control can likewise influence the acceptability of high-risk systems [23, 24].

Based on these results, this paper suggests limiting the actual deployment interface of a sentencing assistance system to four types of outputs. First, candidate offenses and candidate

statutes are given, but are required to be aligned with the factual elements of the case on an item-by-item basis. Second, it gives the interval-projected sentencing recommendation, rather than a single "point projection." Third, it presents the class of cases and the chain of reasons that support the recommendation and allows the judge to quickly check them. Fourth, it outputs risk alerts, including statute update alerts, not guilty verdict alerts, evidence conflict alerts and confidence levels. In all cases where the statute has just been revised, where there are clear clues of innocence in the case, where there is a complex structure of multiple offenses and penalties, and where there are significant conflicts in the core evidence, the system should default to manual review and no longer pursue automatic coverage. This deployment mode is not so much a restriction on AI, but rather a prevention of "high-scoring models" from overstepping the boundaries of adjudication that it should not have transgressed.

Thus, the value of sentencing assistance comes not from getting judges out, but from allowing them to see the boundaries of the statute, the coordinates of the case, the interpretive basis, and the signals of risk more quickly. As long as the ultimate discretion remains in human hands, the best role for AI is to move high-frequency, repetitive, and structurable judgments forward, and to make high-risk, uncertain, and value-measuring aspects visible, rather than completing irreplaceable discretionary actions for the judge.

4 Conclusion

In this article, we re-give the definition of the "accuracy" meaning of AI systems surrounding the high-risk situation of criminal sentence aid, and carry out a deployment-focused and integrated evaluation for open technical paths. When compared with the traditional method that compresses accuracy into one single classification or regression measuring index, this article decomposes it into five dimensions: legal element matching, sentencing range correctness, dynamic stability, reason traceability, and operation safety, and therefore builds a path-level evidence corpus that contains publicly published benchmark data, representative methods, and management documents.

(1) This paper completes the unification of the evidence corpus with difficult scenarios at the object organization level by putting CAIL2018, LawBench, LexEval, LJPCheck, LawShift, LJPIV and other public resources and judicial governance documents into the same evaluation panel, avoiding the problem of "high model score" and "deployment security". This avoids the disconnection between "high model score" and "deployment security".

(2) This present paper puts forward a closed-loop optimization route which is made up of legal knowledge base placement, case-similar searching, section restrictions, step-by-step verification, and uncertainty controlling on method layer and result layer. The outcome of comparison shows that the path of VC+UG obtains a comprehensive accuracy score of 79.3, which is 27.1 and 28.6 points higher than the direct prompting large model in the law change and innocence determination scenes, respectively, thus the serious error rate can be controlled at 8.9% with 63% automatic coverage.

(3) This paper also retains clear boundaries for its conclusions: the results here are based on standardized scores from open literature and open benchmarks, not prospective A/B validation of a real court online system; local sentencing practices, differences in rules of evidence, and ongoing revisions of substantive law may still change the optimal weights and thresholds. Follow-up work should incorporate more detailed plot annotations, longer-term tracking of statutory updates, and prospective human-computer collaborative validation in conjunction with real trial business processes in order to determine the boundaries of stable application of the Sentencing Assistance System in different jurisdictions.

About the Author

Jun Hu was born in Hezhou, Guangxi, China, in 1982. She obtained her doctoral degree from Dong-A University in South Korea. Currently, she is employed at Hezhou University. Her research focuses on Civil and Commercial Law, Food Safety Law, Consumer Law, and Intellectual Property Law.

References

[1] Supreme People's Court. (2022). Opinion on regulating and strengthening artificial intelligence judicial application.

[2] CEPEJ. (2018). European ethical charter on the use of artificial intelligence in judicial systems and their environment. Council of Europe.

[3] UNESCO. (2025). Guidelines for the use of AI systems in courts and tribunals. Paris: UNESCO.

[4] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).

[5] Cui, J., Shen, X., & Wen, S. (2023). A survey on legal judgment prediction: Datasets, metrics, models and challenges. IEEE Access, 11, 102050-102071.

[6] Feng, Y., Li, C., & Ng, V. (2022). Legal judgment prediction: A survey of the state of the art. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (pp. 5461-5469).

[7] Xiao, C., Zhong, H., Guo, Z., et al. (2018). CAIL2018: A large-scale legal dataset for judgment prediction. arXiv, arXiv:1807.02478.

[8] Liu, Y., Wu, Y., Zhang, Y., et al. (2023). ML-LJP: Multi-law aware legal judgment prediction. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1023-1034).

[9] Wu, Y., Zhou, S., Liu, Y., et al. (2023). Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 12060-12075).

[10] Fei, Z., Shen, X., Zhu, D., et al. (2024). LawBench: Benchmarking legal knowledge of large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 7933-7962).

[11] Li, H., Chen, Y., Ai, Q., et al. (2024). LexEval: A comprehensive Chinese legal benchmark for evaluating large language models. Advances in Neural Information Processing Systems, 37, 25061-25094.

[12] Zhang, Y., Huang, W., Feng, Y., et al. (2024). LJPCheck: Functional tests for legal judgment prediction. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5878-5894).

[13] Han, Z., Yang, Y., Feng, Y., et al. (2025). LawShift: Benchmarking legal judgment prediction under statute shifts. In NeurIPS 2025 Datasets and Benchmarks Track.

[14] Zhang, K., Yang, H., Tang, X., et al. (2025). Beyond guilt: Legal judgment prediction with trichotomous reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 1815-1826).

[15] Ryberg, J. (2025). Criminal sentencing and artificial intelligence: What is the input problem? Criminal Law and Philosophy, 19(2), 203-220.

[16] Ryberg, J. (2025). Artificial intelligence and criminal justice: How to use algorithmic sentencing support in real life (and ethically non-ideal) penal systems? AI and Ethics, 5, 3255-3263.

[17] Kiejnich-Kruk, K., Twardawa, M., & Formanowicz, P. (2025). Overcoming sentencing inconsistency: A proposal for algorithmic guidelines and juridical misalignment index. Artificial Intelligence and Law.

[18] Rodger, H., Lensen, A., & Betkier, M. (2023). Explainable artificial intelligence for assault sentence prediction in New Zealand. Journal of the Royal Society of New Zealand, 53(1), 133-147.

[19] Zhao, Q. (2025). Legal judgment prediction via legal knowledge extraction and fusion. Journal of King Saud University - Computer and Information Sciences, 37, 31.

[20] Shen, Y., Wei, H., & Tian, X. (2025). TA-LJP: Term-aware legal judgment prediction. Information, 17(1), 17.

[21] Shi, W., Zhu, H., Ji, J., et al. (2025). LegalReasoner: Step-wised verification-correction for legal judgment reasoning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (pp. 7297-7313).

[22] Mei, Z., Zhan, C., Ye, W., et al. (2026). Legal judgment prediction based on charge-anchored graph constraints. Expert Systems with Applications, 302, 130480.

[23] Berk, R. A., Kuchibhotla, A. K., & Tchetgen Tchetgen, E. (2024). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. Sociological Methods & Research, 53(4), 1629-1675.

[24] Santosh, T. Y. S. S., & Chowdhury, I. (2025). Fairness beyond performance: Revealing reliability disparities across groups in legal NLP. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (pp. 24376-24390).