JurisMA-CQAD: A Multi-Agent Framework and Dataset for Legal Consultation Question Answering

Anonymous ACL submission

Abstract

Legal consultation question answering (Legal CQA) presents unique challenges that differ substantially from traditional legal QA tasks, including high contextual dependency, multi-stage reasoning, and a lack of large-scale annotated datasets. To address these issues, we propose JURISMA, a modular multi-agent collaborative framework to decompose complex legal queries into interpretable subtasks. Our system integrates a structured legal element graph for semantic grounding, a Draft Agent for initial opinion generation, and a Manager Agent to dynamically coordinate refinement through auxiliary agents such as FormatCheck and LawSearch. To facilitate training and evaluation, we construct JURISCQAD, a novel dataset comprising over 43,000 real-world Chinese legal consultations, annotated with both positive and adversarial responses under expert supervision. Experiments on the LawBench benchmark demonstrate that our approach significantly outperforms state-of-the-art general and legal-domain LLMs across multiple lexical and semantic metrics.

1 Introduction

Legal consultation is a key focus in legal natural language processing (NLP) (Zhong et al., 2020). Its queries are complex, span **multiple sub-tasks**, and involve **numerous legal entities and relationships**. Such a complexity makes the accurate extraction and interpretation of involved elements critical to the quality of system responses as illustrated in Figure 1. Moreover, Legal Consultation Question Answering (Legal CQA) differs significantly from traditional legal QA in terms of its objectives, data characteristics, modeling approaches, and application scenarios. The specific differences are seen in Appendix B.

Previous studies on Legal CQA generally follow the two workflows: the first focuses on legal knowledge enhancement through continued pretraining of large language models (LLMs) on legal statutes and related materials(Huang et al., 2023). However, due to the lack of high-quality domain-specific training data, such pretraining yields limited improvements and imposes significant computational costs. The second workflow emphasizes finer-grained input structuring to guide generation, such as sentence-level statute retrieval (Ma et al., 2023; Ni et al., 2025). Yet, legal questions are inherently multi-structured and case-specific, re-

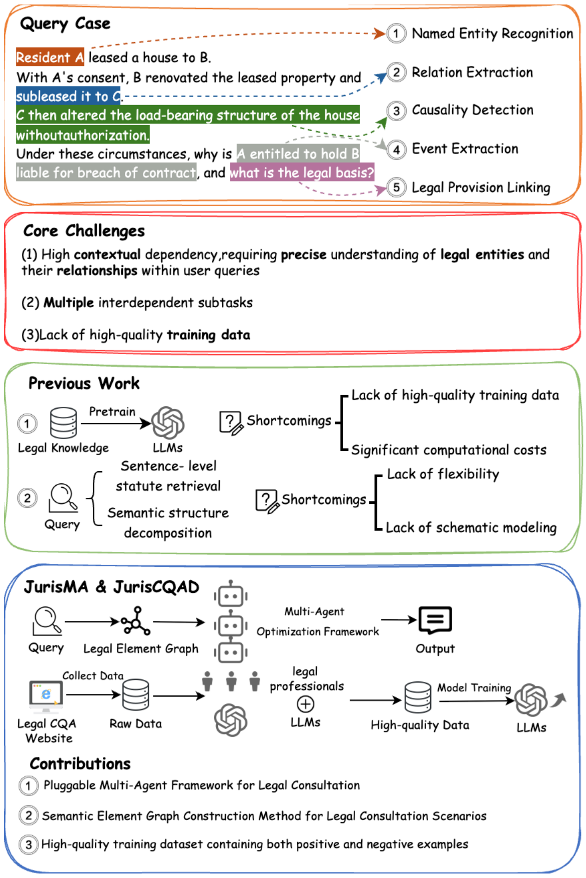

Figure 1: An illustrative example of legal consultation task decomposition, highlighting key challenges, limitations of prior approaches, and the proposed solution via our multi-agent framework JURISMA.

quiring flexible decomposition and context-aware modeling. Existing approaches often lack the capability to dynamically identify and reason over entities and events from a legal perspective.

We thus summarize the three **core challenges** presented in Legal CQA: (1) **high contextual dependency**, requiring precise understanding of legal entities and their relationships within user queries; (2) **complex task composition**, involving multiple interdependent subtasks; and (3) **lack of large-scale, high-quality training data** that reflect real consultation scenarios.

In this paper, we propose a multi-agent collaboration framework for Legal CQA to address the first two challenges. Our framework, entitled JURISMA, stands for Judicial Multi-Agent, reflecting its design philosophy of **simulating legal decision-making** through a collaborative multi-agent architecture. As illustrated in **Figure 2**, our system decomposes the legal consultation workflow into modular stages, supported by a cooperative multi-agent architecture.

We evaluate our method against both general-purpose LLMs (e.g., GPT-4o (Hurst et al., 2024), Qwen-3B (Yang et al., 2025)) and legal-domain models on widely-used datasets. JURISMA achieves **state-of-the-art performance** across multiple metrics, demonstrating superior capability in producing accurate, legally grounded, and user-aligned responses.

We construct JURISCQAD, a dataset of **43K+** real-world legal consultation instances collected from online legal platforms, to address the third challenge of insufficient training data, as illustrated in **Figure 3**. Most existing Chinese legal QA datasets focus on statute extraction, judgment prediction, or multiple-choice formats, falling short in supporting open-ended, generative consultation tasks or providing large-scale, well-structured responses with high-quality annotations (Li et al., 2024b). JURISCQAD organizes each instance as a **triplet** (question, positive answer, negative answer). To ensure data quality, we leverage our team's legal expertise and collaborate closely with large language models for initial answer generation and refinement. The dataset spans a wide range of high-frequency legal domains, capturing both the linguistic styles and the practical concerns of real users. Experiments demonstrate that models trained on JURISCQAD achieve **substantial improvement** in generating accurate and context-aware responses in legal consultation settings.

Our main contributions are summarized as follows:

(1) To address the **high contextual dependency** in legal consultation, we introduce JURISMA, a pluggable multi-agent framework that decomposes the task into modular subtasks and coordinates them via a Manager Agent for dynamic routing and multi-round refinement.

(2) To tackle the **complex task composition**, we propose a structured semantic graph representation that captures legal entities, events, relations, user intents, and legal issues from free-form queries, serving as a shared context for graph-driven generation.

(3) To alleviate the **data scarcity** in this domain, we construct JURISCQAD dataset with 43K+ real-world queries, each paired with expert-verified positive and negative answers, ensuring high-quality supervision for both training and evaluation.

2 Related Work

Evolution of Legal QA Methodologies. Early systems relied on traditional retrieval methods such as BM25 (Shao et al., 2020; Jayawardena et al., 2024), which performed well on structured statute lookup but struggled with ambiguity and long-form queries. With the rise of generative LLMs, domain-adapted models like LawGPT (Zhou et al., 2024) enhanced semantic understanding via pretraining, yet suffered from uncontrollable reasoning and legal inconsistency. Hybrid pipelines (e.g., retrieve-then-read (Louis et al., 2024)) improved grounding but lacked support for dynamic legal knowledge integration. In contrast, our multi-agent framework leverages a Manager Agent to coordinate subtasks and synchronize legal basis updates during iterative review, improving both completeness and validity.

Legal Knowledge Representation and Augmentation. Static injection methods (e.g., LEGAL-BERT (Shao et al., 2020)) enrich legal embeddings but struggle with evolving laws. Retrieval-augmented methods (e.g., LSIM (Yao et al., 2025)) offer real-time updates via semantic similarity, but often confuse legally distinct yet linguistically similar terms. Our structured semantic graph explicitly models entity–relation–fact chains, reducing ambiguity and enhancing interpretability.

Multi-Agent Approaches in Legal Tasks. Existing legal multi-agent systems often follow rigid pipelines (e.g., LawLuo (Zhang et al., 2025)), limiting adaptability in real consultations. General

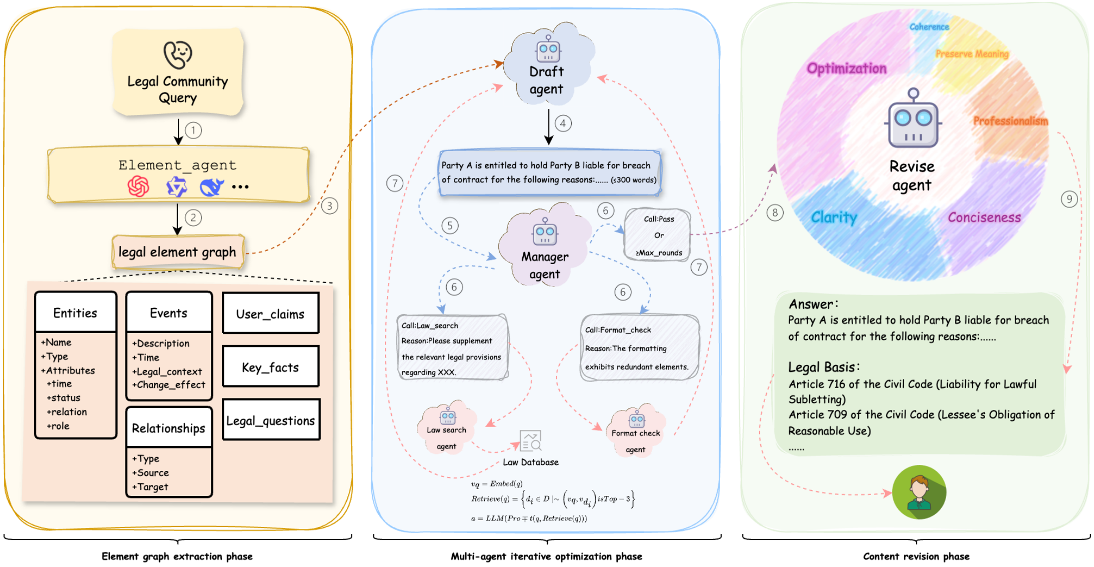

Figure 2: Overview of JURISMA. Legal queries are first parsed into structured graphs, then processed by a set of cooperating agents under the control of a Manager Agent, enabling multi-stage legal response generation.

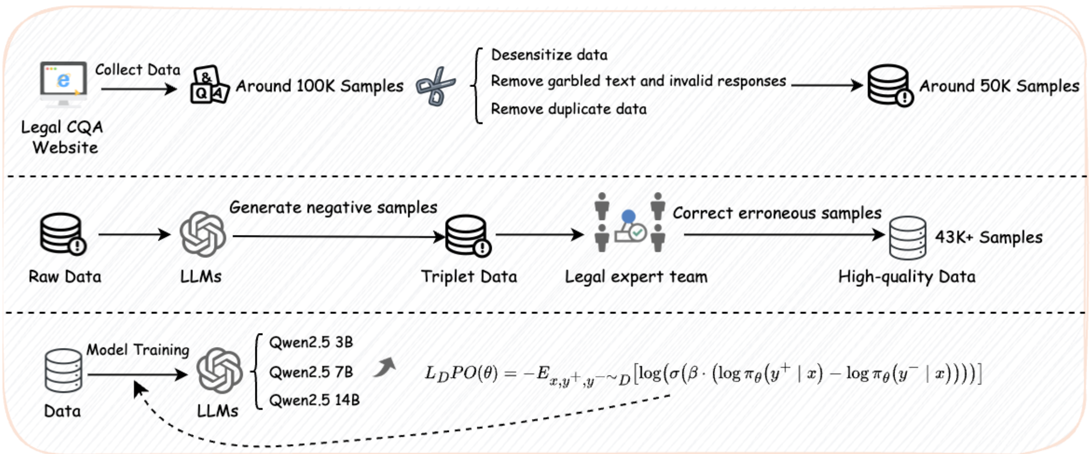

Figure 3: Overview of the JURISCQAD construction pipeline, including real-case collection, negative sample generation, expert validation, and final DPO-based model training.

frameworks like ReAct (Yao et al., 2023) support dynamic reasoning, but lack legal-domain compliance checks. We introduce a Manager Agent that dynamically assesses draft quality and coordinates cooperative repair via FormatCheck and LawSearch agents, mitigating error propagation and enhancing legal robustness.

Data Resources and Evaluation. Datasets like LegalQA (Nigam et al., 2023) and LLeQA (Louis et al., 2024) focus on statute retrieval or synthetic Q&A, lacking realism and linguistic diversity. This problem is acute in Chinese legal NLP, where most models (e.g., LawGPT) are trained on artificial data, leading to domain shift. We address this gap by constructing a 43K-scale dataset of real-world legal consultations with expert-verified triplet annotations (query, positive, negative), covering high-frequency domains and supporting robust, grounded evaluation.

3 Methodology

In this section, we introduce the task formulation, followed by a detailed description of the three primary phases in our multi-agent system: element graph extraction phase, multi-agent iterative optimization phase, and content revision phase.We then describe our test set correction procedure, the construction of a high-quality training dataset, JURISCQAD, and the model training process using Direct Preference Optimization (DPO).

3.1 Task Formulation

Unlike traditional legal QA tasks such as statute matching or multiple-choice assessment, Legal CQA focuses on open-domain user queries posed in natural language that often embed factual complexity and personalized legal dilemmas. These queries are typically informal, structurally loose, and semantically ambiguous, frequently involving multiple overlapping legal relations and reasoning steps. In response to these demands, an effective system must satisfy three essential goals: (1) accurately capture user intent, (2) generate legally sound and context-relevant responses, and (3) present legal references with clarity and professionalism.

Formally, given a user query $q \in \mathcal{Q}$, the goal of Legal CQA is to generate a response $r \in \mathcal{R}$ such that: $\mathrm{Align}(r, q)$ semantic alignment with user's factual context and legal intent $\mathrm{Legal}(r)$ compliance with current Chinese legal provisions $\mathrm{Express}(r)$ professional,accurate, and clear language experssion.

While most existing systems adopt an end-to-end generation paradigm, they often overlook the cognitive and procedural decomposition underlying real-world legal reasoning. In this work, we propose a cognitively inspired, decoupled pipeline that decomposes the task into three phases, as shown in Figure 2. These phases are coordinated through a shared semantic context in the form of an element graph and are orchestrated by a centralized controller, enabling iterative and controllable reasoning workflows.

3.2 Phase I: Element Graph Extraction

Motivated by the observation that legal reasoning revolves around identifying key facts, actors, and their legal relations, we design a *Legal Element Recognition Agent* to extract a graph-based representation G, which explicitly encodes legal entities, events, and their semantic relationships.

This graph-based formulation is theoretically grounded in jurisprudential conceptions of legal reasoning, which view law as a structured system composed of legal subjects, facts, relationships, and norms. Such a perspective is consistent with Hart's theory of primary and secondary rules (Hart, 2012) and Kelsen's hierarchical model of normative systems (Kelsen, 1967).

From a computational standpoint, our design is also informed by advances in structured semantic parsing, particularly Abstract Meaning Representation (AMR) (Banarescu et al., 2013), Semantic Role Labeling (SRL), and knowledge-graph-based question answering.

Specifically, as shown in Figure 2, we define the legal element graph as $G = (V, E)$, where V and E represent the set of nodes and edges, respectively. The node set V includes:

Entities, such as plaintiffs, defendants, or organizations, annotated with attributes like roles, statuses, and temporal markers;

Events, representing legal actions or disputes (e.g., signing, breach of contract);

User claims, **key facts**, and derived **legal questions**.

The edge set E captures **Relations** among these elements, such as kinship, contractual obligations, or causal dependencies.

The resulting graph G is serialized in JSON format and acts as a global contextual abstraction, supporting both interpretability and controllability in downstream modules. For an illustration of this graph, please refer to Appendix C.

3.3 Phase II: Multi-agent Iterative Optimization

Given the constructed legal element graph G, we generate an initial legal opinion draft r_0 using a dedicated *Draft Agent*. To achieve high-quality generation, we convert the structured graph G into a serialized prompt P_G. This prompt is then concatenated with the original user query q to form the model input, providing a rich semantic context for conditioned generation:

$$x = [P_G; q] \quad \Rightarrow \quad r_0 = f(x) \tag{1}$$

Unlike black-box generation, this design emphasizes structure-aware prompting, which aligns the generation trajectory with legal logic and domain expectations. The output r_0 serves as a preliminary response and enters the refinement loop governed by downstream agents.

To ensure the draft response satisfies legal adequacy and linguistic clarity, we introduce a centralized decision-making module, termed the *Manager Agent*. Rather than following a fixed pipeline, the manager dynamically assesses the quality of r_0 and routes it to appropriate sub-agents for refinement. The detailed procedure is illustrated in Algorithm 1.

The design of the Manager Agent is grounded in both cognitive and computational theories. It mirrors human legal writing workflows, where

Algorithm 1: Multi-Agent Controlled Draft Optimization

Input : User query q; initial draft $r_0 = f([P_G; q])$
Output : Final legal response r_{final}
▷ r_t: draft at iteration t
▷ a_t: routing action from ManagerAgent
▷ T: max number of iterations
▷ f: initial generation model with structured input

$t \leftarrow 0$;
while $t < T$ **do**
  $a_t \leftarrow$ ManagerAgent(r_t);
  if $a_t = \emptyset$ ***or*** $a_t =$ `Pass` **then**
    break;
  if a_t *includes* `FormatCheck` **then**
    $r_t \leftarrow$ FormatAgent(r_t);
  if a_t *includes* `LawSearch` **then**
    $r_t \leftarrow$ LawSearchAgent(r_t, q);
  $t \leftarrow t + 1$;
$r_{\text{final}} \leftarrow$ ContentCheckAgent(r_t);
return r_{final};

a draft typically undergoes structural review, legal verification, and stylistic polishing by different experts (Ashley, 2017). Technically, our approach aligns with modular control architectures in multi-agent systems, where a centralized planner dynamically activates task-specific modules based on intermediate outputs (Russell and Norvig, 2016). Moreover, the iterative feedback loop used for draft refinement parallels recent advances in planning-based generation and multi-pass text optimization (Zhang et al., 2020).This multi-agent loop enables dynamic correction and incremental quality improvement until the response meets standards for legal compliance, factual adequacy, and stylistic fluency.

3.4 Phase III: Content Revision

Upon completion of all revision phases, a *Content Optimization Agent* finalizes the output by rewriting it into a professionally formatted, user-facing response. This process adheres to several standards—such as clarity, conciseness, professionalism, and coherence—to ensure the response aligns with practical expectations in legal communication. The output is explicitly partitioned into two segments:

Response: the main advisory opinion directed toward the user;

Legal Basis: the referenced statutory articles supporting the recommendation.

This dual-structured output not only facilitates user comprehension but also ensures legal accountability and transparency, aligning the system's output with real-world consultation norms.

3.5 Test Set Correction and Data Curation

High-quality open-source evaluation datasets for Chinese legal consultation remain scarce. For evaluation, we adopt the widely used LawBench dataset (Fei et al., 2024), which provides question-answer pairs sourced from real-world legal consultation websites. However, upon close inspection, we observed a significant number of responses in the test set containing legal inaccuracies, including but not limited to incorrect conclusions, irrelevant or misleading answers, and misinterpretation of legal provisions. Such issues substantially undermine the reliability of evaluation results.

To address this, we conduct a systematic correction of the test set using a collaborative strategy involving both model-assisted review and expert verification. Specifically, we employ LLMs to automatically identify legal inaccuracies and generate revision suggestions, which are then reviewed line-by-line by licensed legal professionals to ensure linguistic clarity and legal validity. Detailed justifications for each revision are provided in Appendix D.We fully acknowledge and appreciate the contributions of the LawBench team to the field of legal NLP; our corrections are intended solely to provide a more reliable and equitable evaluation environment for fair model comparison.

More significantly, we construct a large-scale Chinese legal consultation dataset, JurisCQAD, consisting of over 43K examples sourced from real-world user queries. As shown in Figure 3, we first perform data cleaning to remove noise and inconsistencies. Next, a LLM is used to generate challenging negative answers that are fluent but legally inaccurate. Finally, we perform expert-guided verification to ensure the accuracy, validity, and consistency of both positive and negative samples, leveraging our team's legal training and domain expertise. This process results in high-quality training triplets (query,positive,negative).

We adopt the DPO framework for training on this dataset. Each training instance is organized as a triple (x, y^+, y^-), where x denotes the user query, y^+ is a high-quality human reference response, and y^- is a model-generated suboptimal response. The training objective is defined as:

$$\Delta_\theta(x, y^+, y^-) = \log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x) \quad (2)$$

$$\mathcal{P}_\theta(x, y^+, y^-) = \sigma\left(\beta \cdot \Delta_\theta(x, y^+, y^-)\right) \quad (3)$$

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\log \mathcal{P}_\theta(x, y^+, y^-)\right] \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid function and β is a temperature scaling parameter.

Section 4.3 provides strong empirical evidence for the effectiveness of this dataset.

4 Experiment Details

4.1 Experimental Setup

Dataset and evaluation metrics. We conduct comprehensive evaluations using a revised version of LawBench (Fei et al., 2024), one of the most widely used benchmarks in Chinese legal consultation. As detailed in Section 3, we identified factual and legal issues in the original answers and applied a hybrid correction process to improve reliability. The resulting test set offers legally accurate, well-structured responses and serves as the gold standard for all experiments.

We report performance using multiple commonly adopted metrics: Rouge-1, Rouge-2, Rouge-L (Lin, 2004), Bleu-1, Bleu-2, Bleu-N (Papineni et al., 2002), BertScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020). All scores are computed using official or widely adopted implementations with default parameters.

Baselines. We compare our method against a diverse set of strong baselines, grouped into two categories: (1) General LLMs, including GPT-4o (Hurst et al., 2024) and Qwen3-14B (Yang et al., 2025), represent state-of-the-art general-purpose chat models without legal domain specialization. Notably, Qwen3-14B was released in late April 2025, reflecting the latest advances in open-source instruction-tuned LLMs. (2) Legal LLMs, including Chatlaw-33b (Cui et al., 2024), Fuzi-mingcha (Deng et al., 2023), Hanfei (He et al., 2023), Lawgpt (Zhou et al., 2024), Lawyer-llama (Huang et al., 2023), Lexilaw (Li et al., 2024a), and Wisdom-Interrogatory (Wu et al., 2024). These models are specifically trained or adapted for Chinese legal tasks, with various sources of legal corpora. All models are evaluated in zero-shot using the same prompt template. All reported results are averaged over 5 runs with different random seeds to ensure robustness and reduce variance.

Implement Details. We introduce the selected LLMs, hyperparameter settings and other details in Appendix A.

4.2 Main Results

The main results are presented in Table 1, and we summarize our key findings as follows:

(1) Our method achieves the best overall performance across most evaluation metrics. Compared to both general-purpose and legal-specialized LLMs, our approach consistently outperforms all baselines on ROUGE, BLEU-2, BLEU-N, and BERTSCORE, and ranks second in BLEURT. By explicitly modeling legal factors and user intent, our system improves both factual accuracy and legal relevance in the generated responses.

(2) Legal LLMs vary significantly in performance depending on their training data and alignment strategies. Models such as WISDOM and HANFEI perform relatively well on several metrics, while others like LAWGPT and CHATLAW-33B consistently underperform. These discrepancies may stem from differences in the quality of legal corpora, the presence (or absence) of legal reasoning supervision, and the degree of alignment with user-oriented tasks.

(3) Our method excels in both lexical precision and semantic fidelity. Its superior BERTSCORE and near-best BLEURT scores demonstrate the system's ability to generate legally sound, contextually appropriate responses that go beyond superficial token overlap. This is especially important in legal consultation scenarios, where semantic correctness and normative justification are more critical than surface-level phrasing.

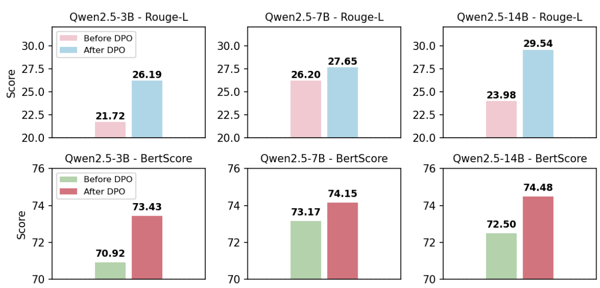

Figure 4: Rouge-L and BertScore comparison before and after DPO across Qwen2.5 models (3B/7B/14B), with consistent gains across model sizes validating the generalizability of the dataset.

4.3 Dataset Evaluation

To evaluate the effectiveness of our constructed legal consultation dataset, we conduct DPO on three

Table 1: Main results on the revised LawBench test set. "†" indicates statistically significant improvement over all baselines under a paired t-test with $p < 0.05$. Bold numbers denote the best performance. Underlined numbers indicate the second-best results.

Models	Rouge (%)			Bleu (%)			BertScore (%)	Bleurt (%)
	Rouge-1	Rouge-2	Rouge-L	Bleu-1	Bleu-2	Bleu-N		
General LLM								
GPT4o	40.24	15.27	24.50	**34.94**	12.34	8.89	73.22	55.16
Qwen3-14b	42.55	19.24	27.27	21.93	10.25	8.10	74.64	**62.38**
Legal LLM								
Chatlaw-33b	27.78	7.40	18.15	17.84	3.81	1.46	67.77	57.54
Fuzi-mingcha	32.31	9.92	17.41	23.13	7.07	5.19	70.47	52.46
Hanfei	30.79	10.21	18.37	13.47	4.26	2.68	69.68	58.37
Lawgpt	20.52	5.22	7.26	5.15	1.01	0.49	63.44	46.94
Lawyer-llama	30.61	9.50	18.82	24.65	6.28	3.35	69.17	58.47
Lexilaw	31.23	9.61	18.15	14.40	4.76	3.33	70.00	58.12
Wisdom	36.60	15.74	23.12	34.89	10.16	8.04	71.97	56.09
Our Method								
JurisMA	**44.68**†	**23.42**†	**31.14**†	32.54	**16.18**†	**14.25**†	**75.05**	58.63

general-purpose LLMs from the Qwen2.5 series: 3B, 7B, and 14B. Figure 4 summarizes the performance before and after DPO fine-tuning across multiple metrics.

(1) DPO consistently improves performance across all model scales. All models exhibit notable gains in both lexical metrics (e.g., ROUGE and BLEU) and semantic metrics (e.g., BERTSCORE). For example, Qwen2.5-3B sees an improvement of +5.15 on ROUGE-L and +3.26 on BLEU-2. The improvements are more pronounced in larger models, with Qwen2.5-14B achieving a +5.56 point gain on ROUGE-L and +4.45 on BLEU-N.

(2) The results validate the usefulness of our dataset for DPO-style supervision. Despite being constructed from real-world queries and weakly supervised negative samples, the dataset proves effective in aligning model outputs with legal correctness and stylistic clarity, even without additional reward modeling or prompt engineering.

4.4 Ablation Study

To assess the contribution of each module in our system, we conduct an ablation study by removing key components individually: KG (legal element graph), Manager (decision routing), and Revise process. The results for both Qwen2.5-7B and 14B models are presented in Table 2.

(1) All modules contribute to performance. The full system achieves the best results, with Qwen2.5-14B reaching 31.14 ROUGE-L, 14.25 BLEU-N, and 75.05 BERTSCORE. Removing any module leads to clear degradation.

Table 2: Ablation study showing performance drops when removing key modules. Results confirm the importance of structured graph input, dynamic task routing, and iterative refinement.

Models	Rouge-L (%)	Bleu-N (%)	BertScore (%)
w/o KG			
Qwen2.5:7b	21.33 (-6.68)	7.72 (-3.6)	71.23 (-2.59)
Qwen2.5:14b	21.57 (-9.57)	7.21 (-7.04)	71.74 (-3.31)
w/o Manager			
Qwen2.5:7b	21.41 (-6.6)	7.81 (-3.51)	71.62 (-2.2)
Qwen2.5:14b	21.30 (-9.84)	6.89 (-7.36)	71.68 (-3.37)
w/o Revision			
Qwen2.5:7b	20.58 (-7.43)	7.10 (-4.22)	70.64 (-3.18)
Qwen2.5:14b	19.96 (-11.18)	6.10 (-8.15)	69.56 (-5.49)
Full			
Qwen2.5:7b	**28.01**	**11.32**	**73.82**
Qwen2.5:14b	**31.14**	**14.25**	**75.05**

(2) Revision is critical for semantic quality. Excluding the revision stage results in the largest BERTSCORE drop (–5.49 for 14B), showing that iterative polishing improves fluency and legal clarity.

(3) The manager enhances consistency. Removing the manager reduces BLEU-N and BERTSCORE, confirming its role in dynamic feedback and controllable generation.

(4) The legal graph boosts factual grounding. Without the legal graph, both ROUGE-L and BERTSCORE decline, indicating its importance in encoding legally salient facts.

Question	我在2011年有过一次醉酒驾驶，那时候醉驾还没有入刑，从那年的5.1号以后醉驾才入刑的请问影响孩子报考航空类学校吗? I had a DUI (driving under the influence) incident in 2011, but at that time, drunk driving had not yet been classified as a criminal offense — it only became a criminal offense after May 1st of that year. May I ask if this would affect my child's eligibility to apply for an aviation-related school?
GPT4o	回答：一般来说，醉酒驾驶的记录可能会影响到孩子报考航空类学校，尤其是在一些对学生背景审核较严格的学校或机构，因为航空类专业对纪律和个人品德有较高的要求。然而，由于您的醉酒驾驶行为发生在《刑法修正案（八）》生效之前，它不构成刑事犯罪记录，只是行政处罚，所以影响可能有限。这仍然取决于具体学校的招生政策。法律依据：**1.《中华人民共和国刑法修正案（八）》**：根据2011年施行的这项法律，醉酒驾驶才被正式纳入刑法，因此您在2011年4月30日前的醉驾行为不构成刑事犯罪。**2.《中华人民共和国道路交通安全法》**：在醉酒驾驶入刑之前，醉驾属于行政违法行为，受行政处罚。**3. 各航空类学校的招生条例**：这些条例会具体规定申请者的背景要求，可能会涉及到无犯罪记录或良好的品德记录，具体条款因学校而异。建议您直接咨询具体航空类学校的招生部门以获得详细信息。 Answer:In general, a record of drunk driving may affect a child's eligibility to apply to aviation-related schools, ..., as aviation programs place high importance on discipline and personal integrity. However, since your drunk driving incident occurred before the Amendment (VIII) to the Criminal Law of the People's Republic of China came into effect, it does not constitute a criminal record but only an administrative penalty, so the impact may be limited. This still depends on the specific admission policies of each school.Legal Basis:**Amendment (VIII) to the Criminal Law of the People's Republic of China: ......Road Traffic Safety Law of the People's Republic of China: ......Admissions regulations of aviation schools: ......**
LawGPT	答：根据**《中华人民共和国道路交通安全法》第九十九条规定**，饮酒后驾驶机动车的，公安机关交通管理部门应当约束处罚，并记3个记分分：一次饮酒后驾驶机动车的，记3个记分分，并处以一千元以上二千元以下的罚款，并暂扣六个月机动车驾驶证。 Answer:According to **Article 99 of the Road Traffic Safety Law of the People's Republic of China**, anyone who drives a motor vehicle after consuming alcohol shall be subject to administrative penalties imposed by the public security traffic management department, and 3 demerit points shall be recorded. Specifically, for a single instance of driving after drinking, 3 demerit points will be added, a fine of no less than 1,000 yuan and no more than 2,000 yuan will be imposed, and the driver's license will be temporarily suspended for six months.
JudicaMA	回答：您在2011年5月1日前醉酒驾驶的行为不构成刑事犯罪，因此不会影响孩子报考航空类学校。\n **法律依据：《中华人民共和国刑法》第一百三十三条之一（危险驾驶罪）规定**：\"在道路上驾驶机动车，有下列情形之一的，处拘役，并处罚金：（一）追逐竞驶，情节恶劣的；（二）醉酒驾驶机动车的；......\"。该条款自2011年5月1日起生效，根据\"法不溯及既往\"原则，您的行为不构成犯罪。" Answer:Your act of drunk driving before May 1, 2011, does not constitute a criminal offense, and therefore will not affect your child's eligibility to apply to aviation-related schools.Legal Basis:**Article 133-1 of the Criminal Law of the People's Republic of China (Crime of Dangerous Driving) states:**"Anyone who drives a motor vehicle on the road under any of the following circumstances shall be sentenced to criminal detention and fined: (1) engaging in racing with other vehicles under aggravated circumstances; (2) driving a motor vehicle while intoxicated; ..."This provision came into effect on May 1, 2011. According to the legal principle of "non-retroactivity of the law," your behavior does not constitute a crime.

Figure 5: Case study comparing model outputs on a time-sensitive legal query. Color highlights indicate legal focus, conclusion, statute, and principle

4.5 Generalization Evaluation

To assess the robustness and generalizability of our method beyond the Chinese-language test sets, we further evaluate JURISMA on the English RULEQA subset from the LegalBench benchmark. This dataset contains factually complex, rule-based legal questions that require precise statutory reasoning. We randomly sample 50 instances and perform evaluation across 5 different random seeds to obtain averaged results. As detailed in Appendix E, JURISMA consistently outperforms both general-purpose and legal-domain LLMs, demonstrating strong cross-lingual and task-level generalization capability.

4.6 Case Study

To qualitatively assess interpretability and legal reasoning, we present a case study based on a real-world consultation: whether a man's drunk driving incident in early 2011 would affect his child's eligibility for applying an aviation school. The core legal issue is time-sensitive, as drunk driving was not classified as a criminal offense until May 1, 2011.

As shown in Figure 5, we annotate model responses with four highlight colors: Dark red sentences indicate the core legal focus; Green sentences denote each model's final conclusion; Light red sentences mark the cited statutory basis; Blue sentences highlight references to underlying legal principles.

Our method delivers the most accurate and concise judgment: the incident occurred before the law came into effect, hence no criminal record, and no eligibility issue. It explicitly cites Article 133-1 of the Criminal Law and invokes the principle of non-retroactivity. In contrast, GPT-4o provides a lengthy, less focused explanation, and LAWGPT fails to cite the most directly applicable statute, omitting precise legal grounding.

This example demonstrates our system's strength in aligning factual analysis with legal authority and reasoning principles, enhancing trustworthiness in high-stakes consultation scenarios.

5 Conclusion

In this paper, we propose JURISMA, a cognitively inspired multi-agent framework for Legal Consultation Question Answering (Legal CQA). By converting complex legal queries into structured semantic graphs and coordinating agents via a centralized Manager Agent, our method enables controllable, interpretable, and legally compliant reasoning.

We also introduce JURISCQAD, a 43K-scale expert-validated dataset of real-world legal consultations, supporting preference-aligned DPO training and robust evaluation.

Experiments show that JURISMA significantly outperforms both general and legal-specialized LLMs. Ablations confirm the impact of structured prompting and dynamic routing, while case studies demonstrate alignment with statutory principles and legal reasoning norms.

Limitations

While JURISMA achieves strong performance, several limitations remain. First, the multi-agent architecture introduces additional latency, which may hinder real-time deployment.Second, although JURISCQAD covers diverse legal scenarios, it may still be biased toward high-frequency consultation topics, limiting performance on rare or complex cases. Finally, our evaluation is based on automatic metrics and expert-verified datasets; integration of human-in-the-loop feedback and real-world user studies remains an important direction for future work.

References

Kevin D Ashley. 2017. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Herbert Lionel Adolphus Hart. 2012. *The concept of law*. OUP Oxford.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0. `https://github.com/siat-nlp/HanFei`.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *Preprint*, arXiv:2305.15062.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lasal Jayawardena, Nirmalie Wiratunga, Ramitha Abeyratne, Kyle Martin, Ikechukwu Nkisi-Orji, and Ruvan Weerasinghe. 2024. Scale-qa: Sri lankan case law embeddings for legal qa.

Hans Kelsen. 1967. *Pure theory of law*. Univ of California Press.

Haitao Li, Qingyao Ai, Qian Dong, and Yiqun Liu. 2024a. Lexilaw: A scalable legal language model for comprehensive legal understanding.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024b. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Preprint*, arXiv:2409.20288.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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, pages 22266–22275.

Yixiao Ma, Yueyue Wu, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. Caseencoder: A knowledge-enhanced pre-trained model for legal case encoding. *arXiv preprint arXiv:2305.05393*.

Shiwen Ni, Hao Cheng, and Min Yang. 2025. Pre-training, fine-tuning and re-ranking: A three-stage framework for legal question answering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

SK Nigam, SK Mishra, AK Mishra, N Shallum, and A Bhattacharya. 2023. Legal question-answering in the indian context: Efficacy, challenges, and potential of modern ai models. arxiv, 1–15.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. pearson.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, volume 2020, pages 3501–3507.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, and Kun Kuang. 2024. wisdominterrogatory. Available at GitHub.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025. Elevating legal llm responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning. *arXiv preprint arXiv:2502.07912*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yu Zhang, Shutong Qiao, Jiaqi Zhang, Tzu-Heng Lin, Chen Gao, and Yong Li. 2025. A survey of large language model empowered agents for recommendation and search: Towards next-generation information retrieval. *arXiv preprint arXiv:2503.05659*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2406.04614.

A More Details for Experimental Setup

A.1 JurisCQAD Dataset Details

The core structure and statistics of the JurisCQAD dataset are summarized in Table 3.

Table 3: Summary of JurisCQAD Dataset

Property	Description
Source	Real-world legal consultation platforms
Language	Chinese
Size	43,126 triplets
Data Format	(query, positive answer, negative answer)
Annotation Method	LLM-assisted generation + expert verification (4 legal experts)
Negative Sample Strategy	LLM-generated distractors with legal/semantic flaws
Domains Covered	Contract law, tort liability, family law, labor disputes, etc.
Average Query Length	15.14 tokens
Average Positive Answer Length	264.97 tokens
Average Negative Answer Length	194.79 tokens
Usage	Used for supervised fine-tuning and DPO training

A.2 Implementation Details

We perform DPO fine-tuning on Qwen2.5 models of three different sizes (3B, 7B, 14B). All models are trained with LoRA adapters (rank=8, α=16) using the HuggingFace + Deepspeed framework (Stage 2) on up to 2 × A100 80GB GPUs. Gradient accumulation is set to 8, and we use a batch size of 8 per device, for an effective batch size of 128. Mixed precision training is enabled via bf16. All training runs use AdamW with a cosine learning rate schedule, an initial learning rate of 1×10^{-5}, and no warm-up.

The dataset used is JurisCQAD, comprising 43K+ real-world consultation queries with expert-verified (query, positive, negative) triplets. We train for 3 epochs with max sequence length 1024. The DPO β is set to 0.1, and the loss is computed using sigmoid preference loss without reward normalization.

All prompts follow the Qwen dialogue template, with system instructions embedded. We do not apply quantization or offloading, and DeepSpeed offload is disabled. Model checkpoints are saved every 100 steps. No external reward model or RLHF phase is used. Evaluation is performed in zero-shot mode using the same prompt template across all models.

In the JURISMA, we employ Qwen2.5-14B-Instruct as the underlying model for all agent components. Although newer models such as Qwen3-14B-Instruct have been released with stronger base capabilities, our method—when built upon Qwen2.5—still consistently outperforms Qwen3-14B across all metrics in legal consultation tasks. This highlights the effectiveness of our framework design, independent of backbone improvements. We deliberately avoid using larger models to ensure reproducibility and reduce computational costs, thereby demonstrating that strong performance can be achieved through structural innovation rather than model scaling alone. All datasets and model baselines used in this study are publicly available under licenses that permit academic use. We ensure that our use is consistent with their intended purpose, strictly limited to research contexts.

Our proposed dataset, JURISCQAD, was constructed from publicly accessible legal consultation forums. All collected samples underwent careful anonymization and manual screening to eliminate personally identifiable information (PII) and potentially offensive content. To ensure ethical integrity, all data was processed solely for non-commercial research use, in line with prevailing data use policies and licensing norms. The dataset will be released for academic purposes only under a research-friendly license.

B Legal QA & Legal CQA Comparison

Table 4 provides a comparative overview of traditional Legal QA tasks and the more complex Legal Consultation QA (Legal CQA), highlighting their differences in task objectives, data sources, and evaluation metrics.

Table 4: Comparison between Legal QA and Legal Consultation QA (Legal CQA)

Comparison Dimension	**Legal QA**	**Legal CQA**
Task Goal	Answer exam questions or legal provisions	Respond to real-world legal concerns from users
Task Type	Mostly multiple-choice or extraction tasks	Requires generation of context-relevant legal suggestions
Data Source	Legal exams, statutory texts	Legal forums and Q&A communities
Data Characteristics	Standardized answers, concise questions	Long, complex questions with diverse factual scenarios
Question Structure	Short, standardized text	Long, unstructured, and often informal expressions
Legal Context	Involves a single legal provision	Involves multiple statutes and factual elements
Evaluation Metrics	Accuracy, F1 score	BLEU, METEOR, human evaluation (completeness, professionalism, relevance)
Answer Diversity	Single correct answer	Multiple plausible answers depending on context

C Example of Element Graph

Table 5 presents each node and its detailed attributes in the element graph extracted from the case study Figure 5 question: "I had a drunk driving incident in 2011. At that time, drunk driving had not yet been criminalized. It was only after May 1st of that year that it became a criminal offense. Will this affect my child's application to an aviation school?"

D Test Set Correction Cases

To enhance the reliability and legal validity of evaluation, we manually reviewed and revised a subset of LawBench test cases. Among 500 test queries, 340 were found to contain legal or factual errors and were subsequently corrected. Each correction involved identifying flaws in the original answer, followed by regeneration using expert-reviewed LLMs. Table 6 summarizes representative examples and reasons for revision.

E Generalization study

Table 7 shows generalization results on the LEGALBENCH-RuleQA subset (50 samples, averaged over 5 seeds).

Table 5: Example of Element Graph

Section	Content
Entities	• **User (Person)**: Drunk driving record in 2011; the user themself committed the act. • **Child (Person)**: Child of the user; intends to apply for an aviation school. • **Drunk Driving (Illegal Act)**: Occurred in 2011; not criminalized at the time (before May 1, 2011). • **Aviation School (Institution)**: Has specific eligibility requirements for applicants.
Event	• **Description**: The user committed drunk driving in 2011. • **Time**: 2011 • **Legal Context**: Drunk driving was not yet criminalized before May 1, 2011. • **Change Effect**: Criminalization started after May 1, 2011, but the user's act occurred earlier.
Relationships	• **Kinship**: User → Child • **Application Target**: Child → Aviation School • **Legal Involvement**: User → Drunk Driving
User Claims	• What should I do? • Is it illegal? • Can my child apply to an aviation school?
Key Facts	• The user committed drunk driving in 2011. • Drunk driving was not criminalized before May 1, 2011. • The user's child intends to apply for an aviation school. • Aviation schools have specific background requirements.
Legal Questions	• Will the drunk driving record affect the child's application to aviation school? • Was drunk driving a criminal offense in 2011? • What are the background screening standards for aviation school applicants?

Table 6: Representative Examples of Test Set Corrections

Case ID	Revision Reason	Key Correction
1	Original answer failed to distinguish pre-/post-May 1, 2011 legal status of drunk driving and omitted aviation-specific background check regulations.	Added analysis of non-criminal administrative penalty and cited *Article 133-1 of the Criminal Law* and aviation review guidelines.
2	Original answer discussed unrelated payment default topic and lacked any applicable law to the real estate recovery dispute.	Rewritten answer clarified legal ownership transfer, invoked *Civil Code* articles on registration, good faith acquisition, inheritance, and statute of limitations.
3	Original answer incorrectly stated that all owners must sign service contracts. It misunderstood the legal effect of contracts signed by the owners' committee and confused public and private contracting rules.	Clarified that a legally signed service contract by the owners' committee is binding for all owners under *Civil Code Article 939* and *Property Management Regulations*.
4	Original answer cited outdated or inaccurate medical insurance provisions and failed to reflect local retirement policies.	Updated answer clarified retirement exemption from further payments, citing *Social Insurance Law* and regional cumulative contribution rules.
5	Original answer misunderstood liability in sublease and construction. Misapplied contract law and omitted tenant's liability for subtenants' actions.	Added correct explanation using *Articles 714, 716, 577 of Civil Code*, showing tenant's liability for third-party damages and breach of duty to maintain the property.
6	Original answer failed to cite core law on execution exemption and missed user's intent to reserve minimum livelihood funds.	Correction referenced *Civil Procedure Law Article 243* and Supreme Court regulations on exempt property and basic living standards.
7	Original answer did not address user's question on how to reserve part of pension funds during execution. It also missed the legal basis for such exemption.	Clarified court must reserve necessary funds during pension account freeze, citing *Civil Procedure Law* and relevant enforcement provisions.
8	Original answer failed to answer whether the drawer could stop check payment. Misquoted irrelevant provisions and missed core check law rules.	Corrected to include legal conditions under which a drawer may suspend payment, citing *Negotiable Instruments Law*, *Payment and Settlement Measures*, and Supreme Court judicial interpretations.

Table 7: Generalization results on the LegalBench-RuleQA subset (50 samples, averaged over 5 seeds). "†" indicates statistically significant improvement over all baselines under a paired t-test with $p < 0.05$. Bold numbers denote the best performance. Underlined numbers indicate the second-best results.

Models	Rouge (%)			Bleu (%)			BertScore (%)
	Rouge-1	**Rouge-2**	**Rouge-L**	**Bleu-1**	**Bleu-2**	**Bleu-N**	
			General LLM				
GPT4o	10.10	1.40	8.11	4.66	0.49	0.21	57.88
			Legal LLM				
Lawyer-llama	8.42	0.07	7.70	3.70	0.03	0.97	57.13
			Our Method				
JurisMA	**20.25**†	**10.72**†	**12.81**†	**16.94**†	**6.72**†	**5.41**†	**70.48**†