Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.35434

Research on Dynamic Judicial Hallucination Risk Identification Mechanism for Legal Domain Large Language Models

Jingyu Yang

National Security and Legal Education Research Centre (NSLERC), The Education University of Hong Kong, Hong Kong, China
rara481846778@gmail.com

Abstract. Legal domain large language models are increasingly used in legal consultation, precedent retrieval, document generation, and AI-assisted adjudication. While these models improve the efficiency of legal information processing, legal hallucinations such as fabricated cases, misquoted statutes, distorted holdings, and broken reasoning chains may undermine the reliability of judicial grounds, the clarity of liability allocation, and procedural legitimacy. To identify hallucination risks in AI-generated legal opinions, this study constructs a dynamic risk identification mechanism based on verifiable legal corpora collected from public statutes, judicial interpretations, judicial documents, and typical cases. The mechanism integrates retrieval-augmented generation, semantic consistency detection, citation validity assessment, and reasoning-chain completeness scoring to classify model outputs into different risk levels. The experimental results show that the proposed model outperforms ordinary prompting, RAG prompting, and self-consistency detection in accuracy, recall, F1-score, and AUC, achieving an overall accuracy of 0.86 and a high-risk recall of 0.89. The mechanism transforms legal hallucination from an opaque generation error into a computable, explainable, and reviewable judicial risk, providing technical support for algorithmic transparency, human oversight, and responsibility allocation in AI-assisted justice.

Keywords: Legal Domain Large Language Models, Judicial Hallucination, Risk Identification, Explainable AI, Human Oversight

1. Introduction

Legal domain large language models are moving from legal consultation, contract review, and case retrieval to AI-assisted adjudication. Judicial activities require accuracy, traceability, and clear responsibility boundaries, which gives model-generated risks stronger normative consequences. Legal hallucination mainly appears as fabricated cases, misquoted statutes, distorted holdings, and broken reasoning chains. Existing research has shown that legal large language models still produce a considerable proportion of erroneous outputs in verifiable legal questions [1]. Current studies mainly focus on model performance, task adaptation, or ethical principles, while limited attention

DOI: 10.54254/2755-2721/2026.35434

has been paid to how hallucination risks can be dynamically identified, graded, and incorporated into human oversight within judicial procedures [2]. Focusing on risk identification in AI-generated legal content, this study combines retrieval-augmented generation, legal knowledge graphs, semantic consistency detection, and explainable scoring to construct a dynamic judicial hallucination risk identification mechanism for AI-assisted adjudication.

2. Literature review

2.1. Legal domain large language models and judicial adaptation

The core advantage of legal domain large language models lies in their ability to process dense normative texts, judicial documents, and legal question-answering tasks. Their performance improvement usually depends on legal corpus fine-tuning, professional knowledge injection, and task-specific reconstruction [3]. In judicial applications, models need to understand factual descriptions, legal norms, and adjudicative logic at the same time. Reliance on general language generation may lead to errors in the transfer of legal concepts. Retrieval-augmented generation introduces external authoritative materials to reduce outdated knowledge and missing evidence, allowing model answers to establish more direct evidential connections with statutes, cases, and judicial interpretations [4]. The difficulty of judicial adaptation also comes from the model's ability to handle normative conflicts, case distinctions, and discretionary reasoning.

2.2. Legal hallucination types and liability risks

Legal hallucination in AI-generated legal opinions has multiple layers. It may involve fabricated cases at the factual level, mismatched statutes at the normative level, or reasoning jumps in judicial argumentation. Research such as TruthfulQA shows that large language models may reproduce or amplify false knowledge in fluent answers, and this mechanism becomes more misleading when transferred to professional legal texts [5]. SelfCheckGPT provides a method for identifying hallucination through multi-sample consistency detection, which helps assess whether an answer remains internally stable without extensive manual annotation [6]. Liability risks vary with hallucination types. Fabricated legal grounds may affect system provider liability, erroneous judicial adoption may involve the review duty of legal professionals, and repeated circulation of wrong answers may extend responsibility to platform deployment and user reliance.

2.3. Explainability, transparency, and human oversight

Judicial algorithmic transparency requires model conclusions to be reviewable, challengeable, and verifiable. When AI participates in judicial decision support, procedural legitimacy must cover evidence sources, reasoning processes, and human intervention points. Research on procedural algorithmic justice points out that parties may face insufficient voice and limited remedies in algorithm-assisted decision-making, which gives explainability a procedural participation function [7]. The European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems emphasizes fundamental rights, non-discrimination, quality and security, transparency, and user control, providing a normative framework for judicial AI governance [8]. The EU AI Act further assigns transparency, human oversight, and risk management obligations to high-risk systems, which requires legal large language models in judicial contexts to develop verifiable, recordable, and accountable technical procedures [9].

DOI: 10.54254/2755-2721/2026.35434

3. Experimental methods

3.1. Dataset construction and task design

Data collection focuses on four types of information: legal questions, model-generated answers, authoritative legal evidence, and risk labels. Public sources include the National Laws and Regulations Database, guiding and typical cases released by the Supreme People's Court, publicly available judicial documents, and open judicial interpretation texts. The time span is set from 2019 to 2024. Four high-frequency judicial scenarios are selected, including civil contract disputes, labor disputes, tort liability, and administrative penalties, so that the samples cover fact determination, statutory application, precedent reference, and judicial reasoning generation. During data cleaning, samples with incomplete facts, unverifiable legal bases, or missing reasoning are removed. Each sample is organized into a structured format of "question text, standard evidence, reference answer, and risk type." A total of 1,200 samples are constructed, including 720 training samples, 240 validation samples, and 240 testing samples. The risk labels include fabricated cases, mismatched statutes, distorted rules, and reasoning jumps, providing unified inputs for semantic consistency calculation, citation validity judgment, and dynamic risk scoring.

3.2. Dynamic risk identification model

Dynamic risk identification first decomposes the model answer into several legal claims and then matches each claim with retrieved statutes, cases, judicial interpretations, and judgment summaries [10]. Evidence consistency measures whether the model answer can be supported by authoritative legal texts, as shown in Formula (1).

$$C_i=\frac{1}{m}\sum_{j=1}^{m}\max_{k\in E_i}\cos\left(s_{ij},e_{ik}\right) \tag{1}$$

Here, C_i denotes the evidence consistency score of the i-th sample, s_{ij} denotes the j-th legal claim in the generated answer, e_{ik} denotes the k-th authoritative text segment in the evidence set, and m denotes the number of legal claims. On this basis, citation validity, normative conflict intensity, and reasoning-chain completeness are further integrated into the risk judgment, forming the dynamic judicial hallucination risk score, as shown in Formula (2).

$$R_i=\sigma\left(\alpha_1\left(1-C_i\right)+\alpha_2\left(1-V_i\right)+\alpha_3 D_i+\alpha_4\left(1-L_i\right)\right) \tag{2}$$

Here, R_i denotes the judicial hallucination risk value, V_i denotes citation validity, D_i denotes normative conflict intensity, L_i denotes reasoning-chain completeness, and σ is the Sigmoid function. The risk value is divided into three thresholds: $R_i<0.35$ indicates low risk, $0.35\leq R_i<0.65$ indicates medium risk, and $R_i\geq 0.65$ indicates high risk.

3.3. Evaluation metrics and comparative experiments

The experiment follows six steps: question input, evidence retrieval, answer generation, claim decomposition, risk calculation, and human review. First, legal questions in the test set are input into the legal domain large language model, with the generation temperature fixed at 0.2 to reduce random output variation [11]. Then, BM25 keyword retrieval and dense vector retrieval are combined to return the top 10 candidate legal evidence items from the authoritative corpus. These

DOI: 10.54254/2755-2721/2026.35434

candidates are further filtered by source, publication date, and text similarity. After the model generates an answer, the system decomposes it into factual judgment, normative application, and judicial reasoning claims, and then calculates evidence consistency, citation validity, normative conflict intensity, and reasoning-chain completeness. Four comparative settings are used: ordinary prompting, RAG prompting, self-consistency detection, and dynamic risk identification. Ordinary prompting only uses the legal question, RAG prompting adds retrieved evidence, self-consistency detection compares multiple generated answers, and dynamic risk identification adds risk scoring and explanation output on the basis of RAG. Finally, two annotators with legal training review the risk level. Disputed samples are decided by a third annotator, and the human-reviewed labels serve as the evaluation benchmark.

4. Results

4.1. Risk identification performance

Among the 240 test samples, 92 are high-risk hallucination samples, 81 are medium-risk samples, and 67 are low-risk samples. The identification results of different methods are shown in Table 1. The dynamic risk identification model outperforms the three comparison methods in accuracy, recall, F1-score, AUC, and high-risk recall. The ordinary prompting group obtains an accuracy of 0.68 and a high-risk recall of 0.58, indicating that the model tends to generate complete but insufficiently grounded answers without external evidence constraints. The RAG prompting group improves the accuracy to 0.75, but its ability to identify reasoning jumps and distorted rules remains limited. The self-consistency detection group reaches an F1-score of 0.75 and can detect some unstable generations. The dynamic risk identification model achieves an accuracy of 0.86 and a high-risk recall of 0.89, showing that joint scoring based on evidence consistency, citation validity, and reasoning-chain completeness is more suitable for identifying compound hallucination risks in judicial contexts.

Table 1. Judicial hallucination risk identification results of different methods

Method	Accuracy	Recall	F1-score	AUC	High-risk Recall
Ordinary prompting	0.68	0.61	0.64	0.72	0.58
RAG prompting	0.75	0.70	0.72	0.80	0.69
Self-consistency detection	0.78	0.73	0.75	0.83	0.74
Dynamic risk identification model	0.86	0.84	0.85	0.91	0.89

4.2. Explainability and judicial applicability

After cross-analyzing judicial tasks and hallucination types, different risk sources show clear differences, as shown in Figure 1. In the case existence verification task, the identification rate of fabricated cases reaches 91.4%, indicating that source verification and case-number matching can effectively detect non-existent judicial authorities. In the statutory matching task, the identification rate of incorrect statutes reaches 88.9%, mainly relying on statutory text similarity and applicability-condition comparison. In the holding generation task, the identification rate of rule distortion reaches 83.6%, showing that the model can identify incorrect summaries through normative keywords, judicial reasoning, and outcome correspondence. In the precedent reasoning task, the identification rate of reasoning jumps reaches 86.7%, indicating that the completeness of the chain

DOI: 10.54254/2755-2721/2026.35434

among factual similarity, disputed issues, and judicial outcomes can reflect precedent mismatch risks. These results support the configuration of different review thresholds for case citation, statutory application, and precedent reasoning in AI-assisted judicial systems.

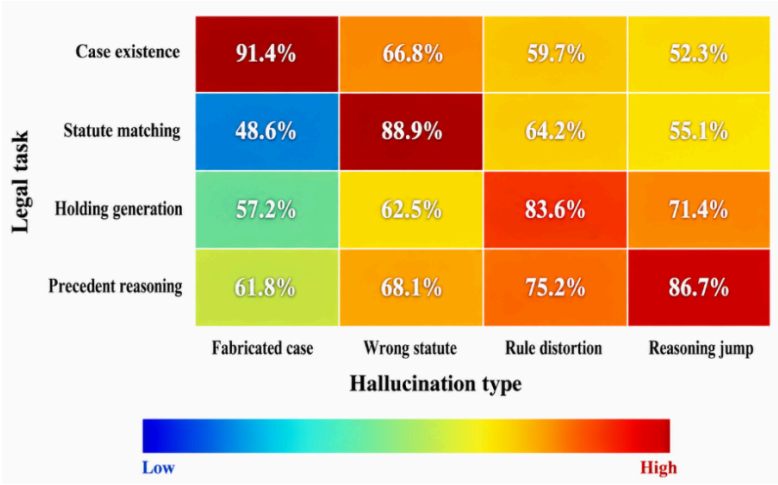

Figure 1. Detection rate matrix of judicial hallucination risks

5. Discussion

The experimental results indicate that legal hallucination risks are not evenly distributed across judicial tasks. They are closely related to task type, evidence structure, and reasoning complexity. Case existence verification can identify fabricated authorities through source checking and case-number matching, while statutory matching depends more on textual similarity and applicability-condition comparison. Precedent reasoning requires the system to examine the logical chain among factual similarity, disputed issues, and judicial outcomes. The advantage of the dynamic risk identification mechanism lies in extending single-text judgment into a combined assessment of evidence consistency, citation validity, normative conflict, and reasoning-chain completeness. Model outputs are therefore not simply judged as correct or incorrect, but classified into different levels, including directly usable, requiring human review, and prohibited from direct output. This also shows that explainability in judicial contexts should serve procedural control. The system must indicate risk sources, evidential gaps, and review priorities so that judges and legal professionals can retain substantive decision-making authority.

6. Conclusion

Focusing on hallucination risks in legal domain large language models used for judicial assistance, this study proposes a dynamic judicial hallucination risk identification mechanism and verifies its feasibility through public legal corpora, task-based sample construction, and comparative experiments. The mechanism integrates legal knowledge retrieval, semantic consistency calculation, citation validity assessment, and reasoning-chain completeness analysis into a unified scoring framework. It can identify major risk types, including fabricated cases, incorrect statutes, distorted rules, and reasoning jumps. The experimental results show that the dynamic risk identification model performs well in high-risk sample recall and can be embedded into warning and human review procedures in AI-assisted judicial systems. Future work may expand test samples across jurisdictions, case types, and real trial contexts, while incorporating feedback from judicial

DOI: 10.54254/2755-2721/2026.35434

professionals to optimize risk thresholds. This will help make the use of legal large language models in digital justice more controllable, transparent, and accountable.

References

[1] Dahl, M., et al. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93.
[2] Magesh, V., et al. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242.
[3] Liu, J. Z., & Li, X. (2024). How do judges use large language models? Evidence from Shenzhen. *Journal of Legal Analysis, 16*(1), 235–262.
[4] Dehghani, F., et al. (2025). Large language models in legal systems: A survey. *Humanities and Social Sciences Communications, 12*(1), 1977.
[5] Lai, J., et al. (2024). Large language models in law: A survey. *AI Open, 5*, 181–196.
[6] Guo, X., et al. (2025). Specialized or general AI? A comparative evaluation of LLMs' performance in legal tasks: X. Guo et al. *Artificial Intelligence and Law*, 1–37.
[7] Dugac, G., & Altwicker, T. (2025). Classifying legal interpretations using large language models: G. Dugac, T. Altwicker. *Artificial Intelligence and Law*, 1–19.
[8] Ribeiro de Faria, J., Xie, H., & Steffek, F. (2025). Information extraction from employment tribunal judgments using a large language model. *Artificial Intelligence and Law*, 1–22.
[9] Buffa, M., et al. (2025). Enhancing legal document building with retrieval-augmented generation. *Computer Law & Security Review, 59*, 106229.
[10] De La Osa, D. U. S., & Remolina, N. (2024). Artificial intelligence at the bench: Legal and ethical challenges of informing—or misinforming—judicial decision-making through generative AI. *Data & Policy, 6*, e59.
[11] Kinchin, N. (2024). "Voiceless": The procedural gap in algorithmic justice. *International Journal of Law and Information Technology, 32*(1), eaae024.