LegalGraphRAG: Multi-Agent Graph Retrieval-Augmented Generation for Reliable Legal Reasoning

Anonymous ACL submission

Abstract

Graph-based Retrieval-Augmented Generation (GraphRAG) advances flat document retrieval by structuring knowledge as relational graphs, enabling more coherent and effective reasoning. However, applying it to specific domains like legal reasoning faces critical challenges. (i) Legal corpora are heterogeneous, containing multi-granular knowledge from cases, articles, and interpretations. A flat knowledge graph cannot adequately differentiate between factual details, applied rules, and abstract principles, limiting accurate retrieval. (ii) Reliable legal judgment demands transparent, evidence-based reasoning. Traditional RAG passes retrieved context directly to an LLM without verification, resulting in opaque, error-prone reasoning. To this end, we propose **LegalGraphRAG**, a framework designed for reliable legal reasoning. Our approach introduces two core components: a hierarchical legal graph that hierarchically organizes legal sources to enable retrieval at appropriate abstraction levels, and a multi-agent system for reliable legal reasoning, where a Researcher retrieves candidate evidence, an Auditor rigorously verifies its validity against source documents, and an Adjudicator synthesizes the set of verified evidence to render a final judgment. Extensive experiments show that LegalGraphRAG achieves the state-of-the-art performance, outperforming existing GraphRAG baselines in accurate and trustworthy legal analysis. Our code, datasets and implementation details are available at https://anonymous.4open.science/r/LegalGraphRAG-E845.

1 Introduction

The rapid advancement of Large Language Models (LLMs), like GPT (Achiam et al., 2023), Gemini (Comanici et al., 2025) and Qwen (Yang et al., 2025a) series, has driven significant progress in intelligent decision-making across various real-world tasks (Zhao et al., 2023; Naveed et al., 2025). However, deploying these models in specialized, knowledge-intensive fields like legal reasoning remains challenging due to the domain's demanding standards of rigor and reliability (Lai et al., 2024; Hou et al., 2025; Siino et al., 2025). Domain-specific tasks necessitate a comprehensive understanding and multi-step reasoning across a vast knowledge base of specialized concepts, rigorous rules, and complex dependencies (Wang et al., 2023; Kim et al., 2025), which requires strict logical reasoning and domain expertise that exceed the capabilities of general-purpose LLMs. While Supervised Fine-Tuning (SFT) (Ouyang et al., 2022; Hu et al., 2022) on domain corpora enables models to internalize that expertise, this approach incurs substantial computational costs and often risks critical catastrophic forgetting in many real-world scenarios (Yue et al., 2024; Luo et al., 2025).

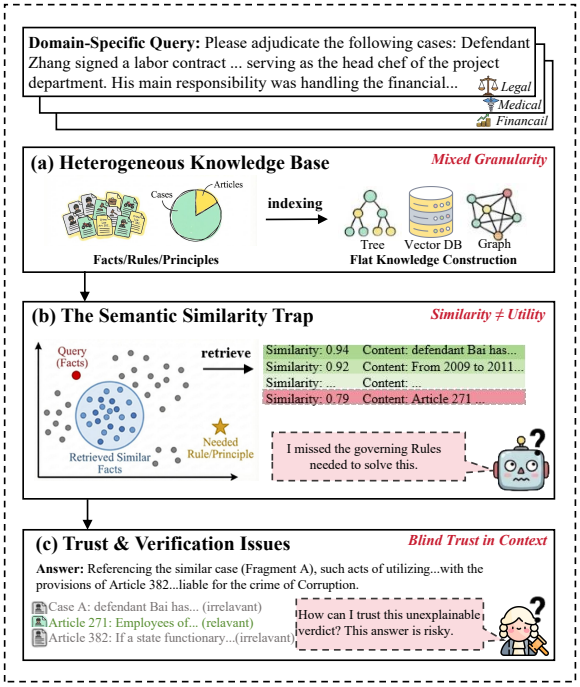

Figure 1: Challenges of Traditional RAG in Domain-Specific Tasks. (i) **Flat Graph Structure**: Struggles to handle heterogeneous documents. (ii) **Unverified Retrieval**: Contains excessive irrelevant information.

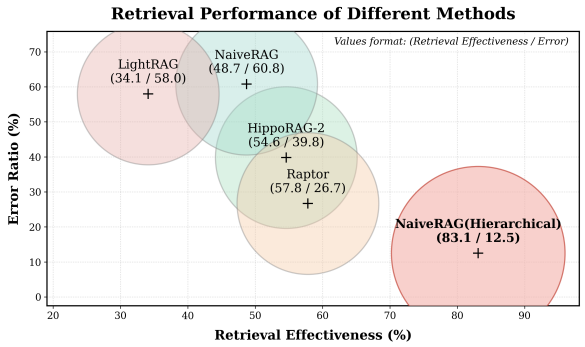

Figure 2: Retrieval performance comparison revealing that conventional RAG methods struggle with heterogeneous domain documents, suffering from high error rates and limited effectiveness. detailed experimental setup is introduced in Section 3.1 and Appendix A.3.

Recently, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Borgeaud et al., 2022; Li et al., 2025) offers a practical solution to adapt LLMs for specific domains. RAG systems enable LLMs to generate responses by leveraging not only their parametric knowledge but also real-time retrieved domain knowledge, thereby providing more accurate and reliable answers (Mallen et al., 2023; Zhang et al., 2025b). However, standard RAG systems typically retrieve information based on semantic similarity (Karpukhin et al., 2020; Chen et al., 2024), treating documents as independent text segments. This hinders complex multi-hop reasoning over hierarchical legal concepts and multiple documents, limiting effectiveness in legal analysis.

Graph-based Retrieval-Augmented Generation (GraphRAG) (Edge et al., 2024; Zhang et al., 2025a) advances this paradigm by organizing domain corpora into structured relational graphs. This structural awareness captures hierarchical relationships between different concepts, thereby enabling more precise retrieval and supporting the multi-hop reasoning required for complex queries. However, directly applying standard GraphRAG to the legal domain faces critical challenges: ❶ A flat graph structure cannot capture the multi-granular hierarchies present in legal corpora, which span factual details, applied rules, and abstract principles across legal cases, articles, and interpretations, thereby limiting accurate retrieval. ❷ Lack of verifiable, evidence-based reasoning. Traditional RAG passes retrieved context directly to an LLM without any verification. This "retrieve-then-generate" pipeline often results in opaque, error-prone reasoning.

In this paper, we propose LegalGraphRAG, a novel framework that synergizes graph-based retrieval with the multi-agent reasoning system for reliable legal reasoning. Specifically, LegalGraphRAG consists of two key components: (i) Hierarchical legal graph (HierarGraph), which organizes legal knowledge into a hierarchical graph to effectively decouple historical cases, relevant statutes, and judicial interpretations, and (ii) a multi-agent system for evidence-based reasoning, where the legal judgment process is structured as a transparent pipeline that retrieves, verifies, and reasons over graph-grounded evidence to produce interpretable decisions. Generally, our contributions are summarized as follows:

- We propose LegalGraphRAG, an evidence-based legal reasoning framework driven by a multi-agent system operating on a hierarchical knowledge graph, which address legal heterogeneity and ensure reliable reasoning.
- We design a hierarchical legal knowledge graph with Ontology, Fact, and Rule layers to model multi-granular legal knowledge and support accurate retrieval.
- We establish a multi-agent system for evidence-based reasoning that performs adjudication through a transparent pipeline of retrieval, validation, and synthesis, grounding judgments in verifiable evidence chains.
- Extensive experiments show that LegalGraphRAG consistently outperforms existing GraphRAG baselines and legal language models in accurate and trustworthy legal analysis.

2 Problem Statement

Complex legal reasoning is formulated as an open-ended generation task evaluating the decision-making capabilities of LLMs within the legal domain. Formally, given a criminal fact description f and a defendant d, a LLM is tasked with predicting the applicable charges y. In this paper, we focus on integrating this reasoning framework with RAG to assess the model's ability to leverage external legal knowledge for judicial reasoning. This task can be organized into the following stages:

Knowledge Organization. Given an offline corpus of legal documents $\mathcal{D}$, including historical cases, articles and interpretations we construct a domain-specific legal knowledge graph:

$$KG = \Phi(\mathcal{D}), \tag{1}$$

where $\Phi(\cdot)$ denotes the organization function.

Knowledge Retrieval. For a legal query characterized by criminal facts f and a defendant d, we retrieve relevant evidence from KG to form a contextual reference:

$$\mathcal{C} = \mathcal{R}(f, d, KG), \tag{2}$$

where $\mathcal{R}(\cdot)$ represents the retrieval operator.

Judgment Generation. Finally, the legal judgment (e.g., charge) y is inferred by reasoning over the query and retrieved evidence:

$$P(y \mid f, d, \mathcal{C}) = \mathcal{G}(f, d, \mathcal{C}), \tag{3}$$

where $\mathcal{G}(\cdot)$ denotes the generator LLM.

3 Preliminary Study

Applying standard retrieval paradigms to the specialized, knowledge-intensive legal domain faces critical challenges due to the inherent structural complexity and rigorous standards of such fields. To illustrate these challenges, we conduct two preliminary experiments to empirically investigate the specific limitations of existing methods regarding knowledge granularity and generation quality.

3.1 Investigation on Knowledge Granularity

Complex domain knowledge possesses an inherent hierarchy. In the legal context, this necessitates distinguishing between abstract statutory principles and concrete case facts. We hypothesize that standard retrieval strategies fail to distinguish between these semantic granularities because they treat all text segments in the same way. To verify this, we compare a *Flat Strategy* against a naive *Hierarchical Strategy* that explicitly segregates articles from case narratives (detailed in Appendix A.3).

As illustrated in Figure 2, the empirical results confirm our hypothesis. *Flat Strategy* exhibit a distinct "granularity bias", frequently prioritizing high-frequency factual details due to surface-level semantic overlaps, often at the expense of essential abstract principles. Conversely, *Hierarchical Strategy* aligns better with the domain's logical structure, improving retrieval performance by 25.3%. This observation suggests that structural flatness constitutes a fundamental bottleneck for standard RAG when handling multi-granular knowledge.

3.2 Investigation on Generation Quality

Reliable domain reasoning demands not only information retrieval but also evidence verification. Real-world legal environments often contain documents that share similar keywords but differ fundamentally in their domain applicability. To simulate this realistic challenge, we conduct a test (detailed in Appendix A.4). Specifically, we inject legally plausible but factually irrelevant documents into the retrieval context to evaluate the model's ability to focus on relevant evidence.

Method	Charge		Articles		Term of Penalty	
	ACC↑ (%)	Δ	ACC↑ (%)	Δ	MAE↓ (months)	Δ
RAG (Correct Context)	42.8	–	74.7	–	24.3	–
RAG + 2 Irrelevant Docs	34.9	↓ 7.9	57.2	↓ 17.5	27.7	↑ 3.4
RAG + 4 Irrelevant Docs	32.9	↓ 9.9	51.1	↓ 23.6	28.4	↑ 4.1
RAG + 6 Irrelevant Docs	29.8	↓ 13.0	46.8	↓ 27.9	31.7	↑ 7.4

Table 1: Performance degradation under varying levels of simulated retrieval noise. ACC (↑) denotes Accuracy for Charge and Articles prediction. MAE (↓) represents Mean Absolute Error for Term of Penalty.

As summarized in Table 1, standard RAG models exhibit significant sensitivity to context purity. The inclusion of irrelevant information precipitates a sharp performance drop. This observation shows that without a dedicated verification mechanism to filter irrelevant content, the model struggles to distinguish valid evidence from misleading information, which undermines reasoning reliability.

3.3 Discussion and Motivation

The findings from these two studies highlight fundamental limitations in applying standard RAG to complex domains: ❶ Flat retrieval mechanisms fail to navigate the hierarchical nature of domain knowledge (e.g., distinguishing rules from facts), resulting in biased context. ❷ The lack of an explicit verification step makes the system fragile to misleading information, which is unacceptable in rigorous fields like law. These insights motivate the design of LegalGraphRAG, which incorporates a *Hierarchical Legal Graph* to resolve granularity conflicts and a *Evidence-based Legal Reasoning* (Researcher-Auditor-Adjudicator) framework to enforce rigorous verification.

4 The Framework of LegalGraphRAG

4.1 Overview

Traditional GraphRAG approaches face limitations in legal judgment due to the heterogeneous and multi-granular nature of legal corpora. To address this challenge, We propose LegalGraphRAG, an evidence-based legal reasoning framework driven by a multi-agent system operating on a hierarchical

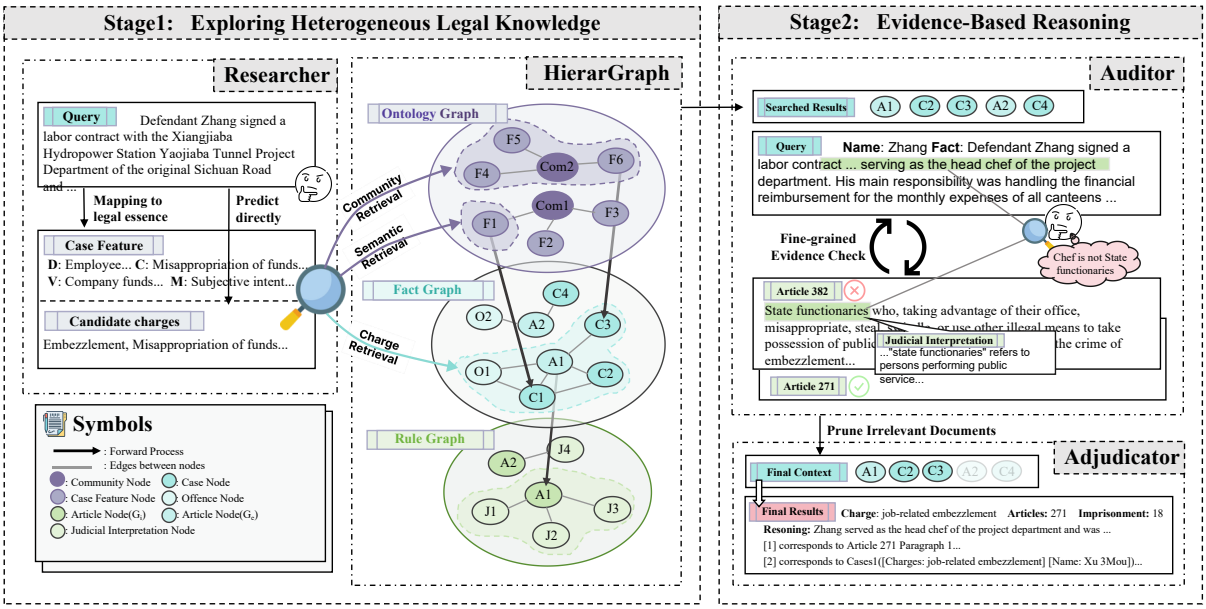

Figure 3: **The architecture of LegalGraphRAG.** The framework consists of two main phases: (1) **Hierarchical Knowledge Construction**, which builds a Hierarchical Legal Graph (HierarGraph) comprising an Fact Graph, Ontology Graph and Rule Graph to organize heterogeneous legal knowledge; and (2) **Evidence-based Legal Reasoning**, where a multi-agent system (Researcher, Auditor, and Adjudicator) performs structured retrieval, validation, and synthesis over the HierarGraph to generate interpretable legal decisions.

knowledge graph. The framework operates in two distinct phases: (i) **Hierarchical Knowledge Construction**, which organizes legal knowledge into a layered graph structure to effectively decouple historical cases, relevant statutes, and judicial interpretations, and (ii) **Evidence-based Legal Reasoning**, structures the legal judgment process as a transparent pipeline that retrieves, verifies, and reasons over graph-grounded evidence to produce interpretable decisions. The whole framework is illustrated in Figure 3.

4.2 Hierarchical Knowledge Construction

Legal reasoning involves heterogeneous information sources, including historical cases, abstract legal articles, and interpretations. Employing a flat storage structure is not enough to handle the inherent structural differences of these data sources, leading to disorganized information and inefficient retrieval. To address this challenge, we construct a Hierarchical Legal Graph (HierarGraph) $\mathcal{H}$ that organizes legal knowledge into distinct semantic layers, enabling explicit differentiation among legal concepts and providing a structured basis for reliable reasoning. The HierarGraph is composed of three specialized subgraphs:

Fact Graph ($\mathcal{G}_{fac}$), which serves as a structured collection of verified legal precedents, providing the essential factual basis for ensuring legally grounded judgments. Accordingly, $\mathcal{G}_{\text{fac}}$ models the natural structure of legal documents by explicitly connecting *Cases* ($\mathcal{C}$), *Articles* ($\mathcal{A}$), and *Offense* ($\mathcal{O}$) nodes. Relationships are established via e_{ca}, linking a case c to its cited article a, and e_{co}, linking a case c to its convicted offense o. This structure provides the factual granularity required for evidence gathering. Formally, it is defined as:

$$\mathcal{G}_{fac} = (\mathcal{V}_{fac}, \mathcal{E}_{fac}) = \left(\{c_i, a_i, o_i\}_{i=1}^{|\mathcal{G}_{\text{fac}}|}\right). \quad (4)$$

Ontology Graph ($\mathcal{G}_{ont}$), which bridges the semantic gap and mitigates noise by abstracting case features. $\mathcal{G}_{ont}$ distills raw narratives containing instance-specific details (e.g., dates and locations) into a purified semantic space that reflects the "legal essence". Specifically, we design a domain-specific legal ontology based on legal theory (Rüthers et al., 2013), encompassing four key dimensions: *Defendant Attributes*, *Criminal Behaviors*, *Victim Characteristics* and *Subjective Mental States*. Keywords and entities are extracted and aligned with these properties to form structured embeddings, serving as indices for *Case Feature Nodes* ($\mathcal{F}$).

To reveal hidden connections between different cases, we employ the k-Nearest Neighbors (k-NN) algorithm to connect nodes with high semantic similarity. We then apply the Leiden algorithm (Traag

et al., 2019) to group related cases into communities, each treated as a *Community Node* ($\mathcal{K}$). Each k contains the summarized information of the cases inside it, facilitating hierarchical retrieval that navigates from broad contexts to specific details. Formally, this subgraph is defined as:

$$\mathcal{G}_{ont} = (\mathcal{V}_{ont}, \mathcal{E}_{ont}) = \left(\{c_i, k_j\}_{i=1,j=1}^{|\mathcal{G}_{ont}|}\right). \quad (5)$$

Rule Graph ($\mathcal{G}_{rul}$), which resolves statutory ambiguities by systematically linking *Articles* ($\mathcal{A}$) with its corresponding *Judicial Interpretations* ($\mathcal{J}$). This explicit alignment establishes the contextual grounding necessary for precise legal reasoning.

Moreover, applying the correct article often depends on specific conditions. A small difference can lead to a completely different judgment for the same crime. (e.g. whether the defendant is an adult or a minor). Simple semantic matching often fails to distinguish these subtle differences. To address this, we equip each a with a *Diagnostic Checklist* ($\mathcal{D}$). This mechanism breaks down complex legal rules into specific verification steps. Formally, this subgraph is defined as:

$$\mathcal{G}_{rul} = (\mathcal{V}_{rul}, \mathcal{E}_{rul}) = \left(\{a_i, j_i\}_{i=1}^{|\mathcal{G}_{rul}|}\right). \quad (6)$$

where

$$\mathcal{D}(a_i) = \{d_1, \ldots, d_{|C|}\} \quad (7)$$

By integrating these three layers, HierarGraph $\mathcal{H}$ transforms heterogeneous legal corpora into a structured ecosystem. This architecture directly addresses the limitations of flat retrieval by offering multi-granular support for following evidence-based legal reasoning. The detailed construction procedures are provided in Appendix B.1.

4.3 Evidence-based Legal Reasoning

To leverage the multi-granular knowledge encoded in our HierarGraph, we propose a multi-agent system for evidence-based reasoning, in which specialized agents sequentially traverse the graph to perform evidence retrieval, validation, and synthesis. Specifically, the workflow consists of three agents:1) *Researcher*, 2) *Auditor*, and 3) *Adjudicator*. Through structured graph traversal and logical analysis, the framework resolves the raw case query by constructing a final, verifiable judgment.

4.3.1 Evidence Retrieval

A reliable evidence-based reasoning process begins with grounding a raw case description in relevant legal evidence. To this end, *Researcher* perform structured evidence retrieval over the $\mathcal{G}_{ont}$ and the $\mathcal{G}_{\text{fac}}$, transforming unstructured case narratives into a coherent set of related *Cases* ($\mathcal{C}$) and *Articles* ($\mathcal{A}$).

Specifically, the Researcher aligns the case description with the ontological dimensions defined in Section 4.2. Based on these features, We formulate the evidence retrieval process $\mathcal{R}(q)$ as the union of three operators, where q is the legal query:

$$\mathcal{R}(q) = \mathcal{R}_{\text{sem}}(q) \cup \mathcal{R}_{\text{com}}(q) \cup \mathcal{R}_{\text{chg}}(q) \quad (8)$$

First, we employ *Semantic Match Retrieval* to locate direct evidence via semantic similarity, where $\phi(\cdot)$ denotes ontology-aligned embeddings:

$$\mathcal{R}_{\text{sem}}(q) = \underset{c \in \mathcal{G}_{ont}}{\text{Top-k}}\, \text{sim}\big(\phi(q), \phi(c)\big) \quad (9)$$

Next, to capture structural context, we conduct *Community Expansion Retrieval*. We first identify the top-ranked communities by topic $\mathcal{S}_{\mathcal{K}}$ aligned with the query, and then retrieve the most similar cases within these communities:

$$\begin{aligned} \mathcal{K}^* &= \underset{\mathcal{K} \in \mathcal{G}_{ont}}{\text{argmax}}\, \text{sim}\big(\phi(q), \phi(\mathcal{K})\big) \\ \mathcal{R}_{\text{com}}(q) &= \underset{c \in \mathcal{K}^*}{\text{Top-k}}\, \text{sim}\big(\phi(q), \phi(c)\big) \end{aligned} \quad (10)$$

Finally, we implement *Charge-Anchored Retrieval* to anchor the legal basis by collecting cases linked to inferred charges. Here, $\mathcal{O}(q)$ denotes the set of predicted charges and $\mathcal{N}$ represents the neighboring cases connected to charge o in $\mathcal{G}_{fac}$:

$$\mathcal{R}_{\text{chg}}(q) = \bigcup_{o \in \mathcal{O}(q)} \mathcal{N}_{\mathcal{G}_{fac}}(o) \quad (11)$$

The specific retrieval algorithms and parameter settings are detailed in Appendix B.2.

4.3.2 Evidence Validation

Given the candidate evidence retrieved in the Evidence Retrieval, this stage focuses on validating whether the case facts genuinely satisfy the conditions required by the law, rather than relying on surface-level semantic relevance.

Specifically, for each candidate article, we verify its applicability by evaluating the case facts using the associated *Diagnostic Checklist* and *Judicial Interpretations* encoded in the $\mathcal{G}_{\text{rul}}$. The verification outcomes are then aggregated to produce a definitive applicability judgment for each article.

Based on these judgments, *Auditor* filters the retrieval subgraph by pruning inapplicable articles and their associated case and charge nodes. Finally, it organizes the remaining nodes into a legally consistent and evidence-supported subgraph, which

Model	Size	CAIL								CMDL								Average	
		Public Safety		Economic		Social Order		Person Rights		Public Safety		Economic		Social Order		Person Rights			
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	All	Δ
Open-Source Models																			
Qwen-2.5-7B-Instruct	7B-Inst	24.0	45.8	23.1	42.5	22.9	36.7	27.4	46.0	25.8	32.4	28.7	35.8	27.2	42.1	32.8	49.6	26.7	↑ 22.8
Qwen-3-8B	8B-Inst	31.7	49.2	25.8	42.7	26.3	39.8	27.6	47.8	44.0	52.3	44.7	53.1	42.7	51.9	53.0	57.7	35.2	↑ 19.9
Internlm3-8b-instruct	8B-Inst	29.8	49.1	26.7	42.0	25.2	34.3	28.1	47.3	25.4	32.1	35.7	37.0	27.5	36.2	34.1	53.6	26.6	↑ 22.9
Glm-4-9b-chat	9B-Inst	18.4	33.7	19.7	36.1	15.8	32.1	26.0	44.5	23.5	34.2	23.6	40.8	19.1	37.0	41.5	47.0	21.2	↑ 28.2
Closed-Source Models																			
GPT-4o-mini	∼8B	19.7	35.5	19.6	33.3	15.5	35.2	29.0	46.3	18.0	28.0	22.7	31.9	21.9	32.0	35.9	50.3	28.4	↑ 21.1
DeepSeek-V3.1	∼200B	31.0	51.3	29.0	48.4	29.8	50.2	35.2	54.8	35.0	64.0	54.7	62.7	58.2	61.9	62.5	71.6	42.8	↑ 6.7
Legal Specific Methods																			
DISC-LawLLM-7B	7B-Inst	40.1	50.9	31.0	51.5	34.8	47.7	34.5	56.0	49.7	53.6	39.6	52.1	30.3	49.5	48.4	63.3	30.3	↑ 19.1
ADAPT	7B-Inst	38.7	43.7	32.7	43.4	27.6	41.7	35.2	50.7	54.5	58.8	57.1	59.4	40.9	43.4	61.5	62.1	42.8	↑ 6.7
Legal Δ	7B-Inst	40.8	50.6	25.1	37.4	32.1	43.7	34.1	53.6	58.3	61.5	51.8	55.8	50.2	54.8	65.8	64.4	42.4	↑ 7.1
RAG Based Methods																			
Naive RAG	8B-Inst	31.0	45.7	24.4	38.7	28.1	38.4	34.5	46.8	45.8	57.3	44.8	55.2	46.8	58.5	49.6	57.8	33.3	↑ 16.1
G-retriever	8B-Inst	33.8	48.0	26.0	39.8	23.8	39.3	32.6	50.1	36.8	40.0	42.5	48.8	45.3	50.7	46.2	52.4	34.4	↑ 13.2
LightRAG	8B-Inst	20.4	43.6	21.7	42.5	19.0	42.5	26.9	50.6	37.9	50.1	43.2	45.1	44.2	51.3	43.7	46.9	30.5	↑ 19.0
RAPTOR	8B-Inst	34.6	50.4	31.6	43.9	32.1	45.6	32.4	45.7	53.8	62.6	53.6	60.1	52.5	62.8	52.1	66.9	43.1	↑ 6.3
HippoRAG2	8B-Inst	34.5	38.2	24.0	33.5	28.8	35.0	31.0	36.3	53.5	56.5	50.6	52.7	53.5	55.0	62.4	62.8	43.1	↑ 6.3
LegalGraphRAG (Ours)	8B-Inst	**42.9**	**54.3**	**38.5**	**53.6**	**37.6**	**51.1**	**37.2**	**58.3**	**65.5**	**66.5**	**59.8**	**65.1**	**58.5**	**63.7**	**70.1**	**72.7**	**49.5**	–

Table 2: **Performance comparison on CAIL and CMDL.** We employ Qwen3-8B as the default backbone model. The best results are highlighted in **bold**, and the second-best are underlined. We visualize the gains of LegalGraphRAG over each baseline in the Δ columns.

serves as a validated knowledge basis for subsequent decision-making. Further implementation details can be found in Appendix B.2.

4.3.3 Evidence Synthesis

In the final stage, the validated evidence produced in the previous steps is synthesized to derive a legally grounded judgment. Based on the verified subgraph, *Adjudicator* integrates the confirmed articles ($\mathcal{A}^f$), cases ($\mathcal{C}^f$), and offense information ($\mathcal{O}^f$) to determine the applicable charges and their statutory basis. This process is formulated as:

$$\mathcal{J} = Adjudicator(q \oplus \mathcal{A}^f \oplus \mathcal{C}^f \oplus \mathcal{O}^f) \quad (12)$$

Crucially, the judgment is not produced as a direct verdict. Instead, it is accompanied by explicit citations to the statutory articles and judicial interpretations used in the reasoning process, ensuring that every conclusion is directly traceable to verified evidence in the HierarGraph.

Overall, LegalGraphRAG formulates legal judgment as a transparent, evidence-based reasoning pipeline rather than a black-box generation process. Through sequential evidence grounding, validation, and synthesis, the system enforces stepwise verification and ensures that every conclusion is explicitly derived from and supported by verified legal evidence, resulting in reliable judicial decisions.

5 Experiment

This section presents a comprehensive evaluation of LegalGraphRAG on two legal judgment benchmarks. Our experiments are designed to answer the following three questions. **Q1 (Generation Accuracy)**: Does LegalGraphRAG outperform SOTA GraphRAG methods and leading legal-domain LLMs in generation quality? **Q2 (Case Study)**: How does LegalGraphRAG handle specific legal cases, and does it provide more interpretable outputs compared to baselines? **Q3 (Ablation Study)**: What is the contribution of each core component to the final performance of LegalGraphRAG? More additional experiments are provided in Appendix C

5.1 Experiment Setup

Datasets We evaluate on two widely used legal benchmarks: CAIL2018 (Xiao et al., 2018) and CMDL (Huang et al., 2024), covering diverse criminal sub-fields such as Public Safety, Social Order, Economic Offenses, and Person Rights. The retrieval knowledge base is built from a collection of authoritative legal sources, including case datasets and statutory texts. Further dataset details are provided in Appendix D.1 & D.2.

Baselines To ensure a comprehensive evaluation, we categorize our comparative experiments into four distinct groups: (i) Open-Source Models, utilizing Qwen-series (Yang et al., 2025a), InternLM (Fei et al., 2025) and GLM (GLM et al., 2024) as foundational backbones; (ii) Closed-Source Models, represented by GPT-4o-mini (Achiam et al., 2023) and DeepSeek-V3.1 (Liu et al., 2024). (iii) Legal-Specific Methods, which include domain-specialized approaches

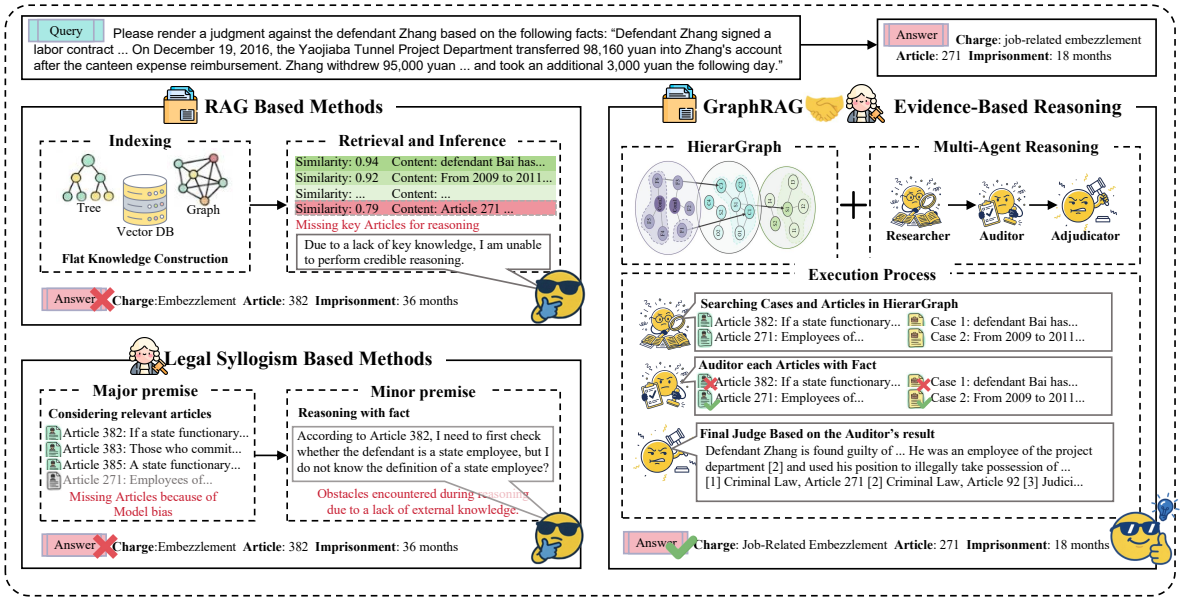

Figure 4: **A comparative case study** illustrating the reasoning trajectories of different methods. While Naive RAG fails due to missing legal articles and syllogism-based methods struggle with ambiguities, LegalGraphRAG derives the correct judgment. By leveraging the HierarGraph and Evidence-based Legal Reasoning, our framework demonstrates transparency and reliability, providing a verifiable reasoning chain grounded in legal evidence.

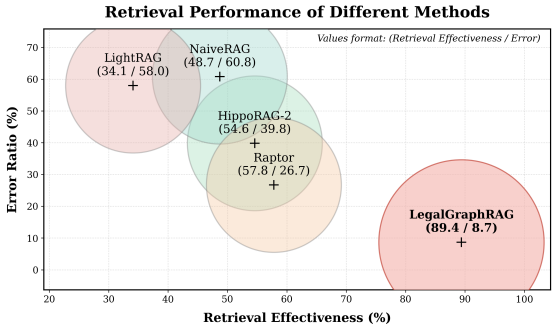

Figure 5: **Retrieval Performance Comparison.** LegalGraphRAG demonstrates superior retrieval effectiveness and significantly lower error ratios compared to conventional flat graph baselines.

such as Disc-LLM (Yue et al., 2024), LegalΔ (Dai et al., 2025), and ADAPT (Deng et al., 2024b); and (iv) RAG-Based Methods, encompassing Naive RAG and advanced graph-augmented strategies like G-retriever (He et al., 2024a), RAPTOR (Sarthi et al., 2024), LightRAG (Guo et al., 2024), and HippoRAG2 (Gutiérrez et al., 2025). Detailed configurations are provided in Appendix D.4.

Evaluation Metrics We employ Accuracy and Micro-F1 score to evaluate prediction performance. Detailed definitions are provided in Appendix D.3.

Implementation Details We utilize GPT-4o-mini for graph construction and BGE-m3 (Chen et al., 2024) for embedding generation. Various LLMs serve as backbone models for the reasoning phase. We employ Qwen3-8B (Yang et al., 2025a) as the default backbone model for our main experiments. Full hyperparameter settings and hardware specifications are detailed in Appendix D.5.

5.2 Generation Accuracy (Q1)

To address Q1, we evaluate LegalGraphRAG against SOTA RAG methods and specialized legal LLMs on two legal judgment datasets. The primary comparison results for charge prediction are reported in Table 2, with extended analyses in Tables 4, 5, and 6 in Appendix. We summarize the key observations below.

Obs.1. LegalGraphRAG consistently outperforms baselines in legal datasets. Our method achieves the best results on most evaluation metrics across both datasets. Notably, LegalGraphRAG delivers significant improvements ranging from 6.3% to 19.1% over the strongest baselines. Unlike standard GraphRAG methods that struggle in the legal domain, our approach effectively structures heterogeneous knowledge, thereby enhancing legal reasoning capabilities and improving charge prediction accuracy overall.

Obs.2. LegalGraphRAG substantially surpasses existing specialized legal LLMs. Our approach outperforms Legal Δ and ADAPT by an average of 7.1% and 6.7%, respectively. Moreover, as shown in Table 4 in Appendix, LegalGraphRAG integrates flexibly with different backbone models, achieving a peak performance of 78.7% on CMDL when com-

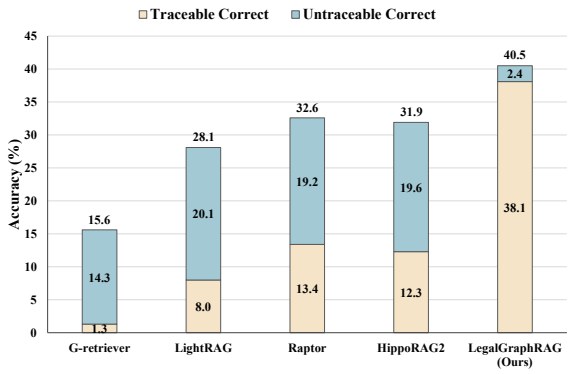

Figure 6: **Reliability Analysis.** LegalGraphRAG significantly increases the proportion of Traceable Correct samples, effectively minimizing Untraceable Correct predictions where the answer is correct but lacks supporting evidence in the retrieved context.

Settings	CAIL ACC	CAIL Δ
LegalGraphRAG (Full)	**40.9**	–
w/o HierarGraph	33.7	↓ 7.2
w/o Researcher	36.9	↓ 4.0
w/o Semantic Match	39.1	↓ 1.8
w/o Community Exp.	38.5	↓ 2.4
w/o Charge-Anchored	39.3	↓ 1.6
w/o Auditor	37.5	↓ 3.4

Table 3: **Ablation study** of LegalGraphRAG components on the CAIL dataset. Results underscore the indispensable role of the HierarGraph for knowledge organization and the synergy between the Researcher and Auditor agents in ensuring reasoning accuracy.

bined with strong backbones. This demonstrates strong adaptability and robust reasoning compared to specialized legal-domain baselines.

5.3 Case Study (Q2)

To demonstrate the superior interpretability of our framework, we present a qualitative analysis of a representative criminal case in Figure 4. More cases are provided in Appendix E.

Obs.3. LegalGraphRAG retrieves significantly more relevant and comprehensive evidence. As illustrated in Figure 5, conventional flat graph structures (e.g., HippoRAG2) struggle to handle heterogeneous legal documents, often failing to capture essential statutes. This structural limitation leads to fragmented context. In contrast, our hierarchical organization effectively structures legal knowledge, ensuring that the retrieved context is sufficient to support robust reasoning.

Obs.4. LegalGraphRAG guarantees decision traceability through rigorous evidence grounding. While baseline models often achieve correct predictions, our reliability analysis (Figure 6) reveals a critical issue of "unsupported correctness", where the model predicts the right charge but fails to retrieve the necessary supporting evidence. This implies that the prediction is not supported by relevant evidence or a valid reasoning chain. LegalGraphRAG significantly increases the ratio of "Traceable Correct" samples (defined in Appendix A.5). By enforcing strict verification, our system ensures that every statute cited in the judgment is explicitly present in the retrieved context, transforming opaque predictions into transparent, traceable decisions.

5.4 Ablation Study (Q3)

To quantify the impact of each component, we performed a systematic ablation study by removing specific modules from the full LegalGraphRAG framework. Results are detailed in Table 3.

Obs.5. Hierarchical structure is the cornerstone of performance. Removing the hierarchical graph (*w/o HierarGraph*) causes the sharpest accuracy drop of 7.2%. This confirms that separating concrete facts from abstract rules into distinct granular levels is essential, providing structural precision that flat indexing lacks.

Obs.6. The multi-agent workflow guarantees reasoning reliability. Excluding the *Researcher* and *Auditor* degrades accuracy by 4.0% and 3.4%, respectively. This validates their synergistic roles: the *Researcher* maximizes evidence coverage through diverse retrieval strategies, while the *Auditor* enforces rigorous verification, ensuring only validated evidence supports the judgment.

6 Conclusion

In conclusion, we have presented LegalGraphRAG, an evidence-based legal reasoning framework that addresses the critical challenges of legal heterogeneity and reasoning reliability. By integrating a hierarchical knowledge graph with a collaborative multi-agent system, our approach transforms the legal reasoning process into a transparent pipeline of retrieval, verification, and synthesis. Extensive experiments on legal judgment benchmarks validate that LegalGraphRAG establishes a new state-of-the-art, significantly advancing accurate and trustworthy AI for reliable and complex legal analysis.

Limitation

While LegalGraphRAG demonstrates significant proficiency in processing textual legal documents and statutes, its current scope is confined to unimodal textual inputs. Real-world judicial proceedings, however, often rely on a heterogeneity of evidence types, including crime scene photography, surveillance footage, scanned handwritten documents, and audio recordings of court hearings. Currently, our framework requires all non-textual evidence to be transcribed or described textually before processing, which may result in the loss of critical visual or auditory nuances essential for fact verification. For instance, distinguishing between "inten" and "negligence" might sometimes rely on visual cues in surveillance video that textual descriptions fail to capture fully. Extending the *Hierarchical Legal Knowledge Graph* to incorporate multimodal nodes (e.g., embedding visual evidence into the *Fact Graph*) represents a promising avenue for future research. Such an extension would enable the model to perform cross-modal reasoning, verifying textual testimony against visual evidence, thereby moving closer to a holistic and robust "Smart Court" system.

Ethics Statement

We confirm that this study fully complies with the ACL Ethics Policy. Below, we address specific ethical considerations regarding the data and the application of our proposed model, LegalGraphRAG.

Data Privacy and Compliance Our experiments involve four publicly available datasets (CAIL2018, CMDL, JuDGE, and LeCaRDv2) and statutory texts. These resources are established benchmarks in the legal NLP community. We emphasize that all court judgments utilized in this work have been pre-processed and anonymized by the original data providers. Private details, including the real names of defendants and victims, have been removed or masked to ensure no personally identifiable information (PII) is exposed. We strictly use this data for academic research purposes and adhere to their respective data usage licenses.

Bias and Fairness We acknowledge that models trained on historical legal judgment data may inadvertently capture or amplify inherent biases present in the judicial system, such as those related to region or gender. While our work focuses on improving the logical reasoning and retrieval capabilities of legal LLMs through GraphRAG, where the outputs are interpreted with clear evidence.

Intended Use and Misuse The proposed LegalGraphRAG is designed as an assistive tool to support legal professionals and researchers in retrieving precedents and analyzing case facts. It is **not** intended to replace human judges or lawyers, nor should it be deployed as a fully automated decision-making system in real-world judicial scenarios. The "prison term" and "judgment" predictions generated by the model should be viewed as reference probabilities rather than enforceable verdicts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*. PMLR.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*.

Xin Dai, Buqiang Xu, Zhenghao Liu, Yukun Yan, Huiyuan Xie, Xiaoyuan Yi, Shuo Wang, and Ge Yu. 2025. Legal δ: Enhancing legal reasoning in llms via

reinforcement learning with chain-of-thought guided information gain. *arXiv preprint arXiv:2508.12281*.

Hudson de Martim. 2025. Graph rag for legal norms: A hierarchical and temporal approach. *arXiv preprint arXiv:2505.00039*.

Chenlong Deng, Kelong Mao, and Zhicheng Dou. 2024a. Learning interpretable legal case retrieval via knowledge-guided case reformulation. *arXiv preprint arXiv:2406.19760*.

Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024b. Enabling discriminative reasoning in llms for legal judgment prediction. *arXiv preprint arXiv:2407.01964*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Jidong Ge, and Vincent Ng. 2025. Internlm-law: An open-sourced chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Sudipto Ghosh, Devanshu Verma, Balaji Ganesan, Purnima Bindal, Vikas Kumar, and Vasudha Bhatnagar. 2024. Inlegalllama: Indian legal knowledge enhanced large language model. In *International Joint Conference on Artificial Intelligence*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

Zhang Han and Dou Zhicheng. 2023. Case retrieval for legal judgment prediction in legal artificial intelligence. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024a. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024b. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Mengzhe Hei, Qingbao Liu, Sheng Zhang, Honglin Shi, Jiashun Duan, and Xin Zhang. 2024. A heterogeneous graph based on legal documents and legal statute hierarchy for chinese legal case retrieval. *IEEE Access*, 12:93502–93516.

Justin Ho, Alexandra Colby, and William Fisher. 2025. Incorporating legal structure in retrieval-augmented generation: A case study on copyright fair use. *arXiv preprint arXiv:2505.02164*.

Zhitian Hou, Zihan Ye, Nanli Zeng, Tianyong Hao, and Kun Zeng. 2025. Large language models meet legal artificial intelligence: A survey. *arXiv preprint arXiv:2509.09969*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. 2024. Cmdl: A large-scale chinese multi-defendant legal judgment prediction dataset. In *Findings of the Association for Computational Linguistics ACL 2024*.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Hyunjae Kim, Jiwoong Sohn, Aidan Gilson, Nicholas Cochran-Caggiano, Serina Applebaum, Heeju Jin, Seihee Park, Yujin Park, Jiyeong Park, Seoyoung Choi, and 1 others. 2025. Rethinking retrieval-augmented generation for medicine: A large-scale,

systematic expert evaluation and practical insights. *arXiv preprint arXiv:2511.06738*.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025. Lexrag: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. 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*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. 2023. Finding the law: Enhancing statutory article retrieval via graph neural networks. *arXiv preprint arXiv:2301.12847*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Xiao Peng and Liang Chen. 2024. Athena: Retrieval-augmented legal judgment prediction with large language models. *arXiv preprint arXiv:2410.11195*.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

B. Rüthers, C. Fischer, and A. Birk. 2013. *Rechtstheorie mit juristischer Methodenlehre*. Grundrisse des Rechts. C.H. Beck.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on information and knowledge management*.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.

Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. 2025. Judge: Benchmarking judgment document generation for chinese legal system. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2024. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on chinese legal domain. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, and 1 others. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Xuran Wang, Xinguang Zhang, Vanessa Hoo, Zhouhang Shao, and Xuguang Zhang. 2024. Legalreasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration. *IEEE Access*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Hannes Westermann. 2024. Dallma: Semi-structured legal reasoning and drafting with large language models. In *2nd Workshop on Generative AI and Law*.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*. Springer.

Shiguang Wu, Zhongkun Liu, Zhen Zhang, Zheng Chen, Wentao Deng, Wenhao Zhang, Jiyuan Yang, Zhitao Yao, Yougang Lyu, Xin Xin, Shen Gao, Pengjie Ren, Zhaochun Ren, and Zhumin Chen. 2023a. fuzi.mingcha.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023b. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. *arXiv preprint arXiv:2310.09241*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and 1 others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv:2004.02557*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rui Yang. 2024. Casegpt: a case reasoning framework based on language models and retrieval-augmented generation. *Preprint*, arXiv:2407.07913.

Xinyu Yang, Chenlong Deng, and Zhicheng Dou. 2025b. Glare: Agentic reasoning for legal judgment prediction. *arXiv preprint arXiv:2508.16383*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *Preprint*, arXiv:2212.01326.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and 1 others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, and 1 others. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications*. Springer.

Jianqiiu Zhang. 2024. Should we fear large language models? a structural analysis of the human reasoning system for elucidating llm capabilities and risks through the lens of heidegger's philosophy. *arXiv preprint arXiv:2403.03288*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, and 1 others. 2025a. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv preprint arXiv:2406.04614*.

A Frequently Asked Questions (FAQs)

A.1 What are the advantages of LegalGraphRAG?

LegalGraphRAG introduces several key advancements over traditional retrieval-augmented generation methods and specialized legal LLMs, addressing critical challenges in the legal domain through its hierarchical structure and multi-agent workflow.

Superior Retrieval Effectiveness. First, our framework significantly improves how legal information is retrieved. While traditional flat retrieval methods often struggle to differentiate between specific case facts and abstract statutory rules, our hierarchical graph organizes this complex information into distinct levels. This structure ensures that the system captures both detailed evidence and high-level principles, providing a much more comprehensive context than standard baselines.

Trustworthy and Transparent Reasoning. Second, LegalGraphRAG addresses the "black box" issue common in standard LLMs. Instead of generating answers directly, which can lead to hallucinations or correct predictions based on wrong premises, our system employs a multi-agent workflow. This process strictly verifies the retrieved evidence against the facts of the case. Consequently, it constructs a logical chain of evidence, ensuring that the final judgment is grounded in valid legal logic rather than statistical probability.

Flexibility and Model Agnosticism. Finally, the framework offers superior flexibility compared to rigid, specialized legal models. Unlike methods that require extensive and costly fine-tuning on legal datasets, LegalGraphRAG functions as a modular system. It allows users to easily swap the underlying backbone model. As demonstrated in our experiments, this capability enables the integration of powerful closed-source models to achieve state-of-the-art performance without the need for additional training.

A.2 How to Evaluate Legal Reasoning?

We utilize Legal Judgment Prediction (LJP) as the primary experimental testbed because it serves as a rigorous "cognitive touchstone" for evaluating complex reasoning in specialized domains. While our introduction highlights broader challenges in healthcare, finance, and law, LJP uniquely encapsulates the core difficulties of high-stakes reasoning.

Validating Hierarchical Knowledge Alignment Introduction Challenge (i) highlights the difficulty of managing heterogeneous knowledge. LJP exemplifies this struggle by requiring the model to bridge the semantic gap between concrete case facts and abstract statutory rules. Successfully mapping these distinct granularities validates the effectiveness of our hierarchical graph structure in organizing multi-level domain knowledge.

Evaluating Rigorous Logical Deduction Unlike general Question Answering tasks that may rely on surface-level semantic matching, LJP necessitates strict syllogistic reasoning (Major Premise $\rightarrow$ Minor Premise $\rightarrow$ Conclusion). This structural dependency provides an ideal setting to stress-test our Multi-Agent framework, specifically validating whether the *Auditor* can effectively filter irrelevant distractions and enforce the logical consistency.

Benchmarking High-Stakes Reliability In professional domains, plausibility is insufficient; accuracy is paramount. LJP imposes a zero-tolerance standard for hallucination, as every judgment must be supported by cited articles. By demonstrating that LegalGraphRAG can produce verifiable, evidence-based judgments in this demanding context, we establish a strong precedent for its applicability to critical domains like medicine and finance.

A.3 How was the retrieval performance evaluated and compared across different strategies?

To ensure consistent assessment throughout our study (spanning both the preliminary investigation and main comparative experiments), we established a standardized evaluation pipeline based on the legal corpora and CAIL (Xiao et al., 2018) dataset described in Section D.2. The evaluation procedure consists of three steps.

Execution on Test Set For every case query in the test dataset, we executed two representative RAG strategies. The *Flat Strategy* follows the traditional baseline approach, indexing all legal documents in a unified flat repository. The *Hierarchical Strategy*, built upon Naive RAG, adopts a decoupled approach by separately storing and retrieving legal articles and historical cases. Based on these strategies, each model retrieved a set of candidate evidence from the corpus.

Ground Truth Alignment We utilized the articles provided in the dataset (ground truth articles)

as the "Gold Standard," as all cases in the dataset are inherently annotated with their relevant statutory articles. Any retrieved node matching these articles was marked as a *True Positive*.

Metric Calculation Based on the alignment results, we quantified performance using the two key indicators defined in Section D.3:

- *Retrieval Effectiveness:* This measures the Recall of gold-standard evidence, indicating the system's ability to locate legal evidence.
- *Error Rate:* This assesses the proportion of irrelevant or misleading nodes within the retrieved context reflecting the system's ability to filter distractions.

This allows us to objectively compare how different structural approaches (flat vs. hierarchical) impact the precision of legal reasoning.

A.4 How were the "context with irrelevant information" constructed for the Generation Quality investigation?

To rigorously test the model's verification capabilities, we constructed evaluation contexts containing High-Similarity Irrelevant Information. Instead of including randomly selected texts, we curated sets of documents that are semantically similar to the correct evidence but legally inapplicable. This design mirrors real-world scenarios where documents share surface-level keywords but differ fundamentally in domain applicability. The construction process involved two steps:

Ground Truth Context: First, we established the baseline context using the ground truth provided in the CAIL (Xiao et al., 2018) dataset. For each case, this set consists exclusively of the correct applicable articles required for the judgment.

Injection of Irrelevant Distractors: To simulate the presence of legally plausible but factually irrelevant documents, we utilized the entire Criminal Law code as a retrieval corpus. For each correct article in the Ground Truth Context, we performed a vector-based similarity search over this corpus to identify the top-k most similar articles that were not part of the ground truth.

These retrieved articles serve as High-Similarity Distractors: they share significant lexical and semantic overlap with the correct laws (e.g., sharing keywords like "theft" or "fraud") but differ in specific constitutive elements or sentencing standards. By mixing these irrelevant documents into the context, we created a challenging environment that forces the model to discern legal essence from superficial similarity.

A.5 Reliability Analysis Definitions

We analyzed the CAIL test results to categorize correct predictions based on evidence support. A prediction is classified as Traceable Correct if the model correctly predicts the charge and successfully retrieves the ground-truth articles. Conversely, it is Untraceable Correct if the correct charge is predicted despite failing to retrieve the necessary articles.

B Method Details

In this section, we provide the comprehensive technical specifications and implementation details of the proposed **LegalGraphRAG** framework. As outlined in the main text, our approach operates in two distinct phases: (i) **Hierarchical Knowledge Construction**, which organizes legal knowledge into a layered graph structure to effectively decouple historical cases, relevant articles, and judicial interpretations; and (ii) **Evidence-based Legal Reasoning**, which employs a collaborative agent workflow to retrieve relevant evidence and generate verifiable judgments.

B.1 Hierarchical Knowledge Construction

We construct a Hierarchical Legal Graph $\mathcal{G}$ composed of three specialized subgraphs, as illustrated in Figure 3. This multi-layered structure explicitly differentiates between specific precedents, abstract case relationships, and rigorous statutory rules.

Fact Graph ($\mathcal{G}_{fac}$) serves as the repository for ground-truth precedents. It encodes the natural structure of legal documents by explicitly linking Cases ($\mathcal{C}$), Articles ($\mathcal{A}$), and Offenses ($\mathcal{O}$). Edges are established to represent citation relationships ($e_{ca} : \mathcal{C} \to \mathcal{A}$) and conviction outcomes ($e_{co} : \mathcal{C} \to \mathcal{O}$). Formally, it is defined as:

$$\mathcal{G}_{fac} = (\mathcal{V}_{fac}, \mathcal{E}_{fac}) = \left(\{c_i, a_i, o_i\}_{i=1}^{|\mathcal{G}_{fac}|} \right). \tag{13}$$

Ontology Graph ($\mathcal{G}_{ont}$) abstracts case features to model inter-case relationships. To map unstructured narratives into a structured semantic space, we define a domain-specific ontology along four dimensions: *Defendant Attributes*, *Criminal Behaviors*, *Victim Characteristics*, and *Subjective Mental*

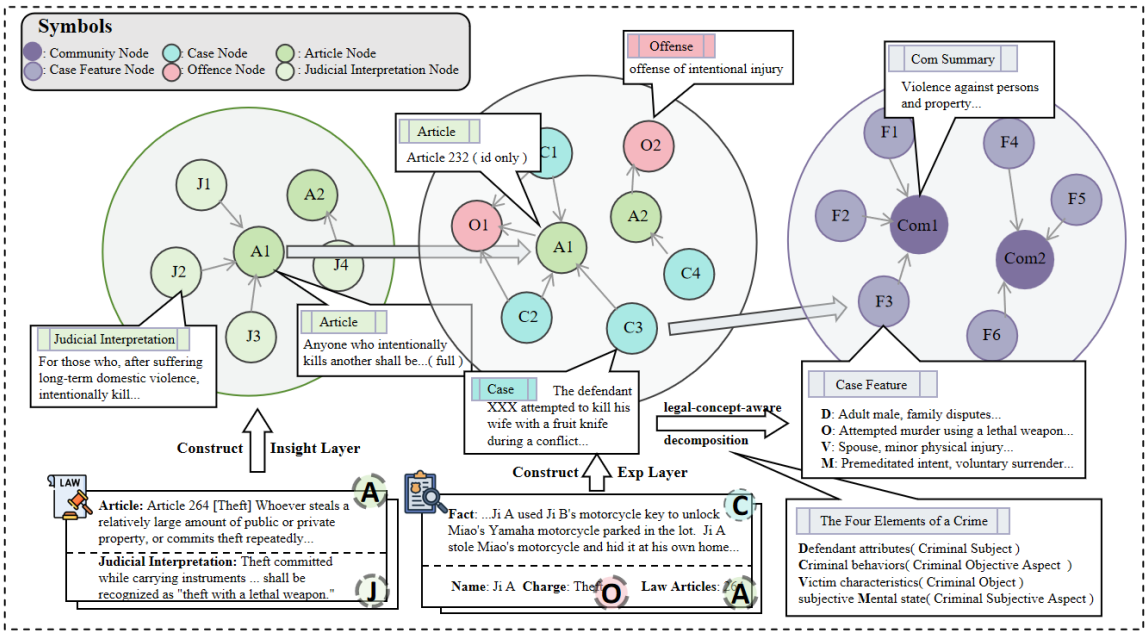

Figure 7: **Case study** of LegalGraphRAG compared to baselines. Their execution processes are compared in detail.

States. Keywords are extracted and aligned with these dimensions to form Case Feature Nodes ($\mathcal{F}$).

Structurally, we utilize the k-Nearest Neighbors (k-NN) algorithm to establish semantic edges between cases. Based on this topology, we apply the Leiden algorithm (Traag et al., 2019) to cluster related cases into Community Nodes ($\mathcal{K}$), facilitating coarse-to-fine retrieval. The subgraph is formally defined as:

$$\mathcal{G}_{ont} = (\mathcal{V}_{ont}, \mathcal{E}_{ont}) = \left(\{c_i, k_j\}_{i=1,j=1}^{|\mathcal{G}_{ont}|}\right). \quad (14)$$

Rule Graph ($\mathcal{G}_{rul}$) incorporates fine-grained legal knowledge to resolve statutory ambiguities. This graph consists of Articles ($\mathcal{A}$) and Judicial Interpretations ($\mathcal{J}$), linked by explicit cross-references. To further enhance precision, each article node a_i is equipped with a **Diagnostic Checklist** $\mathcal{D}(a_i)$.

Generated by parsing statutory texts, this checklist decomposes complex legal provisions into atomic boolean queries. For instance, regarding *Article 266* (Fraud), the checklist validates the logical chain of the crime: "*Did the defendant fabricate facts or conceal the truth?*", "*Did the victim fall into a mistake due to this act?*", and "*Did the victim dispose of property based on this mistake?*". This mechanism forces the model to verify each constitutive element step-by-step, rather than relying on vague semantic overlaps. Formally, this subgraph and its associated checklists are defined as:

$$\mathcal{G}_{rul} = (\mathcal{V}_{rul}, \mathcal{E}_{rul}) = \left(\{a_i, j_i\}_{i=1}^{|\mathcal{G}_{rul}|}\right), \quad (15)$$

where

$$\mathcal{D}(a_i) = \{d_1, \ldots, d_{|C|}\}. \quad (16)$$

By integrating these three layers, HierarGraph $\mathcal{G}$ transforms heterogeneous legal corpora into a structured ecosystem. This architecture directly addresses the limitations of flat retrieval by offering multi-granular support for our multi-agent system.

B.2 Evidence-based Legal Reasoning

We propose a multi-agent framework to emulate the rigorous workflow of legal professionals. This system operates sequentially through three specialized agents (**Researcher**, **Auditor**, and **Adjudicator**) to transform a case query into a verifiable judgment.

Researcher Agent: This agent is responsible for grounding the unstructured case query in relevant legal knowledge. First, it aligns the raw case description with the ontology structure in $\mathcal{G}_{ont}$, extracting standardized evidentiary features (e.g., defendant characteristics and criminal behaviors).

Based on these features, we formulate the evidence retrieval process $\mathcal{R}(q)$ as the union of three parallel strategies, where q is the legal query:

$$\mathcal{R}(q) = \mathcal{R}_{\text{sem}}(q) \cup \mathcal{R}_{\text{com}}(q) \cup \mathcal{R}_{\text{chg}}(q) \quad (17)$$

(i) **Semantic Match Retrieval**: We first locate direct evidentiary analogues via fine-grained semantic similarity. Let $\phi(\cdot)$ denote the ontology-aligned embeddings, this process is defined as:

$$\mathcal{R}_{\text{sem}}(q) = \underset{c \in \mathcal{G}_{ont}}{\text{Top-k}}\, \text{sim}\big(\phi(q), \phi(c)\big) \quad (18)$$

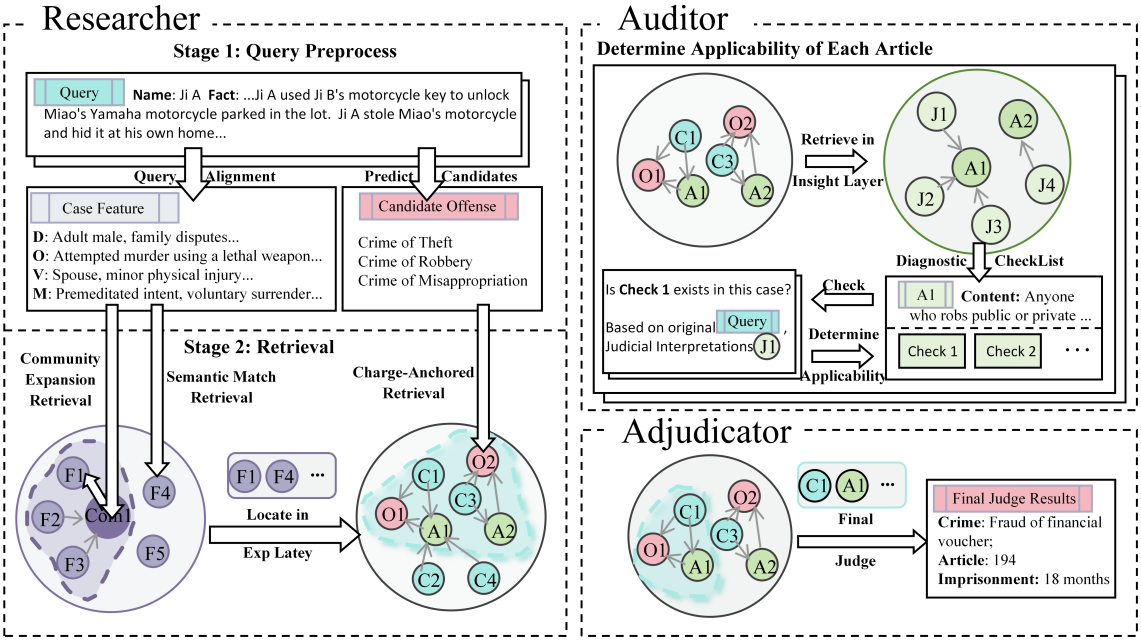

Figure 8: **Case study** of LegalGraphRAG compared to baselines. Their execution processes are compared in detail.

(ii) **Community Expansion Retrieval**: To capture broader structural context, we employ a community-guided strategy. We identify the single most relevant thematic community $\mathcal{K}^*$ aligned with the query, and then retrieve the top-k similar cases restricted within this community:

$$\begin{aligned} \mathcal{K}^* &= \underset{\mathcal{K}\in\mathcal{G}_{ont}}{\operatorname{argmax}} \operatorname{sim}\big(\phi(q), \phi(\mathcal{K})\big) \\ \mathcal{R}_{\text{com}}(q) &= \underset{c\in\mathcal{K}^*}{\text{Top-k}} \operatorname{sim}\big(\phi(q), \phi(c)\big) \end{aligned} \tag{19}$$

(iii) **Charge-Anchored Retrieval**: Finally, we anchor the legal basis by retrieving cases linked to inferred charges. Here, $\mathcal{O}(q)$ denotes the set of predicted charges and $\mathcal{N}_{\mathcal{G}_{fac}}(o)$ represents the neighboring cases connected to charge o in the Fact Graph:

$$\mathcal{R}_{\text{chg}}(q) = \bigcup_{o\in\mathcal{O}(q)} \mathcal{N}_{\mathcal{G}_{fac}}(o) \tag{20}$$

These three retrieval strategies organize the candidate evidence set $\mathcal{S}_{cand}$.

Auditor Agent: Operating on the candidate evidence set $\mathcal{S}_{cand}$, the Auditor validates the applicability of each retrieved article v_a through a rigorous verify-and-prune mechanism. This process proceeds in three specific steps:

(i) **Diagnostic Retrieval**: For each article v_a, the agent retrieves its specific Diagnostic Checklist $\mathcal{D}(v_a) = \{d_1, \ldots, d_{|C|}\}$ and relevant Judicial Interpretations $\mathcal{J}$ from the Rule Graph $\mathcal{G}_{rul}$.

(ii) **Item-wise Verification**: The agent executes a verification loop for each diagnostic item $d_k \in \mathcal{D}$. It evaluates whether the raw case facts q satisfy the specific legal condition d_k, supporting the judgment with the interpretive context $\mathcal{J}$. This produces a set of boolean verification results $V_{results} = \{r_k\}$, where $r_k \leftarrow \text{CheckCondition}(q, d_k, v_a, \mathcal{J})$.

(iii) **Decision and Pruning**: Finally, the Auditor synthesizes the verification results $V_{results}$ to determine the overall applicability of the article. If the article fails to meet the necessary criteria ($IsApplicable$ is False), the Auditor executes a pruning operation:

$$\mathcal{S}_{verified} \leftarrow \text{Prune}(\mathcal{S}_{verified}, v_a) \tag{21}$$

This step removes the inapplicable article node v_a along with its dependent case precedents and charge nodes, ensuring that the final subgraph $\mathcal{S}_{verified}$ contains only logically valid and applicable evidence.

Adjudicator Agent: The Adjudicator synthesizes the verified subgraph $\mathcal{S}_{verified}$ to render the final judgment. Specifically, it organizes the valid nodes extracted from $\mathcal{S}_{verified}$ into sets of confirmed articles (V_A^f), case precedents (V_C^f), and charge information (V_O^f). By integrating these evidence components with the original query q, it generates a response with explicit citations. This process is

Algorithm 1 Evidence-based Legal Reasoning

Require: Raw case query q; Ontology Graph $\mathcal{G}_{ont}$; Fact Graph $\mathcal{G}_{fac}$; Rule Graph $\mathcal{G}_{rul}$.
Ensure: Final Judgment $\mathcal{J}$ with citations.

Stage 1: Researcher Agent (Multi-Strategy Retrieval)
$\phi(q) \leftarrow \text{OntologyAlign}(q, \mathcal{G}_{ont})$ ▷ Align query to ontology features
Parallel Evidence Retrieval Strategies:
$\mathcal{S}_{cand} \leftarrow \mathcal{R}_{\text{sem}} \cup \mathcal{R}_{\text{com}} \cup \mathcal{R}_{\text{chg}}$ ▷ Union of candidate evidence
Stage 2: Auditor Agent (Verification & Pruning)
$\mathcal{S}_{verified} \leftarrow \mathcal{S}_{cand}$
for each article node $v_a \in \mathcal{S}_{cand}$ **do**
  $\mathcal{D} \leftarrow \text{RetrieveChecklist}(v_a, \mathcal{G}_{rul})$ ▷ Get checklist $\mathcal{D}(v_a) = \{d_1, \ldots, d_{|C|}\}$
  $\mathcal{J} \leftarrow \text{RetrieveInterpretations}(v_a, \mathcal{G}_{rul})$ ▷ Get Judicial Interpretations
  $V_{results} \leftarrow \emptyset$
  for each diagnostic item $d_k \in \mathcal{D}$ **do** ▷ Item-wise verification loop
    $r_k \leftarrow \text{CheckCondition}(q, d_k, v_a, \mathcal{J})$ ▷ Verify if fact q satisfies condition d_k
    $V_{results} \leftarrow V_{results} \cup \{r_k\}$
  end for
  $IsApplicable \leftarrow \text{Decide}(V_{results})$ ▷ Final determination for node v_a
  if not $IsApplicable$ **then**
    $\mathcal{S}_{verified} \leftarrow \text{Prune}(\mathcal{S}_{verified}, v_a)$ ▷ Remove article and linked nodes
  end if
end for
$\mathcal{G}^f_{sub} \leftarrow \{V^f_A, V^f_C, V^f_O\} \leftarrow \text{Organize}(\mathcal{S}_{verified})$ ▷ Structure the verified subgraph
Stage 3: Adjudicator Agent (Synthesis)
$\mathcal{Y} \leftarrow \text{Adjudicator}(q \oplus V^f_A \oplus V^f_C \oplus V^f_O)$ ▷ Synthesize judgment with citations
return $\mathcal{Y}$

formulated as:

$$\mathcal{Y} = Adjudicator(q \oplus V^f_A \oplus V^f_C \oplus V^f_O) \quad (22)$$

The output $\mathcal{Y}$ ensures that every conclusion is directly traceable to specific nodes in the knowledge graph, enforcing transparency and evidence-based reasoning.

C Additional Experiments

C.1 Extensions to the Main Experiment (Q4)

In this section, we conduct a series of extended experiments to verify the universality of our framework across different model architectures and its robustness in specific, high-difficulty legal sub-tasks.

Obs.7. Universality across Closed-Source Backbones. To verify the universality of our framework, we extended the evaluation to advanced closed-source large language models, specifically DeepSeek-V3 and GPT-4o-mini. As shown in Table 4, LegalGraphRAG consistently outperforms all baselines across both CAIL and CMDL datasets, regardless of the backbone model employed. Notably, even with the lighter GPT-4o-mini on the CMDL dataset, our method achieves a remarkable performance gain (e.g., significantly exceeding the strong baseline RAPTOR in Accuracy), while maintaining its lead with the more powerful DeepSeek-V3. This demonstrates that LegalGraphRAG's structured reasoning capabilities effectively complement the generation power of various state-of-the-art LLMs, enhancing their precision in complex legal application scenarios independent of the underlying model architecture.

Obs.8. Exactness in Law Article Prediction. Table 5 illustrates the model's capability in Law Article Prediction, a task demanding precise statutory grounding rather than generative flexibility. LegalGraphRAG achieves a superior overall accuracy of 47.9%, establishing a substantial lead over both the strongest RAG baseline, HippoRAG2 (39.8%), and the domain-specific state-of-the-art, ADAPT (41.3%). Remarkably, our 8B-parameter framework even surpasses the massive DeepSeek-V3.1 (44.9%), highlighting that our structured, evidence-based retrieval mechanism is more effective at pinpointing legal provisions than simply scaling model parameters or employing semantic retrieval.

Model	Size	CAIL									
		Public Safety		Economic		Social Order		Person Rights		All	Δ
		ACC	F1	ACC	F1	ACC	F1	ACC	F1		
GPT-4o-mini											
Naive RAG	∼8B	27.5	37.6	18.8	33.6	18.0	28.8	22.1	39.2	22.2	↑ 18.7
G-Retriever	∼8B	17.5	24.8	20.3	32.8	20.5	31.0	24.1	31.6	21.4	↑ 19.5
LightRAG	∼8B	25.4	37.4	21.3	36.9	21.7	38.7	23.4	42.8	23.1	↑ 17.8
RAPTOR (Sarthi et al., 2024)	∼8B	33.1	49.0	29.3	44.2	25.9	39.9	28.3	43.1	30.5	↑ 10.4
HippoRAG2 (Gutiérrez et al., 2025)	∼8B	33.1	50.1	28.1	46.1	21.7	43.6	37.2	54.2	31.9	↑ 9.0
LegalGraphRAG (Ours)	∼8B	**39.6**	**54.8**	**36.3**	**52.9**	**37.3**	**51.2**	**42.1**	**62.4**	**40.9**	–
DeepSeek-V3											
Naive RAG	∼200B	38.0	54.0	32.3	49.9	33.4	47.3	40.7	53.4	37.8	↑ 12.1
G-Retriever	∼200B	36.5	54.6	35.1	49.8	36.2	47.2	39.5	48.3	37.2	↑ 12.7
LightRAG	∼200B	36.6	48.5	26.3	50.2	33.5	46.3	39.7	53.1	45.4	↑ 4.5
RAPTOR (Sarthi et al., 2024)	∼200B	42.2	56.3	37.8	53.0	39.2	50.7	45.5	52.3	44.4	↑ 5.5
HippoRAG2 (Gutiérrez et al., 2025)	∼200B	41.5	49.1	33.1	46.8	34.3	47.0	38.6	46.4	41.2	↑ 8.7
LegalGraphRAG (Ours)	∼200B	**44.4**	**58.8**	**41.9**	**57.8**	**41.9**	**56.8**	**46.2**	**65.1**	**49.9**	–

Model	Size	CMDL									
		Public Safety		Economic		Social Order		Person Rights		All	Δ
		ACC	F1	ACC	F1	ACC	F1	ACC	F1		
GPT-4o-mini											
Naive RAG	∼8B	38.7	50.8	35.6	44.0	30.9	44.2	33.3	43.0	32.9	↑ 17.1
G-Retriever	∼8B	24.6	38.4	29.7	38.8	30.4	40.0	41.1	51.6	28.3	↑ 21.7
LightRAG	∼8B	36.2	43.1	37.5	48.9	46.9	55.1	34.6	50.8	34.2	↑ 15.8
RAPTOR (Sarthi et al., 2024)	∼8B	42.0	49.1	48.2	58.0	57.3	60.9	46.0	59.5	48.7	↑ 11.3
HippoRAG2 (Gutiérrez et al., 2025)	∼8B	45.0	52.5	42.9	60.8	45.3	64.6	59.4	75.0	46.0	↑ 14.0
LegalGraphRAG (Ours)	∼8B	**46.5**	**57.8**	**63.7**	**67.9**	**58.0**	**66.2**	**67.2**	**75.4**	**60.0**	–
DeepSeek-V3											
Naive RAG	∼200B	52.0	65.3	66.1	71.6	69.8	70.7	68.8	80.5	62.9	↑ 12.8
G-Retriever	∼200B	46.2	64.9	58.6	69.6	56.1	67.4	69.7	78.9	55.2	↑ 23.5
LightRAG	∼200B	47.7	58.7	46.5	67.6	47.6	53.2	52.3	64.2	52.4	↑ 26.3
RAPTOR (Sarthi et al., 2024)	∼200B	53.4	64.7	68.2	73.0	66.4	75.4	60.3	71.1	56.4	↑ 12.3
HippoRAG2 (Gutiérrez et al., 2025)	∼200B	62.1	63.5	62.4	65.3	**75.9**	78.6	76.9	78.2	74.0	↑ 4.7
LegalGraphRAG (Ours)	∼200B	**66.7**	**69.9**	**76.0**	**79.3**	72.9	**80.4**	**79.7**	**85.5**	**78.7**	–

Table 4: **Performance comparison on Large Closed-Source Models**. We compared LegalGraphRAG and other baselines utilizing larger closed-source LLMs as backbones.

Obs.9. Precision in Term of Penalty Prediction. Table 6 presents the results on the challenging term of penalty prediction task, which requires fine-grained quantitative reasoning rather than simple classification. LegalGraphRAG demonstrates a significant advantage in minimizing prediction error, consistently achieving the lowest Mean Absolute Error (MAE) across most subdomains compared to other RAG-based methods. For instance, in the Public Safety category, our model achieves an MAE of 20.9, outperforming RAPTOR (21.7) and HippoRAG2 (23.0). This indicates that while exact term matching remains difficult for all models, LegalGraphRAG's evidence-based retrieval strategy effectively locates relevant sentencing guidelines and comparable precedents, thereby constraining the generation to a more precise and legally grounded time range.

C.2 Hyper-parameter Sensitivity (Q5)

To evaluate system stability, we investigated the sensitivity of the Researcher Agent to the retrieval parameter k, which governs the number of semantic concepts retrieved from the ontology graph $\mathcal{G}_{ont}$. We varied k over the set $\{3, 4, 5, 6\}$. The upper bound is restricted to 6, as empirical evidence suggests that exceeding this threshold introduces excessive context noise, which overwhelms the model's effective window and degrades reasoning.

Obs.10. Robustness to Retrieval Hyperparameter Variations. As illustrated in Figure 9, LegalGraphRAG exhibits strong robustness to variations

Model	Size	CAIL									
		Public Safety		Economic		Social Order		Person Rights		All	Δ
		ACC	F1	ACC	F1	ACC	F1	ACC	F1		
Open-Source Models											
Qwen-2.5-7B-Instruct	7B-Inst	28.7	53.8	24.1	48.2	26.6	53.5	36.0	56.2	30.1	↑ 17.8
Qwen-3-8B	8B-Inst	23.9	58.5	27.6	51.2	36.6	62.2	46.2	66.7	35.9	↑ 12.0
Internlm3-8b-instruct	8B-Inst	27.6	59.9	26.4	52.8	30.8	59.0	33.6	57.6	29.9	↑ 18.0
Glm-4-9b-chat	9B-Inst	23.9	60.1	25.6	51.0	34.7	59.4	35.1	58.5	30.8	↑ 17.1
Closed-Source Models											
GPT-4o-mini (Achiam et al., 2023)	∼8B	24.7	57.2	25.7	42.4	24.6	51.8	34.7	55.0	30.9	↑ 17.0
DeepSeek-V3.1 (Liu et al., 2024)	∼200B	42.3	63.2	37.1	**61.3**	44.5	68.8	51.9	67.3	44.9	↑ 3.0
Legal Specific Methods											
DISC-LawLLM-7B (Yue et al., 2024)	7B-Inst	36.5	55.9	29.9	48.8	41.5	61.2	39.3	57.8	38.2	↑ 9.7
ADAPT (Deng et al., 2024b)	7B-Inst	40.1	50.8	31.6	42.4	39.6	54.9	41.0	50.5	41.3	↑ 6.6
Legal Δ (Dai et al., 2025)	7B-Inst	33.0	54.9	27.8	51.0	34.6	59.4	44.5	60.6	37.9	↑ 10.0
RAG Based Methods											
Naive RAG	8B-Inst	30.4	45.7	29.2	48.6	37.5	54.1	36.6	51.1	34.8	↑ 13.1
RAPTOR (Sarthi et al., 2024)	8B-Inst	36.4	59.2	33.4	53.9	38.9	64.1	41.0	61.7	37.2	↑ 10.7
HippoRAG2 (Gutiérrez et al., 2025)	8B-Inst	35.2	60.3	33.1	53.7	41.7	67.0	42.6	63.8	39.8	↑ 8.1
LegalGraphRAG (Ours)	8B-Inst	**43.0**	**64.9**	**37.8**	61.0	**44.6**	**69.4**	**54.5**	**70.6**	**47.9**	–

Table 5: **Extended experiments on Article Prediction**. We evaluated the performance of our model and baselines on the specific sub-task of law article prediction. We visualize the gains of LegalGraphRAG to the each baseline in the Δ columns.

Model	Size	CAIL									
		Public Safety		Economic		Social Order		Person Rights		All	Δ
		ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE		
Open-Source Models											
Qwen-2.5-7B-Instruct	7B-Inst	13.0	23.7	8.1	33.3	6.0	30.3	9.7	29.4	29.5	↑ 8.4
Qwen-3-8B	8B-Inst	8.3	32.6	11.3	31.6	17.9	29.7	11.4	26.4	27.5	↑ 7.4
Internlm3-8b-instruct	8B-Inst	7.1	35.2	5.9	37.2	6.0	32.5	7.2	37.3	33.7	↑ 13.6
Glm-4-9b-chat	9B-Inst	3.6	36.0	3.2	32.9	1.5	35.1	6.8	38.6	33.1	↑ 13.0
Closed-Source Models											
GPT-4o-mini (Achiam et al., 2023)	∼8B	6.9	38.2	7.3	34.2	8.3	31.7	8.0	34.6	33.6	↑ 13.5
DeepSeek-V3.1 (Liu et al., 2024)	∼200B	7.1	31.2	8.1	31.5	10.4	25.9	8.4	29.1	29.1	↑ 8.6
Legal Specific Methods											
DISC-LawLLM-7B (Yue et al., 2024)	7B-Inst	15.5	24.5	5.0	33.2	3.0	37.9	8.4	34.5	31.6	↑ 11.5
ADAPT (Deng et al., 2024b)	7B-Inst	8.3	21.8	10.9	21.9	3.0	24.3	9.7	22.3	20.4	↑ 0.3
Legal Δ (Dai et al., 2025)	7B-Inst	11.9	25.0	9.0	29.0	9.0	27.5	8.9	27.1	26.3	↑ 6.2
RAG Based Methods											
Naive RAG	8B-Inst	10.5	29.8	11.4	30.8	18.1	21.7	12.6	24.3	26.5	↑ 4.4
RAPTOR (Sarthi et al., 2024)	8B-Inst	12.0	21.7	11.7	28.3	15.8	26.1	16.2	**21.8**	24.3	↑ 4.2
HippoRAG2 (Gutiérrez et al., 2025)	8B-Inst	13.1	23.0	12.7	25.8	17.9	23.8	13.5	23.4	23.8	↑ 3.7
LegalGraphRAG (Ours)	8B-Inst	**14.0**	**20.9**	**13.7**	**22.1**	**19.4**	**23.6**	**17.1**	22.7	**20.1**	–

Table 6: **Extended experiments on Term of Penalty Prediction**. We assessed the accuracy and error rates of imprisonment term predictions compared to baselines. We visualize the gains of LegalGraphRAG to the each baseline in the Δ columns.

Method	Indexing Time (s)	Token Consumption ($< 10^6$)		Avg Cost	
		Prompt	Completion	Time	Token
RAPTOR	13696.90	5.64	0.72	5.86	3589
HippoRAG2	4581.60	10.58	2.79	11.2	5199
LegalGraphRAG (Ours)	3687.49	3.97	0.78	46.1	10664

Table 7: Comparison of Computational Efficiency: Offline Indexing vs. Online Inference. We report the total time and token usage for graph construction (Indexing) and the average cost per query (Online).

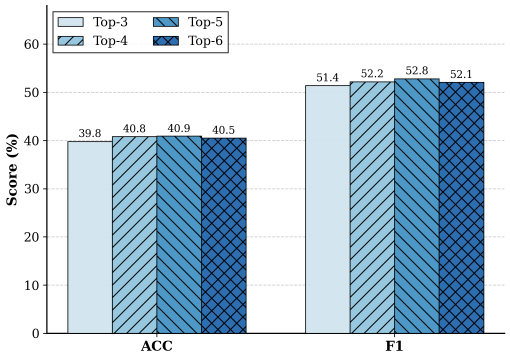

Figure 9: Impact of the retrieval parameter k on charge prediction performance (CAIL dataset). The backbone model is Qwen3-8B.

in k. Although performance peaks at $k = 5$ (achieving 40.9 Accuracy and 52.8 F1), the variance across the tested range is marginal. This indicates that the Researcher Agent reliably captures essential semantic information without being hypersensitive to exact thresholding, provided k remains within a reasonable bound.

C.3 Latency and Token Cost (Q6)

In this section, we shift our focus to the practical efficiency and computational overhead of the compared frameworks. Beyond accuracy, the latency and token consumption are crucial factors for real-world deployment. We provide a detailed breakdown and comparison of the computational costs for RAPTOR, HippoRAG2, and LegalGraphRAG across two primary phases: (i) the offline graph construction stage, reporting both the time and token cost required to build the knowledge base; and (ii) the online query-answering stage, reporting the time and token cost incurred during the retrieval and reasoning process for a given query.

Obs.11. Trade-off between Efficiency and Interpretability. Table 7 presents the computational efficiency. LegalGraphRAG demonstrates superior offline efficiency with the lowest indexing time (3687.49s) and token consumption. However, during the online phase, it incurs higher latency (46.1s) and token usage. This overhead is a necessary trade-off for evidence-based reasoning. Unlike baseline GraphRAG approaches that often operate as opaque "black boxes", our method explicitly constructs credible reasoning chains to support its judgments. While generating such transparent evidence consumes more resources, it is indispensable for ensuring the trustworthiness and interpretability required in legal domains.

D Implementation Details

D.1 Benchmark Dataset

We evaluate LegalGraphRAG on two benchmark datasets, including CAIL2018(Xiao et al., 2018) and CMDL(Huang et al., 2024).

CAIL2018(Xiao et al., 2018): A large-scale Chinese legal dataset designed for the task of Legal Judgment Prediction (LJP), where models predict court outcomes based on factual case descriptions. It comprises over 2.6 million criminal cases published by the Supreme People's Court of China, making it the largest publicly available dataset of its kind. Each case includes a detailed fact description along with structured judgment annotations, namely applicable law articles (183 categories), charges (202 categories), and prison terms. The dataset was created to address the lack of high-quality, large-scale resources in legal AI and to provide a realistic benchmark that reflects the complexity and imbalance inherent in real judicial data, where frequent charges dominate the case distribution. CAIL2018 has since become a foundational resource for evaluating and advancing automated legal judgment prediction systems.

CMDL(Huang et al., 2024): A large-scale, real-world Chinese Multi-Defendant Legal Judgment Prediction dataset designed to address the under-explored challenge of predicting judicial outcomes in cases involving multiple defendants. It comprises 393,945 criminal cases with approximately 1.2 million defendants, covering 321 distinct charges and 275 legal articles. Notably, CMDL in-

Dataset	CAIL	CMDL
# Case Num	568	572
# Charges	168	239
# Average criminal per case	1.25	2.40
# Average defendant per case	1.72	1.14
# Average length per case	654.79	517.13

Table 8: Basic statistics of the test datasets.

troduces case-level evaluation metrics that account for case complexity and varying numbers of defendants, offering a more holistic assessment of model performance in multi-defendant scenarios. For experimental feasibility, the subset **CMDL-small** is often utilized, as it preserves the data distribution while significantly reducing computational costs, making it suitable for preliminary benchmarking and model validation in resource-constrained research settings.

Dataset Construction Due to considerations regarding the generation speed and token cost of the GraphRAG method, we constructed focused subsets from both the CAIL2018 and CMDL datasets using a uniform procedure aimed at controlling input length, balancing charge distribution, and elevating task complexity. The construction first filtered cases to retain only those with factual descriptions under 1,024 characters. To ensure broader coverage of under-represented charges and increase predictive difficulty, the sampling prioritized defendants whose charges included low-frequency offenses and deliberately retained a higher proportion of multi-charge cases. Consequently, the resulting subsets feature a more balanced charge distribution with elevated presence of rare charges and greater average case complexity compared to the original datasets. While this design provides a more challenging testbed for evaluating model performance on complex and low-frequency legal scenarios, it may also lead to lower reported performance for some methods relative to their results on the original, more naturalistic data distribution. Table 8 presents detailed statistics of the subsets.

D.2 Corpus

In this section, we provide detailed descriptions of corpus used in our experiments. To ensure comprehensive coverage of criminal statutes and charges, we construct this knowledge base by aggregating a subset of cases from multiple authoritative legal datasets: JuDGE(Su et al., 2025), CAIL2018(Xiao et al., 2018), CMDL(Huang et al., 2024), and

Dataset	Case
# Num	14049
# Charges*	818
# Average defendant per case	3.39
# Average length per case	399.73

Table 9: Basic statistics of the cases in RAG dataset. *The large number of "Charges" is due to inconsistencies in the descriptions of crimes across different datasets.

Dataset	Article	Judicial interpretations
# Num	452	656
# Average length	128.54	243.95

Table 10: Basic statistics of the legal knowledges in corpus.

LeCaRDv2(Li et al., 2024). The construction follows a procedure similar to that used for the experimental subsets: we first filter cases by fact description length (under 1,024 characters) and apply sampling designed to balance the representation of different charges, thereby creating a broad and diverse collection of historical precedents and factual patterns. Furthermore, to ground the system in authoritative legal provisions, we incorporate the full text of the "Criminal Law of the People's Republic of China" along with its relevant judicial interpretations as a core statutory knowledge library. The combination of this curated historical case library and the official legal provisions library forms the complete corpus, enabling models to retrieve both experiential precedents and statutory knowledge during reasoning.Crucially, we have carefully verified that all cases in the corpus are distinct from those in the test subsets, ensuring no data leakage between the knowledge base and the evaluation benchmarks. The size and composition statistics of the final corpus are detailed in Table 9 & 10.

D.3 Evaluation Metrics

To comprehensively evaluate the Legal Judgment Prediction (LJP) tasks, we employ specific metrics for different sub-tasks: Charge and Article Prediction are evaluated using Accuracy (ACC) and Micro-F1, Term of Penalty Prediction is assessed using ACC and Mean Absolute Error (MAE), and the Retrieval Quality of our RAG system is measured by Retrieval Effectiveness and Error Rate.

Accuracy (ACC) measures the exact match ratio. For classification tasks, it requires the predicted label set O_i to be identical to the ground truth O'_i.

For Term of Penalty, it measures the exact match of the predicted term. It is calculated as:

$$\text{ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(O_i = O_i')$$

Where $\mathbb{I}(\cdot)$ is the indicator function.

Micro-F1 is used for multi-label classification to account for class imbalance. It is the harmonic mean of micro-averaged precision (P_{micro}) and recall (R_{micro}):

$$\text{Micro-F1} = \frac{2 \cdot P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}$$

Mean Absolute Error (MAE) reflects the deviation in the predicted term of penalty. Let T_i and T_i' denote the predicted and ground-truth prison terms (in months), respectively. MAE is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|T_i - T_i'|$$

Retrieval Effectiveness measures how well the retrieved content aligns with the question's intent. Higher values indicate more focused and pertinent information. It is defined as:

$$\text{Retrieval Effectiveness} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\text{R}(c, Q, \mathcal{E})$$

where $\mathcal{C}$ denotes the set of retrieved contexts, Q represents the question, $\mathcal{E}$ denotes the set of evidence, and the operator $\text{R}(\cdot)$ determines the relevance of a context c.

Error Rate quantifies the incompleteness of the retrieval process. Instead of measuring recall directly, we assess the proportion of reference claims not supported by the retrieved context:

$$\text{Error Rate} = 1 - \left(\frac{1}{|\mathcal{R}|}\sum_{c\in\mathcal{R}}\mathbb{I}(\text{S}(c, \mathcal{C}))\right) \quad (23)$$

where $\mathcal{R}$ is the set of reference claims, $\text{S}(\cdot)$ determines whether a claim c is supported by the retrieved context $\mathcal{C}$, and $\mathbb{I}(\cdot)$ is the indicator function. A lower Error Rate indicates a more comprehensive evidence collection.

D.4 Baseline Details

In this section, we provide detailed descriptions of each baseline used in our comparison.

Naive RAG uses the standard RAG paradigm: a retriever model first retrieves relevant context from the corpus based on the given question, and then the question is concatenated with the retrieved context to form a query for the generation model to produce the final answer.

G-retriever(He et al., 2024a) introduces a retrieval augmented generation framework for textual graphs by formulating subgraph retrieval as a Prize-Collecting Steiner Tree optimization problem, enabling conversational question answering across diverse domains like scene understanding and knowledge graphs while mitigating LLM hallucinations and scaling to large graph sizes.

LightRAG(Guo et al., 2024) introduces a graph-enhanced retrieval-augmented generation framework that integrates entity-relationship graphs into text indexing, combining low-level precise entity retrieval with high-level thematic discovery for efficient and adaptive knowledge integration.

HippoRAG2(Gutiérrez et al., 2025) builds on HippoRAG's Personalized PageRank framework by integrating dense-sparse coding for passages and phrases in the knowledge graph, enabling deeper contextualization and recognition memory for triple filtering. Enhances online retrieval with query-to-triple matching and optimized seed node weighting, outperforming standard RAG across factual, sense-making, and associative memory tasks.

RAPTOR(Sarthi et al., 2024) constructs a hierarchical tree by recursively clustering and summarizing embedded text chunks, enabling retrieval of information at multiple levels of abstraction to improve performance on long-document question-answering tasks.

RAG Configuration

```
{
  embedding_model: bge-m3,
  retrieval_topk: 5,
  chunk_token_size: 1000,
  chunk_overlap_token_size: 200
}
```

G-retriever Configuration

```
{
  embedding_model: bge-m3,
  retrieval_topk: 3,
  chunk_token_size: 1200,
  chunk_overlap_token_size: 100,
  entities_max_tokens: 2000,
  relationships_max_tokens: 2000
}
```

LightRAG Configuration

```
{
  embedding_model: bge-m3,
  query_type: hybrid,
  chunk_token_size: 1200,
  retrieval_topk: 20,
  chunk_overlap_token_size: 100,
  max_token_text_unit: 2000,
  max_token_global_context: 2000,
  max_token_local_context: 2000
}
```

HippoRAG2 Configuration

```
{
  embedding_model: bge-m3,
  retrieval_top_k: 5,
  linking_top_k: 5,
  max_qa_steps: 3,
  qa_top_k: 5,
  graph_type: facts_and_sim_passage
  _node_unidirectional
}
```

RAPTOR Configuration

```
{
  embedding_model: bge-m3,
  chunk_token_size: 1200,
  chunk_overlap_token_size: 100,
  num_layers: 5,
  max_length_in_cluster: 3500,
  threshold: 0.1,
  cluster_metric: cosine,
  threshold_cluster_num: 5000
}
```

Disc-LawLLM(Yue et al., 2024) is a retrieval-augmented large language model fine-tuned on Chinese judicial datasets using legal syllogism prompting to provide reasoning-capable legal services, including consultation, judgment prediction, and examination assistance. In this work, we use the officially open-sourced LawLLM-7B, which is fine-tuned from Qwen2.5-Instruct-7B.

Legal Δ(Dai et al., 2025) employs a reinforcement learning framework that enhances legal reasoning in LLMs by maximizing chain-of-thought guided information gain through dual-mode inputs and differential Q-value analysis. In this work, we use the officially open-sourced model, which is fine-tuned from Qwen2.5-Instruct-7B.

ADAPT(Deng et al., 2024b) is a discriminative reasoning framework for LLMs in legal judgment prediction that emulates human judicial processes by asking to decompose case facts into key elements, discriminating among candidate charges for alignment, and predicting final judgments, further improved via multi-task fine-tuning with synthetic trajectories. In this work, we use the officially open-sourced model, which is fine-tuned from Qwen2-7B.

D.5 LegalGraphRAG Setup

In the experimental setup of LegalGraphRAG, hyperparameters are configured to optimize retrieval precision. During graph construction, the Ontology Graph utilizes k-nearest neighbors (kNN) to select the top-3 case feature nodes based on cosine similarity for direct semantic matching.

For the Researcher agent, the evidence-based retrieval strategy operates with specific thresholds: the retrieval parameter is set to $k = 5$. This parameter governs both the number of top-ranked semantic concepts retrieved from the ontology graph and the scope of community expansion, a value selected based on the sensitivity analysis to balance context coverage and noise control.

E Extended Case Examples

In this section, we will walk through several cases to detail the retrieval and reasoning pipeline of LegalGraphRAG.

F Prompt Set

To facilitate reproducibility and provide transparency into the agent behaviors, we present the specific instruction sets designed for the **Researcher**, **Auditor**, and **Adjudicator** agents. These prompts orchestrate the multi-stage reasoning process described in the main text.

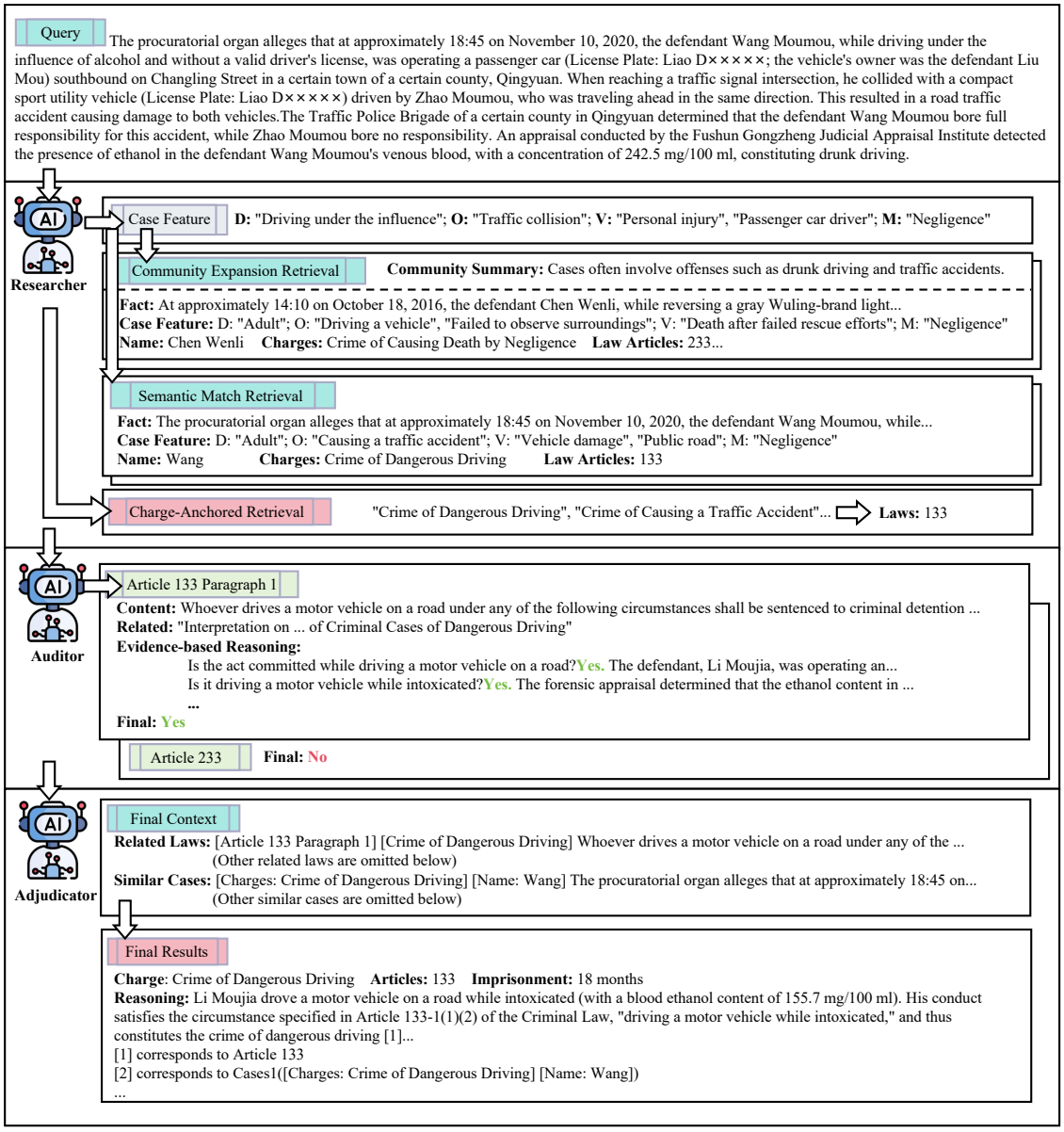

Figure 10: Qualitative analysis of a representative case regarding the crime of **Dangerous Driving**. The visualization highlights the retrieval of evidence related to specific statutory conditions.

G Related Works

G.1 Legal Large Language Models

The rapid evolution of LLMs has catalyzed the development of domain-specific models tailored for the legal sphere. For Chinese law, **ChatLaw** (Cui et al., 2023), **DISC-LawLLM** (Yue et al., 2023), and **InternLM2Law** (Fei et al., 2025) leverage extensive legal corpora, including judicial interpretations and statutes, to handle diverse legal tasks. Other notable models like **LawGPT** (Zhou et al., 2024) and **Fuzi-Mingcha** (Wu et al., 2023a) integrate unsupervised legal texts with supervised fine-tuning to enhance domain understanding. Beyond Chinese, **SaulLM** (Colombo et al., 2024) focuses on English legal texts based on the Mixtral architecture, while **LawLLM** (Shu et al., 2024) addresses US legal tasks such as similar case retrieval. Additionally, specialized models like **InLegalLLaMA** (Ghosh et al., 2024) target Indian and French legal domains respectively. These models provide crucial baselines for downstream tasks but often lack the specific reasoning architecture required for complex judgment prediction.

G.2 Legal judgment prediction

Legal judgment prediction (LJP) has experienced significant development and become an increas-

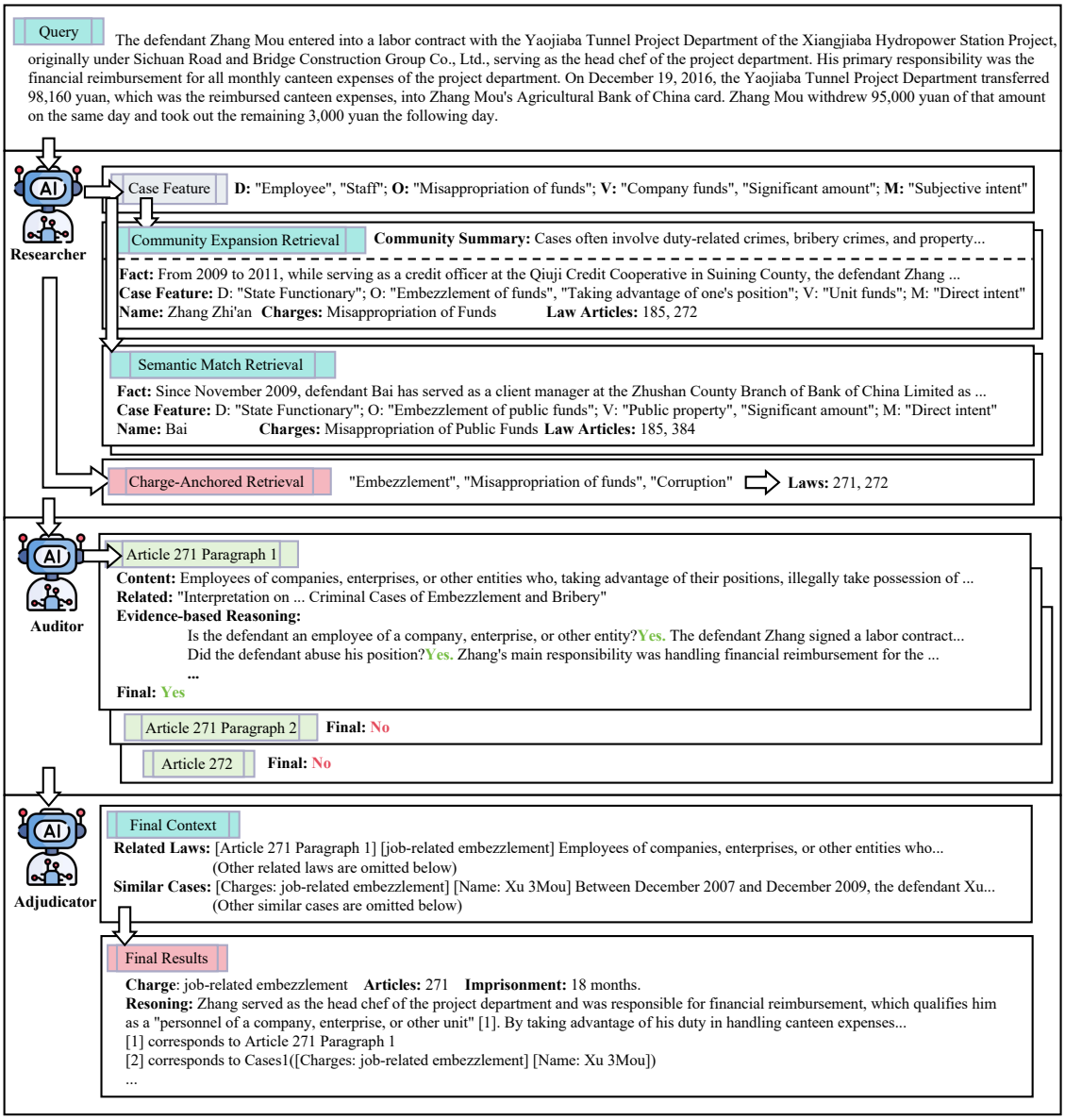

Figure 11: Qualitative analysis of a representative case regarding the crime of **Occupational Embezzlement**. The example demonstrates the model's reasoning in identifying the abuse of professional position.

ingly crucial NLP task. Earlier research (Segal, 1984) relied on artificially designed features, and traditional machine learning methods (Sulea et al., 2017) were applied to predict legal judgments. Recent advances in deep learning (Xu et al., 2020; Han and Zhicheng, 2023) have motivated researchers to leverage neural networks for automated text representation learning. Recently, LLMs have further promoted the progress of LJP (Deng et al., 2024a). Several studies (Wu et al., 2023b; Peng and Chen, 2024) employ Retrieval-Augmented Generation (RAG) to enhance LLMs by incorporating external legal knowledge. To refine decision-making, recent works have introduced structured reasoning frameworks that systematically decompose case facts to distinguish confusing charges (Jiang and Yang, 2023; Deng et al., 2024b; Wang et al., 2024). Furthermore, multi-agent simulation frameworks have been explored to improve performance by simulating court debates and analyzing cases from diverse perspectives (He et al., 2024b). However, existing LLM-based methods still struggle to utilize comprehensive legal knowledge (Fei et al., 2024) effectively. In this context, we make full use of external knowledge and precedents within a unified framework.

G.3 Reasoning skills in legal domain

Recent work has improved LLMs' reasoning through better prompting techniques (Sahoo et al.,

Criminal Case Keyword Extraction

Task Definitions: Extract legal keywords from a criminal case description and classify them into four categories: Defendant Attributes, Criminal Behaviors, Victim Characteristics, and Subjective Mental States. The output must be a strictly valid JSON object without additional text.

Keyword Definitions:

- **Defendant Attributes:** Legal traits (e.g., age group, criminal history, occupation). Avoid specific names or numbers.
- **Criminal Behaviors:** Legal types of acts and significant methods. Exclude specific time/location details.
- **Victim Characteristics:** Nature of the property or location. Generalize specific amounts (e.g., "large amount").
- **Subjective Mental States:** Legal descriptions of intent and remorse.

Output Example: {
"Defendant_Attribute": ["Adult", "Prior Criminal Record"],
"Criminal_Behaviors": ["Theft", "Burglary"],
"Victim_Characteristics": ["Private Residence", "Large Amount"],
"Subjective_Mental_States": ["Direct Intent", "Voluntary Surrender"]
}

Charge Pre-judge(for Charge-Anchored Retrieval)

Task Definitions: Act as a criminal law expert to analyze the provided case ({case_text}).

- Output reasonably possible charges (confidence $> 30\%$) sorted by probability (descending).
- If a dominant charge exists (confidence $> 70\%$), prioritize it; if it is the only certain charge, output it exclusively.
- Exclude charges with probability $< 10\%$.

Format Definitions:

- Output strictly as a Python list: ['Charge 1', 'Charge 2', ...].
- The output must start with [.
- Return an empty list [] if no charge matches.
- **No** additional explanations or text allowed.

2024). Chain-of-thought (CoT) (Wei et al., 2022) prompting can explicitly guide LLMs to reason step by step. In the legal domain, researchers have adapted CoT to legal-specific frameworks. For instance, Yu et al. (Yu et al., 2022) demonstrated that incorporating the IRAC (Issue, Rule, Application, Conclusion) framework significantly enhances reasoning capabilities. LoT (Jiang and Yang, 2023) proposed legal syllogism reasoning to improve performance on LJP tasks, and ADAPT (Deng et al., 2024b) established a workflow enabling discriminative reasoning. Moreover, approaches like MALR (Yuan et al., 2024) utilize parameter-free learning to decompose complex legal tasks, while CaseGPT (Yang, 2024) combines LLMs with RAG to support semi-structured reasoning and legal argumentation (Westermann, 2024). Additionally, GLARE (Yang et al., 2025b) leverages an agentic framework and web data for legal reasoning. However, these approaches primarily rely on intrinsic capabilities or noisy external data, which constraints reasoning depth (Zhang, 2024; Ke et al.,

Auditor Checklist(item)

Task Definitions: Act as a legal AI assistant to assess if the case facts strictly satisfy a specific constituent element of the law.
- Analyze the law_item and case facts.
- Focus exclusively on the target element (e.g., "intent"), using related materials (if provided) for interpretation.
- determine applicability based on facts and logic.

I/O Specifications:
- **Input:** law_item, related (supplementary materials), element, case.
- **Output:** Provide reasoning first, then enclose the final result strictly within tags: <answer>true</answer> or <answer>false</answer>.

Template:

```
law: {law_item}, related: {related}
element: {element}, case: {case}
```

Auditor Checklist(final)

Task Definitions: Act as a legal analysis assistant to determine if the provided law article applies to the specific case (i.e., verify violation or crime).
- Identify all relevant constituent elements from the law text.
- Verify critical elements independently; note that the provided true_list and false_list may be incomplete.

I/O Specifications:
- **Input Variables:** case, law, true_list (proven elements), false_list (disproven elements).
- **Output:** Provide reasoning first, then enclose the final result strictly within tags: <answer>true</answer> or <answer>false</answer>.

Template:

```
case: {case}, law: {law}
true_list: {true_list}, false_list: {false_list}
```

2025). Therefore, we propose an agentic framework to dynamically acquire key legal knowledge, enhancing both breadth and depth.

G.4 RAG in legal domain

In the legal domain, recent studies have adapted RAG and graph-based solutions for specific tasks, particularly Legal Question Answering (LQA) and retrieval pipelines. For instance, recent works optimize LLM outputs by incorporating external case-based information (Wiratunga et al., 2024; Louis et al., 2024) or utilizing adapt-retrieve-revise pipelines (Wan et al., 2024) to combine continual training with evidence revision. Dedicated benchmarks such as LegalBench-RAG show that the retrieval stage remains a bottleneck (Pipitone and Alami, 2024). Works enriching retrieval with structural information (e.g., graphs of articles) demonstrate gains in tasks like statutory article retrieval (Louis et al., 2023; Hei et al., 2024; Ho et al., 2025). Recently, the SAT-Graph RAG framework (de Martim, 2025) was proposed to model the hierarchical structure of legal norms. However, its sophisticated ontology-driven approach requires heavily structured input data, limiting its applicability to less curated corpora.

The Usage of LLMs

In this paper, LLMs were used only to polish the writing and correct grammatical errors for clarity.

Charge & Sentencing (JSON)

Task Definitions: Act as a legal expert to adjudge the defendant based on candidate charges.

- **1. Final Charge Application:** For concurrence, apply the "heavier penalty" rule; for multiple acts, apply combined punishment.
- **2. Sentencing:** Predict the specific law article and a reasonable sentencing range based on facts and judicial practice.

Format Definitions:

- ```
  {
    charge_name: [Charge A, ...],
    law_article: [Art. X, ...],
    term_of_imprisonment: {
      death_penalty: boolean,
      imprisonment: integer (months),
      life_imprisonment: boolean
    }
  }
  ```

Legal Reasoning & Verdict

Task Definitions: Act as a legal consultant to analyze the case using the provided Context documents.

- **Step 1: Fact & Act Analysis:** Analyze how many independent criminal acts exist. Explicitly cite the supporting evidence from the context using [1][2]....
- **Step 2: Law Application:** Resolve any legal concurrence (e.g., Imaginative Concurrence vs. Combined Punishment). Explain why specific articles apply over others.
- **Step 3: Sentencing Prediction:** comprehensive assessment of sentencing based on statutory rules.

Output Format:

- **Structure:** Output in two clear sections: Legal Analysis (reasoning with citations) and Final Verdict (conclusion).
- **Requirement:** You **must** mark the source of your facts or laws using brackets like [1].

Template:

Context: {context_list}, Case: {case_description}

In the preparation of this manuscript, we utilized Large Language Models (LLMs) to assist with the writing process. Specifically, the model was used to refine the English text, including correcting grammatical errors and improving sentence clarity. Additionally, LLMs assisted in the initial formatting of several tables. The authors reviewed all model suggestions and retain full responsibility for the scientific accuracy and integrity of the final content.
  ```