LAWRE: Evaluating Legal Case Retrieval for Practical Legal Reasoning

Xichen Lin, Yi Yang, Shuxian Xie, Yi Feng*, Chuanyi Li
State Key Laboratory for Novel Software Technology, Nanjing University, China
fy@nju.edu.cn

***Abstract*—Legal Case Retrieval (LCR) involves retrieving legally relevant cases from a candidate corpus given a query case. While state-of-the-art (SOTA) LCR models have achieved high recall scores on public datasets (e.g., LeCaRD), such datasets fail to assess key legal reasoning skills needed in practice (e.g., distinguishing confused cases, identifying legal elements). To bridge this gap, we introduce LAWRE, a comprehensive benchmark for evaluating LCR systems, facilitating an understanding of their reasoning processes and offering detailed diagnostic feedback. Testing eight SOTA models on LAWRE exposes notable weaknesses, which in turn motivate analyses of these shortcomings and guide the development of more practically effective LCR systems.**

***Index Terms*—Legal Case Retrieval, AI for Law, Legal Reasoning Benchmark**

I. Introduction

Legal Case Retrieval (LCR) involves retrieving legally relevant cases from a corpus of historical cases given a query case [1]–[6]. A query case involves the factual description of the event, while in the legal database, each historical case includes not only the case facts but also the court ruling and the reasoning applied during the decision process. Figure 1 shows an overview of LCR, illustrating the inherent complexity of retrieving cases that are not only textually similar but also legally comparable. LCR is vital for legal reasoning and decision-making. In common law systems (e.g., the US, UK), it supports arguments through relevant precedents [7], while in civil law systems (e.g., China, Germany), it aids by providing applicable statutes and related past decisions, benefiting both legal professionals and the public.

Despite substantial advances in LCR methods [8]–[10], a key challenge remains insufficiently addressed: how to comprehensively and reliably evaluate these models in real-world judicial practice, where legal correctness and reasoning fidelity are critical? Existing evaluations largely rely on generic metrics such as recall and F1 score [11]–[13]. Yet such single-metric approaches are inadequate, as they fail to uncover a model's underlying reasoning behavior or diagnose its weaknesses. Given that LCR is a high-stakes judicial task, where retrieval errors, such as suggesting irrelevant or incorrect precedents, can undermine legal fairness, evaluation must go beyond surface-level accuracy to assess whether models exhibit the legal reasoning capabilities essential in legal contexts. For instance, identifying legal elements is central to legal reasoning in LCR because it assesses the grounds for similarity between two cases. Ideally, a model should detect key legal elements, such as the intent to *permanently deprive* (Theft) versus *temporarily use* (Embezzlement) property to distinguish theft and embezzlement cases, and retrieve precedents aligned not only in factual content but also in these legal elements. Consider a case where the defendant temporarily used a company laptop without authorization but later returned it. Legally, this constitutes *embezzlement*, characterized by temporary possession, rather than *theft*, which requires the permanent deprivation of property. However, under current metrics, a model that only relies on overall factual similarity could retrieve both embezzlement and theft cases while still achieving a perfect recall [14], [15]. Such mistakes can lead to retrieving incorrect precedents, potentially misguiding judgments in high-stakes judicial contexts. This failure exemplifies a deeper and more systemic flaw: most existing datasets and evaluation protocols prioritize overall performance metrics, while providing little insight into whether a model's retrieval behavior is grounded in sound legal reasoning.

In this paper, we propose a shift in LCR evaluation from focusing solely on overall retrieval performance to a multidimensional paradigm that explicitly assesses the legal reasoning skills essential to the task. Our framework goes beyond generic metrics to assess dimensions like legal element-based retrieval and fairness, bridging the gap between quantitative scores and the qualitative demands of real-world legal decision-making in LCR (Figure 1). Specifically, we present LAWRE[1], a new benchmark for LCR models that enables deeper understanding of model behavior and provides diagnostic insights. Our contributions are three-fold:

- We propose a multidimensional evaluation framework for LCR that goes beyond retrieval performance and assesses legally grounded reasoning skills, including robustness to incomplete information, sensitivity to key elements, verdict consistency, and fairness.
- We develop LAWRE, a large-scale benchmark to operationalize this framework, comprising 51,513 tests covering diverse and relevant legal scenarios.
- We conduct experiments on LAWRE using eight state-of-the-art LCR models, providing analyses that reveal their

* Corresponding author.

[1]LAWRE is available at https://github.com/crazyTomlin/LAWRE.

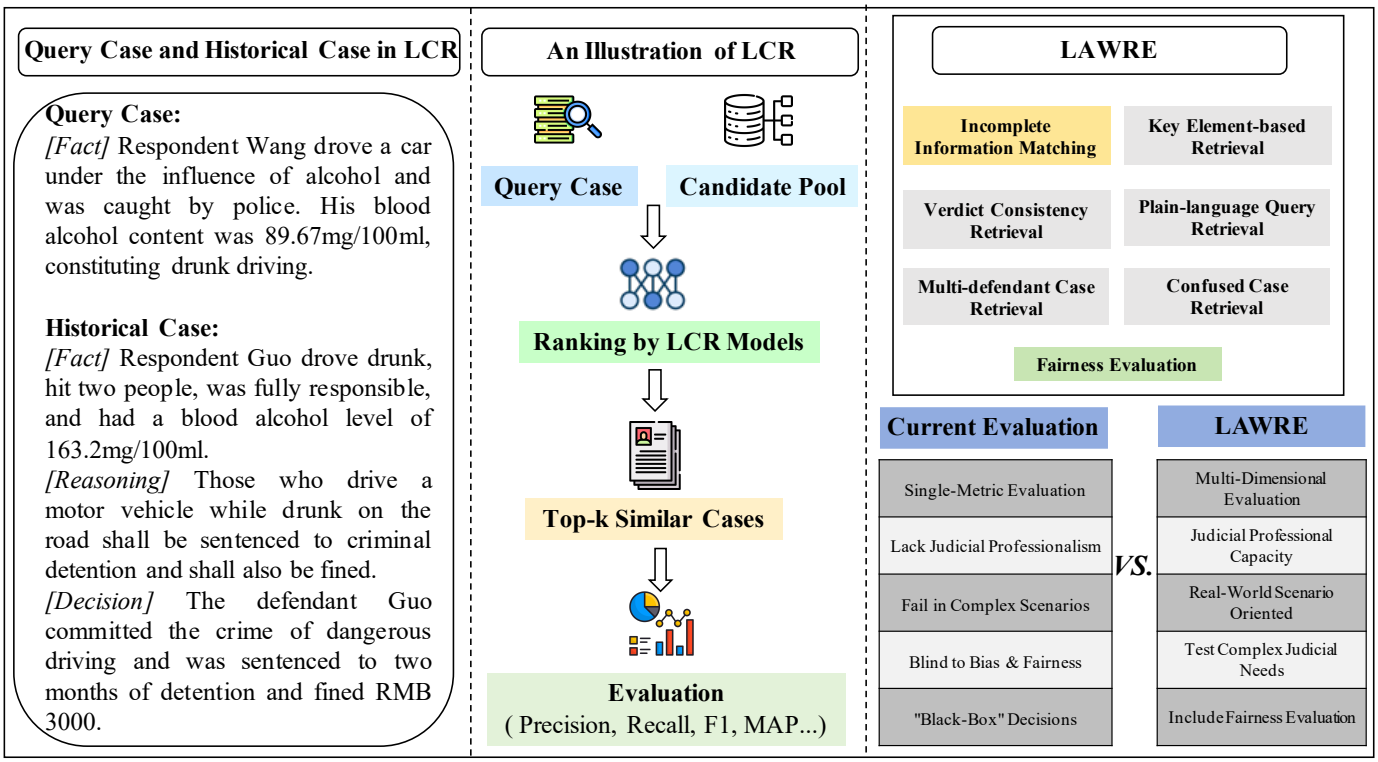

Fig. 1. An illustration of LCR and how LAWRE differs from existing datasets.

strengths, limitations, and failure modes across different reasoning dimensions.

II. Related Work

Legal Case Retrieval has been an active research topic for several decades and is commonly framed as a specialized instance of information retrieval, where the objective is to rank historical cases according to their legal relevance to a given query case. Early LCR methods [16]–[18] relied on statistical approaches such as TF-IDF [19] and BM25 [20], which estimate relevance based on surface-level textual similarity. However, such approaches are insufficient in legal settings where relevance depends on legally decisive factors.

The limitations of traditional IR-based approaches stem from a fundamental mismatch between generic text relevance and legal relevance. In legal reasoning, two cases may appear highly similar at the factual or semantic level while being legally distinct due to differences in statutory thresholds, subjective intent, or legally defined elements. As a result, retrieval models that focus solely on textual similarity may rank legally inappropriate precedents highly, undermining core judicial principles such as consistency and sufficiency of reasoning. Prior work has highlighted that relevance judgments in LCR require external legal knowledge and structured reasoning beyond what can be inferred from raw text alone. This observation motivates the development of retrieval models that move beyond bag-of-words matching and attempt to capture deeper semantic and legal structures.

Recent advances in neural information retrieval have significantly influenced LCR research. Transformer-based architectures [21] enable models to encode long legal documents and capture contextual semantics more effectively than earlier methods. Building on these advances, several studies propose structure-aware or task-specific pretraining strategies tailored to legal documents. For example, models such as SAILER [22] and DELTA [23] explicitly incorporate legal document structure or alignment signals during pretraining, leading to substantial gains over generic language models. In parallel, large language models (LLMs), including both general-purpose and legal-domain-specific variants (e.g., Llama3-8B-Chinese-Law [24] and DISC-LawLLM [25]), have been applied to LCR tasks and shown competitive retrieval performance due to their broad pretraining and emergent reasoning capabilities.

To support model development and evaluation, multiple LCR datasets have been released across jurisdictions, including COLIEE (Canada) [26], ILDC (India) [27], LeCaRD/LeCaRDv2 (China) [28], [29], and CLERC (US) [30]. Given the significant impact of legal decisions on individuals, there are also works focusing on the evaluation of AI systems in the legal domain. A recent survey on Legal Case Retrieval [4] summarizes evaluation practice, reporting commonly used ranking metrics such as Accuracy@K, Precision@K, Recall@K, F1@K, NDCG@K, and MAP [31]–[36], and discussing evaluation challenges specific to legal settings. Protocol-aware evaluation choices [37] make evaluation choices explicit, for example the construction of candidate

TABLE I
EVALUATION DIMENSIONS AND # OF EVALUATION TESTS.

ID	Dimension	# of Test Cases
D1	Incomplete Information Matching	15,414
D2	Key Element-based Retrieval	9,970
D3	Verdict Consistency Retrieval	2,200
D4	Plain-language Query Retrieval	7,707
D5	Multi-defendant Case Retrieval	1,150
D6	Confused Case Retrieval	12,000
D7	Fairness Evaluation	3,072

pools and the thresholds for positive labels, while retaining MAP, P@k, and nDCG@k as primary scores. Controlled user studies [38]–[40] show that the distribution of retrieved precedents can shift participants' sentencing decisions, which underscores fairness and bias considerations. In this paper, we provide an evaluation framework tailored to LCR systems.

Unlike the above works, which only focus on one or a few aspects, we give broader insights with respect to seven different dimensions, including reasoning aspects, ethical dimensions, etc. While our study centers on Chinese LCR, the evaluation methodology is adaptable to other jurisdictions.

III. LAWRE

A. Selecting Dimensions

To define evaluation dimensions for legal reasoning in LCR, we combine (1) a systematic review of existing LCR limitations and (2) structured interviews with legal practitioners. Analysis of 67 studies showed reliance on generic metrics (e.g., recall, F1) that ignore reasoning skills like identifying legal key elements. Interviews with seven legal practitioners (three lawyers and four judges) were conducted to identify key challenges and desired properties of LCR systems. Key needs included robustness to partial/informal queries, consistency with judicial principles (e.g., "similar verdicts for similar facts"), etc. These findings shaped seven evaluation dimensions spanning three aspects as shown in Table I.

B. Constructing LAWRE

During benchmark construction, particular care was taken to avoid trivial retrieval shortcuts based on near-duplicate texts. Candidate pools were inspected where additional annotation was required, ensuring that high performance reflects legal reasoning rather than simple textual overlap. We detail the seven evaluation dimensions of LAWRE and explain how LCR systems are tested against them below.

The first aspect, general retrieval capacity, measures fundamental robustness in accessing relevant information. **Incomplete Information Matching** tests performance on truncated or partial queries, which is common when lawyers face time limits or self-represented litigants recall cases imperfectly. Based on LeCaRD, we select a seed set and create two variants for each query case: (1) truncating case text to the first two-thirds, and (2) generating concise factual summaries via GPT-4o. Robustness is indicated when performance on modified queries remains close to that on the originals.

The second aspect, judicial professional capacity, assesses alignment with legal reasoning across five sub-dimensions. **Key Element-based Retrieval** evaluates whether models can detect key elements that determine verdicts. In drink-driving cases, for instance, an LCR model should accurately identify the defendant's blood alcohol concentration; otherwise, it may retrieve drunk driving precedents for a drink driving case, as both share similar "drinking and driving" semantics. However, drunk driving carries significantly harsher penalties and cannot serve as a valid precedent for drink driving, making precise element recognition essential to uphold the judicial principle of relevance and sufficiency. We construct evaluation tests from ELAM [41], whose cases are annotated with two types of key elements (key facts and key factors). We remove these elements from the facts and perform retrieval. The pass rate (the ratio of correctly matched cases after element removal to those correctly matched before removal) serves as the evaluation metric. It is important to note that a high pass rate in this setting does not indicate desirable robustness; instead, it reveals that a model's similarity judgments are largely insensitive to the presence or absence of legally decisive elements, suggesting reliance on superficial or contextual cues.

Verdict Consistency Retrieval assesses the ability to retrieve cases with similar judgment outcomes (including charges, statutes, and penalty term) to the query case. A LCR model should prioritize precedents where the legal reasoning and judgment outcome both align with the query, thereby reflecting the "similar cases–similar judgments" principle. Note key element-based retrieval is matching the elements for legal qualification, while verdict consistency retrieval goes further by aligning the judicial conclusions. A model may correctly identify key elements yet still fail in verdict consistency retrieval if, for example, the matched legally similar cases have different charges or sentencing. Using LeCaRD cases, we obtain their final verdicts from China Judgments Online and measure query–candidate similarity across three dimensions: charges, statutes, and penalty term. Charge and statute similarities use the Jaccard coefficient. Penalty term similarity is computed by dividing the interval difference between two sentences by the total sentencing range. The final score is the average of these metrics. Since LeCaRD's ground-truth candidates are unordered, we rerank them by judgment similarity. Models that retrieve these reranked cases accurately are considered capable of both legal similarity retrieval and verdict consistency retrieval.

Plain-language Query Retrieval assesses adaptability to colloquial or layperson inputs. This reflects the judicial value of *inclusive access*: LCR systems must serve not only professionals but also laymen. Based on LeCaRD, colloquial versions of facts are generated using GPT-4o, while the candidate cases remained unchanged. The minimal performance gap between colloquial and original versions suggests strong capability in processing plain-language queries.

Multi-defendant Case Retrieval evaluates a model's abil-

ity to handle cases involving multiple parties and complex role relationships (e.g., principal versus accomplice liability). This dimension reflects the doctrine of joint criminality and differentiated culpability, which demands precise reasoning about the roles and responsibilities of different defendants. Failure in this aspect may result in misaligned precedent recommendations in multi-party contexts. Based on ELAM, we select 23 multi-defendant query cases and, for each, construct a pool of 50 candidate cases using the BM25 algorithm. Legal professionals (three lawyers) then annotated the ground-truth matches according to the number of defendants, principal/accessory offender relationships, and verdict similarity. Disagreements were resolved through discussion until consensus was reached, ensuring consistency of the ground-truth labels. Higher retrieval scores in this evaluation indicate stronger capability in multi-defendant case retrieval.

Confused Case Retrieval evaluates precision in distinguishing confused cases, whose facts, charges, or outcomes are so similar to others that even humans might find them hard to distinguish (e.g., Intentional Injury vs. Negligent Injury). Subtle factual distinctions (e.g., subjective intent) can decisively alter charges and sentencing. Failure to capture these nuances lead to wrong precedents. Five groups of commonly confused charges are selected. Based on LeCaRD, each query retains only candidate cases matching its type and its confused counterpart. High retrieval scores reflect strong capability in distinguishing confused cases.

Finally, we include **Fairness Evaluation** to ensure ethical considerations. Specifically, *Fairness Evaluation* assesses robustness to demographic perturbations (names, gender, ethnicity). Retrieval outcomes should be invariant to attributes irrelevant to legal liability. If system performance varies with demographic substitutions, it reveals embedded biases that threaten the legitimacy of AI-assisted judicial systems. Based on CAIL2019-SCM [42], we perturb demographic attributes (names, genders, ethnicities and their combinations) in query cases and evaluated model robustness by comparing retrieval accuracies between the original and perturbed versions.

IV. Experiment

A. Baselines

We evaluate eight SOTA approaches in two categories. LCR-specialized models include Lawformer [43], SAILER [22], LEAD-trained DPR [10], and DELTA [23]. Large language models (LLMs) include two general-purpose LLMs (Qwen2.5-14B [44], DeepSeek-V3 [45]) and two legal-domain LLMs (Llama3-8B-Chinese-Law [24], DISC-LAW-7B [25]). For LCR-specialized models, we employ publicly available checkpoints for testing, while for general LLMs, we use official APIs with prompts to generate similarity scores.

B. Experimental Results

Figure 2 presents a summary of the results. Overall, all models perform poorly on LAWRE, highlighting the gap between generic benchmark performance and the detailed reasoning abilities required in this task. This underscores the limitations of standard benchmarks, where high F1 or recall scores can mask deeper weaknesses. For instance, despite achieving strong scores on official benchmarks, DeepSeek-V3 attains only 31.55% MAP on D3. Generally, LCR-specialized models outperform LLMs, including domain-specific ones, indicating that LCR-oriented model design remains essential in specialized domains. While LLMs excel in general-purpose tasks, incorporating LCR-oriented architectures is still necessary to address domain-specific challenges effectively.

For D1, all models show higher performance on summarized queries than on truncated queries. This finding suggests that, in model design, incorporating content-aware preprocessing, such as automatic legal fact summarization, prior to encoding could be advantageous. From a judicial perspective, deploying retrieval systems that operate on concise yet fact-rich case summaries may better align with how legal professionals review case materials, enhancing both efficiency and precision in legal information systems.

All models exhibit performance degradation on D2, but general LLMs are better than LCR-specialized models, likely due to stronger reasoning and broader knowledge that allow partial inference of missing elements. These results support a hybrid pipeline: use an LLM for broad recall under incomplete information, then apply an LCR-specific re-ranking for precision. Joint training of retrieval and legal element tagging could further emphasize key element, mirroring judicial practice where relevance hinges on the presence or absence of key elements.

Results on D3 expose a gap between current text-similarity objectives and the judicial principle of "similar cases, similar judgments". Such errors undermine legal predictability, suggesting the need for outcome-aware retrieval. Promising directions include multi-task learning that combines retrieval with judgment prediction, or neuro-symbolic architectures that pair neural fact encoding with symbolic reasoning to better align outcomes with doctrinal rules.

LLMs generally outperform LCR-specialized models on D4, likely due to their diverse pre-training data, which enables them to normalize colloquial queries (e.g., "driver drunk hit pedestrian") into formal legal descriptions. Given that inclusive access calls for systems capable of understanding layperson queries, an effective strategy is data augmentation that pairs colloquial expressions with their formal legal equivalents for training.

On D5, DISC-LAW-7B outperforms other SOTA models, likely due to superior relational reasoning and entity tracking from pre-training on diverse, complex narratives. This allows it to capture multi-defendant interactions, role distinctions, and entity-specific actions—crucial for LCR tasks involving joint criminality and differentiated culpability. Future gains may come from explicitly modeling defendant relationships via graph-based encodings or entity-aware transformers.

On D6, models achieved satisfactory performance. Many candidates contain distinctive keywords, allowing superficial matching. For example, distinguishing "robbery" from "theft" was often solved by detecting the token "violence". While

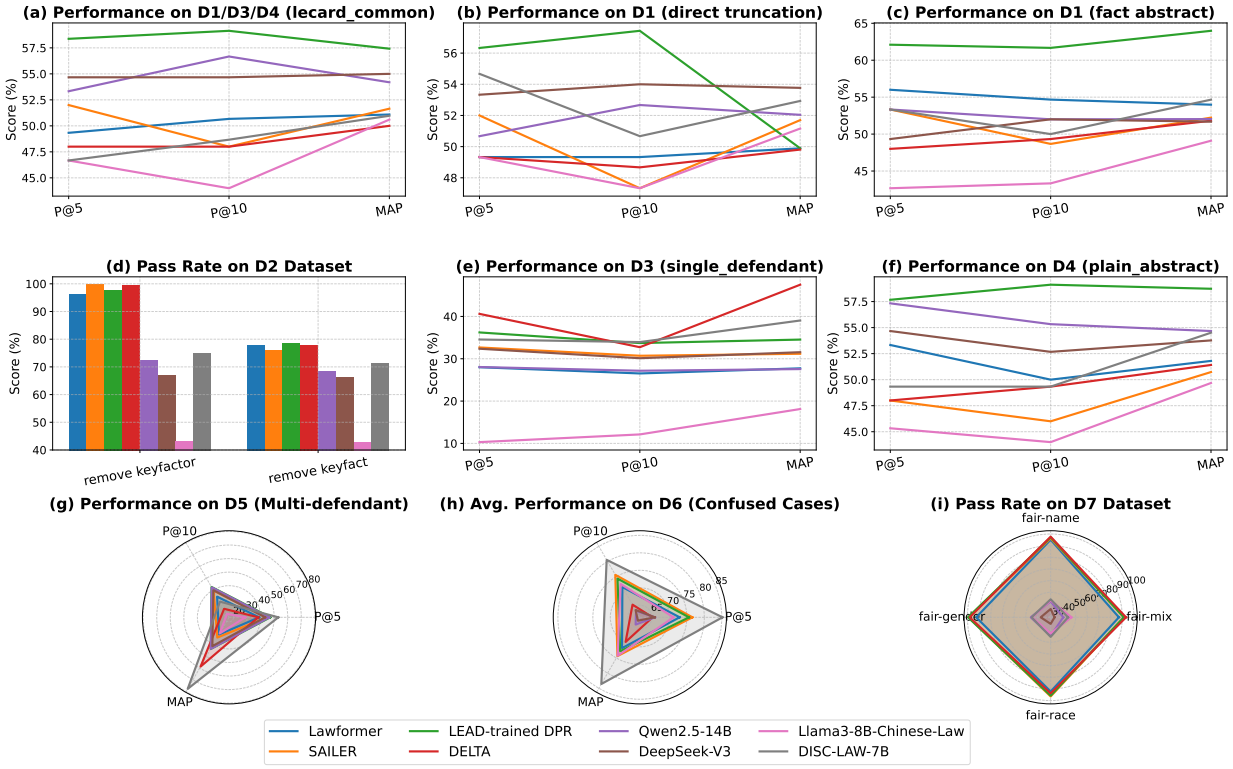

Fig. 2. The results of the experiments.

useful, this underestimates the real difficulty of legal qualification. Future work should develop adversarial counterfactual datasets—cases differing only in one decisive fact—to force models beyond keyword reliance.

A clear ethical gap emerged on D7: LCR-specialized models remain invariant to demographic changes, whereas LLMs are highly sensitive due to training on web text containing social biases. Mitigation calls for fairness-aware learning, e.g., adversarial debiasing or fairness constraints, to ensure LCR relies solely on legally relevant factors.

V. Conclusion

We propose LAWRE, a novel framework for assessing LCR systems' legal reasoning capabilities beyond overall retrieval performance. Through systematic evaluation of eight state-of-the-art models, our results expose notable gaps in robustness, reasoning fidelity, and fairness that are largely overlooked by existing benchmarks. These findings suggest that achieving high retrieval scores does not necessarily imply legally sound behavior, underscoring the need for evaluation frameworks that align more closely with real-world judicial reasoning requirements. By providing fine-grained diagnostic signals, LAWRE offers a practical tool for guiding both model development and evaluation toward legally reliable LCR systems.

Ethical Statements The legal cases used in this study originate from publicly accessible judicial databases with all sensitive personal identifiers (including names, identification numbers, and contact details) rigorously anonymized. Our research strictly adheres to the data protection regulations and ethical guidelines for legal applications of AI. The proposed evaluation framework is designed as an auxiliary tool to enhance the judicial decision support systems rather than replace human legal professionals. We explicitly affirm that no experimental results in this paper should be interpreted as supporting fully automated adjudication systems. The technical objectives focus exclusively on improving the comprehensiveness of legal case retrieval evaluation metrics, not on establishing autonomous judgment capabilities. Even models achieving optimal performance metrics require mandatory human oversight in practical legal workflows according to judicial ethics principles.

Limitations For several dimensions, our evaluation is confined to a restricted set of charges, necessitating broader empirical scrutiny. Additionally, several dimensions are calibrated specifically to the Chinese legal system and are not immediately transferable to alternative jurisdictions. Nevertheless, this limitation does not imply a lack of generalizability; rather, it highlights the capacity for stakeholders to tailor specialized dimensions in accordance with contextual requirements.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62406139), the State Key Laboratory for Novel Software Technology at Nanjing University (KFKT2025A15, ZZKT2025B14, KFKT2024A07, ZZKT2024B02), and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118).

REFERENCES

[1] David C Blair and Melvin E Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM*, vol. 28, no. 3, pp. 289–299, 1985.

[2] Marie-Francine Moens, "Innovative techniques for legal text retrieval," *Artificial Intelligence and Law*, vol. 9, no. 1, pp. 29–57, 2001.

[3] Jieke Li, Min Yang, and Chengming Li, "Clc-rs: a chinese legal case retrieval system with masked language ranking," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 4734–4738.

[4] Yi Feng, Chuanyi Li, and Vincent Ng, "Legal case retrieval: A survey of the state of the art," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 6472–6485.

[5] Vu Tran, Minh Le Nguyen, Satoshi Tojo, and Ken Satoh, "Encoded summarization: Summarizing documents into continuous vector space for legal case retrieval," *Artificial Intelligence and Law*, vol. 28, no. 4, pp. 441–467, 2020.

[6] Soufiane El Jelali, Elisabetta Fersini, and Enza Messina, "Legal retrieval as support to emediation: Matching disputant's case and court decisions," *Artificial Intelligence and Law*, vol. 23, pp. 1–22, 2015.

[7] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner, "Recognizing cited facts and principles in legal judgements," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 107–126, 2017.

[8] Althammer et al., "Dossier@ coliee 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval," in *COLIEE*, 2021, pp. 8–14.

[9] Shao et al., "Bert-pli: modeling paragraph-level interactions for legal case retrieval," in *IJCAI*, 2021, pp. 3501–3507.

[10] Gao et al., "Enhancing legal case retrieval via scaling high-quality synthetic query-candidate pairs," in *EMNLP*, 2024, pp. 7086–7100.

[11] Marc Van Opijnen and Cristiana Santos, "On the concept of relevance in legal information retrieval," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 65–87, 2017.

[12] Carlo Sansone and Giancarlo Sperlí, "Legal information retrieval systems: State-of-the-art and open issues," *Information Systems*, vol. 106, pp. 101967, 2022.

[13] Daniel Locke and Guido Zuccon, "Case law retrieval: Problems, methods, challenges and evaluations in the last 20 years," *arXiv preprint arXiv:2202.07209*, 2022.

[14] Zhou et al., "Boosting legal case retrieval by query content selection with large language models," in *SIGIR*, 2023, pp. 176–184.

[15] Deng et al., "An element is worth a thousand words: Enhancing legal case retrieval by incorporating legal elements," in *ACL*, 2024, pp. 2354–2365.

[16] Ha-Thanh Nguyen, Manh-Kien Phi, Xuan-Bach Ngo, Vu Tran, Le-Minh Nguyen, and Minh-Phuong Tu, "Attentive deep neural networks for legal document retrieval," *Artificial Intelligence and Law*, vol. 32, no. 1, pp. 57–86, 2024.

[17] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera, "Legal document retrieval using document vector embeddings and deep learning," in *Science and information conference*. Springer, 2018, pp. 160–175.

[18] Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith, "Exploring the use of text classification in the legal domain," *arXiv preprint arXiv:1710.09306*, 2017.

[19] Karen Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[20] Robertson et al., "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian, "Sailer: structure-aware pre-trained language model for legal case retrieval," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 1035–1044.

[23] Li et al., "DELTA: pre-train a discriminative encoder for legal case retrieval via structural word alignment," in *AAAI*, 2025, pp. 27072–27080.

[24] mrhua, "llama3-8b-chinese-lora-law_f16_q4_0," https://ollama.com/mrhua/llama3-8b-chinese-lora-law_f16_q4_0.

[25] Yue et al., "Disc-lawllm: Fine-tuning large language models for intelligent legal services," *arXiv preprint arXiv:2309.11325*, 2023.

[26] Baban Gain, Dibyanayan Bandyopadhyay, Tanik Saikh, and Asif Ekbal, "Iitp@ coliee 2019: legal information retrieval using bm25 and bert," *arXiv preprint arXiv:2104.08653*, 2021.

[27] Malik et al., "Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation," *arXiv preprint arXiv:2105.13562*, 2021.

[28] Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma, "Lecard: a legal case retrieval dataset for chinese law system," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 2342–2348.

[29] Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu, "Lecardv2: A large-scale chinese legal case retrieval dataset," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2251–2260.

[30] Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme, "Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation," *arXiv preprint arXiv:2406.17186*, 2024.

[31] Junlin Zhu, Xudong Luo, and Jiaye Wu, "A BERT-based two-stage ranking method for legal case retrieval," in *Proceedings of the 15th KSEM*, 2022, pp. 534–546.

[32] Vu Tran, Minh L. Nguyen, and Ken Satoh, "Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model," in *Proceedings of the 17th ICAIL*, 2019, pp. 275–282.

[33] Yunqiu Shao, Bulou Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma, "THUIR@COLIEE-2020: Leveraging semantic understanding and exact matching for legal case retrieval and entailment," *arXiv preprint arXiv:2012.13102*, 2020.

[34] Yen Thi-Hai Vuong, Quan Minh Bui, Ha-Thanh Nguyen, Thi-Thu-Trang Nguyen, Vu Tran, Xuan-Hieu Phan, Ken Satoh, and Le-Minh Nguyen, "SM-BERT-CR: A deep learning approach for case law retrieval with supporting model," *Artificial Intelligence and Law*, pp. 1–28, 2022.

[35] Dunlu Peng, Jiyin Yang, and Jing Lu, "Similar case matching with explicit knowledge-enhanced text representation," *Applied Soft Computing*, vol. 95, pp. 106514, 2020.

[36] Hui Li, Jin Lu, Yuquan Le, and Jiawei He, "IACN: Interactive attention capsule network for similar case matching," *Intelligent Data Analysis*, vol. 26, no. 2, pp. 525–541, 2022.

[37] Deng et al., "Learning interpretable legal case retrieval via knowledge-guided case reformulation," in *EMNLP*, 2024, pp. 1253–1265.

[38] Wang et al., "Investigating the influence of legal case retrieval systems on users' decision process," in *SIGIR*, 2023, pp. 169–175.

[39] Shao et al., "Understanding relevance judgments in legal case retrieval," *TOIS*, vol. 41, no. 3, pp. 1–32, 2023.

[40] Sebastian Lewis, "Precedent and the rule of law," *Oxford Journal of Legal Studies*, vol. 41, no. 4, pp. 873–898, 2021.

[41] Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen, "Explainable legal case matching via inverse optimal transport-based rationale extraction," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 657–668.

[42] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al., "Cail2019-scm: a dataset of similar case matching in legal domain," *arXiv preprint arXiv:1911.08962*, 2019.

[43] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun, "Lawformer: A pre-trained language model for chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021.

[44] Hui et al., "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186*, 2024.

[45] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al., "Deepseek llm: Scaling open-source language models with longtermism," *arXiv preprint arXiv:2401.02954*, 2024.