From Query to Counsel: Structured Reasoning with a Multi-Agent Framework and Dataset for Legal Consultation

Anonymous ACL submission

Abstract

Legal consultation question answering (Legal CQA) presents unique challenges compared to traditional legal QA tasks, including the scarcity of high-quality training data, complex task composition, and strong contextual dependencies. To address these, we construct JURISCQAD, a large-scale dataset of over 43,000 real-world Chinese legal queries annotated with expert-validated positive and negative responses, and design a structured task decomposition that converts each query into a legal element graph integrating entities, events, intents, and legal issues. We further propose JURISMA, a modular multi-agent framework supporting dynamic routing, statutory grounding, and stylistic optimization. Combined with the element graph, the framework enables strong context-aware reasoning, effectively capturing dependencies across legal facts, norms, and procedural logic. Trained on JURISCQAD and evaluated on a refined LawBench, our system significantly outperforms both general-purpose and legal-domain LLMs across multiple lexical and semantic metrics, demonstrating the benefits of interpretable decomposition and modular collaboration in Legal CQA.

1 Introduction

Legal consultation question answering (Legal CQA) is an emerging core task in legal artificial intelligence. It targets real-world user queries involving personalized legal dilemmas, and aims to generate contextually grounded, executable legal advice (Zhong et al., 2020a; Louis et al., 2024). Its practical value, social impact, and high complexity make it an essential benchmark for evaluating legal reasoning capabilities in language models.

Consider the example in Figure 1. Answering such a query requires: (1) Identifying legal relationships - recognizing A and C lack direct privity while B serves as the contractual intermediary; (2) determining causation - tracing the damage to C's unauthorized structural alteration; (3) identifying the core issue - explaining why A can bypass C to hold B liable; (4) reasoning and legal matching - based on the above analysis, concluding that the principle of contractual relativity and B's supervisory obligations apply, then matching with corresponding civil code provisions. This example illustrates that Legal CQA demands: (1) Strong legal background knowledge, (2) legal-oriented query decomposition, and (3) powerful contextual understanding and reasoning.

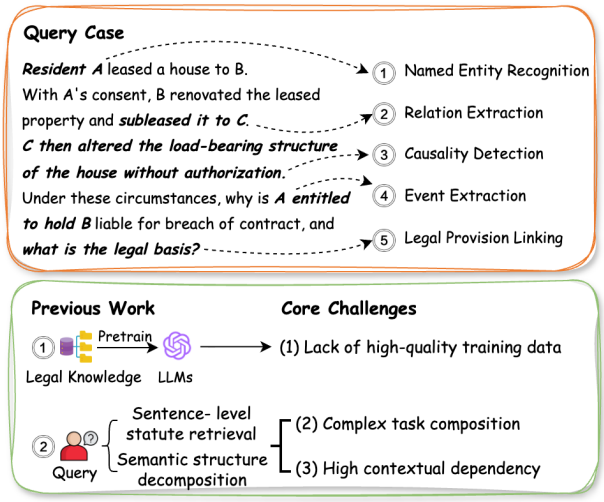

Figure 1: An illustrative example of legal consultation task decomposition, highlighting key challenges, limitations of prior approaches

Previous studies on Legal CQA generally follow two main workflows. The first line of work seeks to enhance legal knowledge within large language models by continuing pretraining on legal statutes, judicial opinions, and domain-specific corpora (Huang et al., 2023). This strategy aims to enrich the model's understanding of legal terminology, doctrinal structures, and statutory patterns in an unsupervised manner—but due to coarse data processing, limited coverage, and non-scenario-based sources, the resulting supervision is low-

quality, leading to marginal performance gains. The second line of work focuses on input restructuring, such as retrieving sentence-level legal provisions to provide explicit legal grounding for generation (Ma et al., 2023; Ni et al., 2025). These methods typically rely on retrieval pipelines to extract relevant statutes or past cases, which are then concatenated with the user query to inform the model's response. However, these methods fall short in handling real-world legal consultations, which are often vague, multi-faceted, and require dynamic interpretation of facts, actors, and legal implications beyond static law matching.

We thus identify three core challenges in Legal CQA: 1) **Lack of high-quality training data** that reflects realistic legal consultation scenarios; 2) **complex task composition**, involving multiple, interdependent subtasks; and 3) **high contextual dependency**, requiring precise interpretation of legal entities, relationships, and user intent.

To address these challenges, we present JURISCQAD, a large-scale benchmark dataset comprising over 43,000 real-world legal consultation instances, each organized as a triplet of (question, positive/negative answer). This dataset supports open-ended, generative Legal CQA tasks and enables high-quality model training and evaluation. Building on JURISCQAD, we propose JURISMA, a multi-agent framework designed to simulate real-world legal decision-making. JURISMA decomposes complex consultation tasks into structured graph-based elements, capturing legal entities, relationships, intents, and issues. Then it employs a cooperative multi-agent architecture to solve these subtasks and reach consensus. The method is evaluated against general large language models (LLMs) and legal-specialized LLMs, achieving state-of-the-art performance across multiple evaluation metrics via training on JURISCQAD.

The contribution of this work is three-fold:

(1) A large-scale, high-quality dataset, tailored for Legal CQA: It enables model training and evaluation, with significant improvement in generating accurate and context-aware legal responses.

(2) A structured graph-based task decomposition strategy: It extracts legal entities, events, relationships, user intents, and legal issues from free-form queries, offering a flexible and effective framework for Legal CQA.

(3) A pluggable and modular multi-agent system where a Manager Agent dynamically coordinates subtasks across specialized agents through multi-round refinement: This framework demonstrates strong performance and adaptability in handling high-context legal consultations.

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 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 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 an-

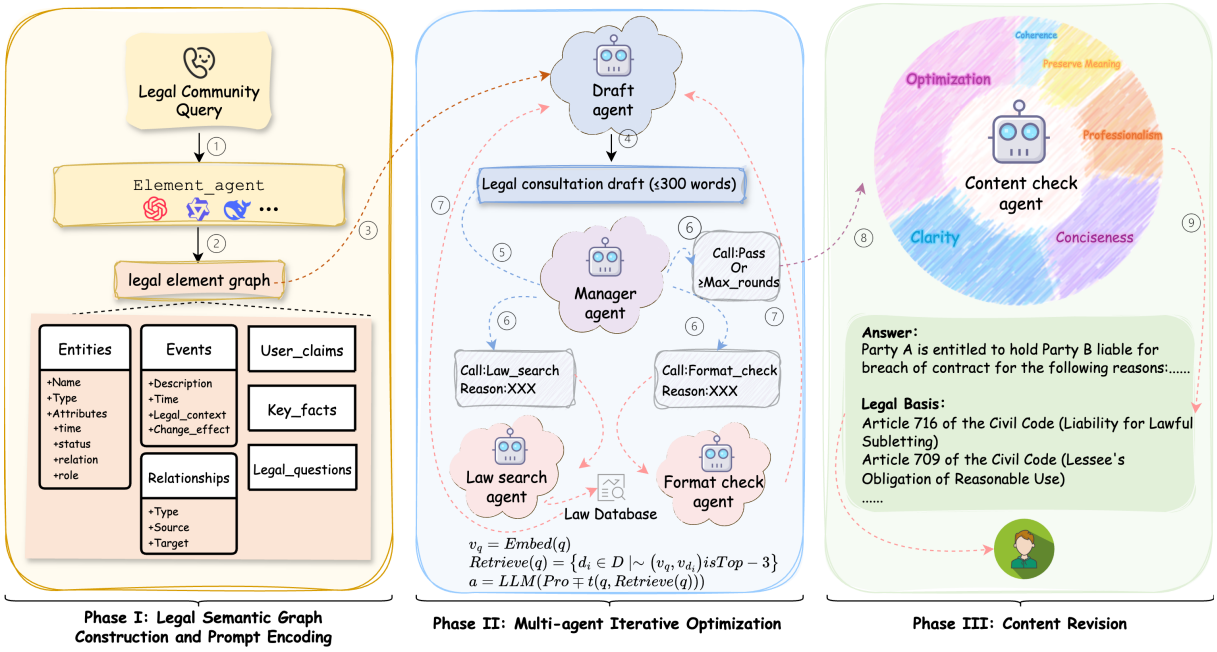

Figure 2: Overview of JURISMA, a multi-agent framework that parses legal queries into element graphs, refines drafts via agent collaboration, and outputs a final legal opinion with supporting statutes.

notations, 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 JURISMA. We then describe the test dataset correction procedure, the construction of JURISCQAD, and the model training process.

3.1 Task Formulation

Legal CQA differs fundamentally from traditional legal QA tasks—such as statute retrieval or multiple-choice assessments—which operate on well-defined inputs and constrained options. In contrast, Legal CQA deals with open-domain user queries expressed in natural language, often characterized by factual ambiguity, informal phrasing, and personalized legal dilemmas. These queries commonly involve multiple entangled legal concepts and reasoning steps. A detailed comparison is provided in Table 8.

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: $\text{Align}(r, q)$ semantic alignment with user's factual context and legal intent $\text{Legal}(r)$ compliance with current Chinese legal provisions $\text{Express}(r)$ professional,accurate, and clear language experssion.

Aspect	Legal QA	Legal CQA
Task Focus	Legal exams, etc.	Real-world legal advice
Input Style	Short, structured	Long, informal
Output Type	Single correct answer, etc.	Contextualized suggestions
Data Source	Statutes, exams	Legal forums
Legal Scope	Narrow	Broad, multi-faceted
Evaluation	Accuracy, F1	BLEU, BERTScore

Table 1: Comparison between traditional Legal QA and Legal Consultation QA (Legal CQA).

Formally, given a user query $q \in \mathcal{Q}$, the objective of Legal CQA is to generate a response $r \in \mathcal{R}$ that satisfies the following criteria: (1) $\text{Align}(r, q)$: Semantic alignment with the user's factual context and legal intent; (2) $\text{Legal}(r)$: Compliance with applicable Chinese legal provisions; (3) $\text{Express}(r)$: Clear, accurate, and professional expression.

Most existing systems use an end-to-end generation paradigm. While effective in general NLP tasks, this approach often ignores the structured and step-by-step nature of legal reasoning. To address this gap, we propose a modular and interpretable pipeline inspired by cognitive theories of legal writing. Our method breaks the task into three sequential phases, as shown in Figure 2. These phases share a common semantic context—captured by an

element graph—and are managed by a centralized controller. This design supports iterative refinement and controllable reasoning.

3.2 Phase I: Legal Semantic Graph Construction and Prompt Encoding

This phase transforms user queries into structured legal element graphs that provide a semantic foundation for downstream reasoning. Legal reasoning revolves around identifying key facts, involved entities, and legal relationships.

Motivated by this observation, we design an *Element Agent* to construct a graph-based representation G that explicitly encodes legal entities, events, and their semantic connections.

This design is grounded in jurisprudential theories that treat law as a structured system of subjects, facts, norms, and relationships. It aligns with Hart's distinction between primary and secondary rules (Hart, 2012), and Kelsen's hierarchical model of normative systems (Kelsen, 1967).

As illustrated in Figure 2, we define the legal element graph as $G = (V, E)$, where V is the set of nodes and E the set of edges. The node set V includes: (1) Entities: individuals or organizations (e.g., plaintiffs), annotated with attributes such as roles, statuses, and timestamps; (2) Events: legal actions or disputes; (3) User claims, key facts, and inferred legal questions.

Edges E represent semantic relations among elements, such as kinship and contractual duties.

The graph G is serialized in JSON and serves as a global contextual abstraction. This structure enhances both the interpretability and controllability of downstream reasoning modules. An example is provided in Appendix C.

To prepare the initial generation, we serialize the element graph G into a prompt P_G, and concatenate it with the user query q to form the input for generation: $u = [P_G; q]$, This structured input provides rich semantic grounding for downstream response generation.

3.3 Phase II: Multi-agent Iterative Optimization

This phase focuses on refining the initially generated legal draft through a multi-agent framework that supports structured, iterative optimization.

Given the semantic input $u = [P_G; q]$ prepared in the previous phase, we obtain an initial response $r_0 = f(u)$ using a *Draft Agent*. This preliminary output is then subject to refinement via downstream agents coordinated by a centralized controller called the *Manager Agent*. It follows the modular control paradigm in multi-agent systems, where a centralized planner selectively activates agents based on intermediate output quality (Russell and Norvig, 2016).

The iterative optimization process emulates professional legal writing workflows, where drafts are collaboratively refined by experts with distinct roles. In our system, *Manager Agent* oversees the process by dynamically assessing whether the current draft requires further improvement. Its decisions are guided by two primary criteria: (1) linguistic adequacy, including clarity and conciseness; and (2) legal completeness, particularly the inclusion of authoritative statutory references.

If deficiencies in structure or expression are detected, the *Format Check Agent* is invoked to generate targeted revision suggestions. These suggestions are then integrated by the *Draft Agent* in the subsequent iteration. If legal references are missing or insufficient, the *Law Search Agent* will retrieve relevant provisions from statutory databases.

This multi-agent refinement loop proceeds for up to five iterations or until the *Manager Agent* returns a "Pass" signal. The modular design ensures that only necessary sub-agents are activated, maintaining both efficiency and controllability.

The refinement leverages an iterative feedback loop similar to recent approaches in planning-based generation and multi-pass text optimization (Zhang et al., 2020). This design allows for gradual correction and quality enhancement until the response meets legal, factual, and stylistic requirements.The full procedure is shown in Algorithm 1.

3.4 Phase III: Content Revision

This phase aims to refine the draft into a coherent, professional, and legally structured output.

After all iterative revisions, a *Content check Agent* is responsible for rewriting the legal opinion to meet professional norms. The agent receives the original user query and the revised draft as input. It then performs a final pass focused on language quality and output structuring, without introducing new legal content or altering legal positions.

Specifically, the *Content Check Agent* preserves the original legal meaning, improves clarity, precision, and conciseness, and outputs the result in a dual-section format.

The finalized output is partitioned as follows:

- Response: a concise and readable advisory opinion tailored to the user's question;
- Legal Basis: authoritative statutory references that support the conclusion, including full legal article content.

This post-processing step enhances the final output's legal readability and user trust, ensuring it aligns with real-world standards for legal communication and accountability.

3.5 Test Set Correction

This section describes our efforts to ensure rigorous evaluation and high-quality supervision by correcting existing benchmarks and constructing a new dataset for preference-based training. For evaluation, we adopt the widely used LawBench dataset (Fei et al., 2024), which provides question-answer pairs sourced from real-world consultation platforms. However, upon detailed inspection, we identified a substantial number of flawed responses in the test set. These included incorrect legal conclusions, irrelevant or misleading content, and misinterpretations of statutory provisions—issues that severely compromise evaluation reliability.

To address this, we performed a correction of the test set using a hybrid process combining LLM assistance with human expert verification. The LLM was first used to flag potential legal inaccuracies and generate preliminary revision suggestions. These were then reviewed line-by-line by legal professionals to ensure both linguistic clarity and legal correctness. All changes and their justifications are documented in Appendix. We acknowledge the foundational value of LawBench and clarify that our modifications are intended solely to improve the fairness and consistency of evaluation.

3.6 Construction of dataset and DPO Training

In addition, we construct a large-scale dataset, JURISCQAD, comprising over 43000 Chinese legal consultation instances derived from real user queries. To train the model's ability to discern legally sound responses from subtly flawed ones, we generate challenging negative examples using a large language model. To ensure the negative samples are meaningful rather than trivial, we follow a controlled prompt design (see Appendix) that instructs the model to introduce subtle flaws—such as incorrect statutory references, flawed legal reasoning, or conclusions lacking legal support—while maintaining fluency and professionalism. All responses are reviewed by licensed legal professionals to ensure legal accuracy, consistency, and stylistic quality. Each training instance is then organized into a triplet (q, y^+, y^-), where q denotes the user query, y^+ is the expert-approved positive response, and y^- is the model-generated negative response.

Dataset	QA Pairs	QLength	ALength (Pos./Neg.)
# Train	39,163	15.11	263.5 / 193.7
# Val	2,176	15.57	285.7 / 211.3
# Test	2,176	15.22	270.8 / 197.5

Table 2: Statistics of JURISCQAD.

The dataset is curated to avoid trivial distractors, such as grammatical errors, misspelled law names, or absurd contradictions. Additionally, we control for potential bias leakage by restricting the model from introducing demographic stereotypes or fabricated statutes.

Compared to existing legal QA datasets such as LawGPT (Zhou et al., 2024), LawBench (Fei et al., 2024), and JEC-QA (Zhong et al., 2020b), JURISCQAD offers three key advantages: (1) all queries are sourced post-2021 Civil Code, ensuring up-to-date legal references; (2) it provides higher annotation quality despite being smaller than LawGPT, as shown in table 5; and (3) it uniquely includes expert-verified contrastive supervision with both positive and adversarial responses. See Appendix for a detailed comparison.

We adopt the Direct Preference Optimization (DPO) to train our model on JURISCQAD. Each training instance consists of a user query x, a preferred response y^+, and a dispreferred response y^-. The objective is to maximize the log-likelihood difference between the two responses:

$$\Delta_\theta(x, y^+, y^-) = \log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x), \tag{1}$$

$$\mathcal{P}_\theta(x, y^+, y^-) = \sigma\left(\beta \cdot \Delta_\theta(x, y^+, y^-)\right), \tag{2}$$

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\log \mathcal{P}_\theta(x, y^+, y^-)\right], \tag{3}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function and β is a temperature scaling parameter. This training paradigm enables the model to distinguish legally sound responses from plausible yet flawed ones, improving its ability to generate accurate, well-grounded, and formulated legal advice.

Models	Rouge (%)			Bleu (%)			BertScore (%)	Bleurt (%)	LLM Score
	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	3.36
Qwen3-14b	42.55	19.24	27.27	21.93	10.25	8.10	74.64	**62.38**	3.43
Legal LLM									
Chatlaw-33b	27.78	7.40	18.15	17.84	3.81	1.46	67.77	57.54	2.30
Fuzi-mingcha	32.31	9.92	17.41	23.13	7.07	5.19	70.47	52.46	2.78
Hanfei	30.79	10.21	18.37	13.47	4.26	2.68	69.68	58.37	2.69
Lawgpt	20.52	5.22	7.26	5.15	1.01	0.49	63.44	46.94	1.52
Lawyer-llama	30.61	9.50	18.82	24.65	6.28	3.35	69.17	58.47	2.74
Lexilaw	31.23	9.61	18.15	14.40	4.76	3.33	70.00	58.12	2.66
Wisdom	36.60	15.74	23.12	34.89	10.16	8.04	71.97	56.09	2.86
Our Method									
JurisMA	**44.68**†	**23.42**†	**31.14**†	32.54	**16.18**†	**14.25**†	**75.05**	58.63	**3.93**

Table 3: Main results on the revised LawBench. "†" 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	Legal Soundness	Reasoning	Completeness
GPT4o	3.54	3.38	2.88
Lawyer-llama	2.72	2.64	2.86
JurisMA	**4.14**	**3.58**	**3.66**

Table 4: Human evaluation results on three key criteria.

4 Experiment Details

4.1 Experimental Setup

Dataset and Metrics. To assess model effectiveness in legal consultation, we evaluate on a revised version of LawBench (Fei et al., 2024), a widely adopted Chinese legal QA benchmark. As discussed in Section 3.5, We corrected Lawbench through a hybrid process of model-assisted detection and expert validation.

We report standard text generation metrics, including ROUGE-1/2/L (Lin, 2004), BLEU-1/2/N (Papineni et al., 2002), BERTScore (Zhang et al., 2019), and BLEURT (Sellam et al., 2020), computed using official or standard implementations with defaults. Additionally, we include LLM score (GPT-4o) and human validation to assess legal soundness, reasoning, and completeness.

Baselines. We compare our method against strong baselines in two groups:

- **General-purpose LLMs**: GPT-4o (Hurst et al., 2024), Qwen3-14B (Yang et al., 2025). These represent state-of-the-art instruction-following LLMs.
- **Legal-domain LLMs**: 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., 2024), and Wisdom-Interrogatory (Wu et al., 2024). These models are specifically fine-tuned on Chinese legal corpora or tasks.

All models are evaluated in a zero-shot setting using a unified prompt schema. Results are averaged over 5 runs with different random seeds to ensure robustness. Details on model configuration, training hyperparameters, and inference setups are provided in Appendix.

4.2 Main Results

Table 3 presents our main evaluation results. We summarize the key findings below:

(1) Our method achieves the best overall performance across most metrics. Compared to both general-purpose and legal-specialized LLMs, our approach consistently outperforms all baselines on ROUGE, BLEU-2/N, BERTSCORE and LLMSCORE, and ranks second on BLEURT. This demonstrates the effectiveness of structured legal modeling and agent-based reasoning in enhancing both factual accuracy and legal alignment.

(2) Legal-domain LLMs show varied performance depending on their training corpora and alignment strategies. Models like WISDOM and HANFEI achieve competitive scores, while others lag behind. These differences likely stem from variation in corpus quality, and the degree of alignment with user-facing tasks.

(3) Our method excels in both lexical precision and semantic fidelity. Strong results on BERTSCORE, BLEURT, and LLM SCORE (GPT-4o), along with human validation, demonstrate the model's ability to produce accurate and contextually sound legal responses.

4.3 Dataset Evaluation

To assess dataset effectiveness, we fine-tune Qwen2.5 models (3B/7B/14B) via DPO and report pre-/post-tuning results in Figure 3. We also compare our dataset with LawGPT's corpus by fine-tuning Qwen2.5-3B on both, results in Table 5.

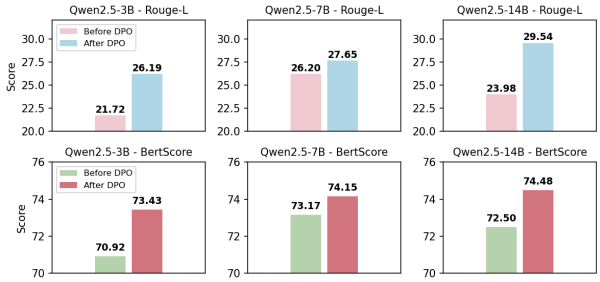

Figure 3: Rouge-L and BertScore comparison before and after DPO across Qwen2.5 models (3B/7B/14B).

Model	Rouge-L	Change	BertScore	Change
Qwen2.5-3B (Base)	21.72	–	70.92	–
+ JurisCQAD	26.19	+4.47	73.43	+2.51
+ LawGPT	20.50	–1.22	68.90	–2.02

Table 5: Performance comparison of Qwen2.5-3B fine-tuned on different training corpora.

(1) Fine-tuning on JURISCQAD leads to consistent performance gains across all model sizes. All models show notable improvements in both lexical and semantic metrics. For instance, Qwen2.5-3B improves by +4.47 on ROUGE-L and +2.51 on BERTSCORE, while Qwen2.5-14B achieves gains of +5.56 and +1.98.

(2) JURISCQAD exhibits high annotation quality and practical effectiveness. As Qwen2.5 models already encode rich Chinese legal knowledge, outdated or noisy data may hinder alignment. In contrast, our curated corpus yields consistent improvements, confirming its superior quality.

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. Table 6 presents results for Qwen2.5-7B and 14B, showing that all modules contribute meaningfully to performance.

Table 6: 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**

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

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

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

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 LegalBench dataset, which contains factually complex, rule-based legal questions. We randomly sample 50 instances and perform evaluation across 5 different random seeds to obtain averaged results. As detailed in Appendix, 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 consultation: whether a man's drunk driving incident

Question	我在2011年有过一次醉酒驾驶，***那时候醉驾还没有入刑***，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***. 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。.....法律依据：**1.《中华人民共和国刑法修正案（八）》**：.....。**2.《中华人民共和国道路交通安全法》**：.....。**3. 各航空类学校的招生条例**：.....。 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.** .....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**，..... Answer:According to **Article 99 of the Road Traffic Safety Law of the People's Republic of China**, ...
JurisMA	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**：.....”。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:**".....”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 4: Case study comparing model-generated responses to a time-sensitive legal query. The example illustrates differences in factual interpretation, statutory grounding, and reasoning structure. To aid analysis, we use the following inline annotations: **bold text** for key legal focus points, underlined text for the conclusion of the response, light red highlighting for cited statutes, and light green highlighting for underlying legal principles.

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.

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 transforming complex legal queries into structured semantic graphs and coordinating specialized agents via a centralized Manager Agent, our system enables controllable, interpretable, and legally grounded reasoning.

We also introduce JurisCQAD, a dataset of over 43,000 expert-validated Chinese legal consultations. It supports preference-aligned DPO training and enables robust evaluation across diverse scenarios.

Experiments on the corrected LawBench benchmark show that JurisMA outperforms both general-purpose and legal-domain LLMs across multiple lexical and semantic metrics. Ablation studies highlight the importance of structured prompting and dynamic agent routing, while case analyses illustrate improvements in legal consistency and statutory reference quality.

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 exhibit bias toward high-frequency consultation topics, limiting generalization to rare or highly specialized cases. Finally, while we incorporate both LLM-based and human evaluations to ensure robustness, broader real-world deployment and continuous user feedback are still needed to assess long-term reliability and practical usability.

Ethics Statement

All data used in this study were derived from publicly available Chinese legal consultation forums. Personally identifiable information was removed through automatic anonymization and manual screening before model training.

The expert annotators involved in dataset creation were all members of our research team, comprising qualified legal professionals. One team member holds a Bachelor's degree in Law and supervised the annotation process. All annotators strictly followed the current laws and regulations of the People's Republic of China during data review and correction. They also received written guidelines covering correctness criteria, bias avoidance, and confidentiality requirements.

During data curation, annotators were instructed not to fabricate legal provisions and to flag any sensitive or ethically questionable content for removal. No private, confidential, or client-specific information was retained. The resulting JurisCQAD dataset is released solely for non-commercial research under an academic license.

References

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. 2024. Lexilaw: A scalable legal language model for comprehensive legal understanding.

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. 2020a. 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.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jecqa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

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 1.

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
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

Table 7: Summary of JurisCQAD Dataset

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.

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

Table 8: Comparison between Legal QA and Legal Consultation QA (Legal CQA)

In the JURISMA, we employ Qwen2.5-14B-Instruct, fine-tuned on the JurisCQAD dataset, 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.

Algorithm 1: Multi-Agent Controlled Draft Optimization

Input: User query q; element graph G; initial draft $r_0 = f([P_G; q])$
Initialize: $t \leftarrow 0$, $r_t \leftarrow r_0$
Output: Final legal response r_{final}

while $t < T$ **do**
  $a_t \leftarrow \text{ManagerAgent}(r_t)$
  if $a_t = \emptyset$ **or** $a_t =$ "Pass" **then**
    break {Stop if draft is acceptable}
  end if
  if "FormatCheck" $\in a_t$ **then**
    $s_t \leftarrow \text{FormatCheckAgent.getSuggestions}(r_t)$
    $r_t \leftarrow \text{DraftAgent.applySuggestions}(r_t, s_t)$
  end if
  if "LawSearch" $\in a_t$ **then**
    $l_t \leftarrow \text{LawSearchAgent.retrieveStatutes}(q, r_t)$
    $r_t \leftarrow \text{DraftAgent.integrateStatutes}(r_t, l_t)$
  end if
  $t \leftarrow t + 1$
end while
$r_{\text{final}} \leftarrow \text{ContentCheckAgent}(r_t)$
return r_{final}

A.3 Prompt

B Legal QA & Legal CQA Comparison

Table 2 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.

C Example of Element Graph

Table 10 presents each node and its detailed attributes in the element graph extracted from the case study 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 11 summarizes representative examples and reasons for revision.

E Generalization study

Table 12 shows generalization results on the LEGALBENCH-RuleQA subset (50 samples, averaged over 5 seeds).

F Multi-Agent Controlled Draft Optimization Algorithm

The detailed optimization process is illustrated in Algorithm 1, which outlines how the *Manager Agent* coordinates auxiliary agents (*FormatCheck* and *LawSearch*) through iterative refinement.

Table 9: Prompts used by different agents.

Agent	Prompt Summary
NEGATIVE EXAMPLE GENERATION AGENT	Please generate a plausible but flawed response to the following legal consultation question. The response should appear reasonable on the surface but contain either incorrect legal references, flawed logical reasoning, or lack critical statutory support. **Requirements:** • The response must be grammatically fluent and professionally styled; • No social biases are allowed (e.g., based on gender, age, ethnicity, or region); • No fabricated or non-existent statute numbers are permitted; • The flaw should be mild—such as subtle legal or factual inaccuracies—not absurd or completely irrelevant; **Instruction:** Answer the following question. First, provide the flawed response, then specify the legal basis.
DRAFT AGENT	You are a legal consultation draft generation agent. When a user submits a legal-related question, your task is to generate a professional response based on existing legal knowledge.

ELEMENT AGENT	You are a professional legal element extraction expert. Your tasks include:
	• Extract key case elements from the user's legal consultation;
	• Identify legal relationships and entities;
	• Clarify the user's legal demands;
	• Output a structured element graph in JSON format.
	Element Graph Format:
	`{`
	`  "entities": [{ "name": "...", "type": "...", "attributes": {...} }],`
	`  "events": [{ "description": "...", "time": "..." }],`
	`  "relationships": [{ "type": "...", "source": ..., "target": ... }],`
	`  "user_claims": ["..."],`
	`  "key_facts": ["..."],`
	`  "legal_questions": ["..."]`
	`}`

MANAGER AGENT	You are the decision-making agent in a multi-agent legal consultation system. Your task is to determine, based on the content of a draft legal response, whether it requires format refinement or legal citation supplementation. **Decision Criteria:** • If the response is not concise, lacks clear logic, or contains redundancy: `Call: FormatCheckAgent`; • If the response lacks statutory references: `Call: LawSearchAgent`; • If both issues apply: `Call: FormatCheckAgent` then `LawSearchAgent`; • If the response is acceptable: `Pass`
FORMATCHECK AGENT	Review the draft for clarity, redundancy, and stylistic issues. Output concrete editing suggestions without changing the legal meaning.
LAWSEARCH AGENT	Retrieve authoritative legal provisions from Chinese law based on the question and draft response. Output only relevant statute text.
CONTENTCHECK AGENT	Rewrite the draft into a fluent, professional legal opinion. Preserve meaning while fusing reasoning and statute into a dual-structured final output.

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 10: Example of Element Graph

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 11: Representative Examples of Test Set Corrections

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**†

Table 12: 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.