MLCQA: A Case Retrieval-Based Question Answering System for Macao Law

Io Cheng Tong[1], Zhe Li[1†], Yuhong Zhang[1†], Hong Io Chu[1], Chuxin Ouyang[1], Ang Li[2†]

[1]University of Macau, China, [2]Zhejiang University, China,
{ictong, zheli, yc37229, yc47254, yc37216}@um.edu.mo, leeyon@zju.edu.cn

Abstract

Despite their advancements in legal AI, large language models (LLMs) continue to struggle with processing court judgments from jurisdictions marked by historical legal pluralism, such as Macao. The rigid translated legal terms, absence of unified structure and complex but varied legal reasoning styles, leading to model hallucinations and comprehension difficulties. In this paper, we introduce the Macao Legal Case-based Question Answering (MLCQA) system, a novel case retrieval augmented generation (RAG) system tailored to this unique legal environment. MLCQA transforms unstructured judgments into structured fields using a hybrid extraction pipeline that combines LLM parsing with regex rules. The LLM is guided by a Legal Syllogism prompt to induce expert-style reasoning, enabling the reconstruction of a clear chain linking legal provisions, judges' interpretation, factual circumstances, and verdict. Unlike standard legal RAG systems that operate on full cases, MLCQA selects different field combinations to drive a multi-stage pipeline for retrieval, reranking, and answer generation, reflecting how legal experts focus on different information at different procedural stages. The system also integrates built-in citations that link answers directly to the referenced legal provisions or precedents. Evaluation shows that MLCQA achieves substantial gains in accuracy, terminology use, and clarity, demonstrating that integrating structured legal knowledge can deliver strong performance.

1 Introduction

Artificial Intelligence and Law (AI & Law) is a research field that is gaining increasing importance. In the early stages, related research mainly relied on rule-based expert systems and traditional machine learning models(Bench-Capon et al. 2012). Recently, LLMs have demonstrated excellent performance in various legal tasks due to its outstanding language understanding and text generation capabilities, significant improvements have been shown in the legal field (Ashley 2017). However, the limitations of LLMs are undeniable, especially the hallucination problem and explainability, which pose particularly acute challenges in the legal profession, as

* These authors contributed equally.
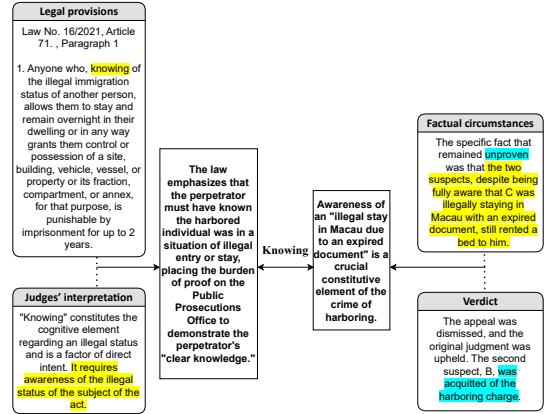† Corresponding authors.

Figure 1: Partial field content extracted by the model from Case No.409/2025 (Criminal Appeal) based on prompts

the field's fundamental requirements for precision and transparency render them especially problematic. These issues could significantly impede professionals' adoption and application of such technologies(Dahl et al. 2024).

To effectively address these challenges, a widely adopted strategy is to enhance large models' legal question-answering capabilities by retrieving specialized legal knowledge. This approach provides more precise, reliable, and interpretable knowledge support, thereby improving its practicality and credibility in legal applications (Lewis et al. 2020). Retrieval Augmented generation (RAG) addresses the issue of LLMs' "hallucinations" by enhancing capabilities in critical domains and improving interpretability through retrieval from external knowledge bases(Hindi et al. 2025). Existing legal applications, such as DRAG-BILQA (Zhang et al. 2024) and CBR-RAG(Wiratunga et al. 2024), introduce legal factor recognition modules or focus on Case-Based Reasoning (CBR) to ensure the recognition. However, shaped by its distinctive colonial legacy and post-handover developments, Macao's judicial documents embody a unique legal amalgamation—integrating Portuguese legal doctrines with the Macao Basic Law and legislation enacted by the Macao SAR since 1999. This hybrid system presents several key challenges for LLM-based analy-

sis. First, Chinese legal terminology is often rigidly translated from Portuguese, carrying nuanced meanings that extend beyond literal interpretation and demanding precise legal definitions for accurate comprehension. Second, the absence of a uniform document format leads to disorganized and inconsistent information structures. The judicial reasoning varies significantly across judges' writing styles, further complicating automated interpretation. This intricate fusion of linguistic complexity and legal traditions not only heightens the risk of model hallucinations and comprehension errors but also renders conventional RAG architectures inadequate, necessitating a purpose-built framework tailored to Macao's unique legal and linguistic landscape.

We have identified a logical pathway within these complex documents that allows AI to decipher Macao's unique legal amalgamation. Macao's judicial documents serve as exemplary models of professional and high-quality legal knowledge. These documents establish a complete and rigorous logical chain: "legal provisions—judges' interpretation—factual circumstances—verdict." The obligation of judges to provide legal reasoning in their rulings requires them to comprehensively present and explain the process and rationale behind their decisions. Therefore, judges fully document in judicial documents their determinations on disputed facts, the reasons for applying specific legal provisions, including basic textual interpretation of the applied laws, and the mapping between constituent elements and specific factual circumstances, as well as distinctions made from other relevant legal provisions. As shown in Fig. 1, using the "knowing" element in the crime of harboring as an example, the judge's interpretation clarifies how factual findings correspond to statutory requirements. It illustrates how the legal provision of "illegal immigration status" connects to the specific case fact of "illegally staying in Macau with an expired document," thereby revealing that the "knowing" requirement is met specifically by the awareness of the document's expired status. Macao's judicial documents not only demonstrate the process of legal application but also contain in-depth analysis of legal texts, including precise textual interpretation, rigorous logical reasoning, and substantive legal reasoning. These elements form a concrete interpretation of Macau's complex legal traditions and provide a reliable foundation for understanding how rules operate in practice. Ignoring or insufficiently engaging with this adjudicative chain can lead to gaps in legal understanding, resulting in misinterpretation of the logic behind legal rules and potentially inaccurate retrieval or reasoning outcomes. This raises a key question: *how can we systematically extract such information and enable LLMs to fully utilize it?*

Building on this insight, we developed the Macao Legal Case-based Question Answering System (MLCQA). Specifically, MLCQA first converts unstructured Macao court judgments into structured fields using a hybrid extraction pipeline that combines LLM parsing with targeted regex rules. The LLM is guided by a Legal Syllogism prompt, simulating legal experts' reasoning to explicitly map legal provisions, judges' interpretation, factual circumstances, and verdict. Using Macao's judgments as data sources, we precisely extracted 20 core fields to construct a high-quality structured

dataset, forming a solid foundation for downstream retrieval and reasoning. To retrieve relevant legal cases, the system employs a combination of text-based and vector-based methods. In the reranking stage, a cross-encoder model performs deep interaction between the query and legal case to produce relevance scores. Additional weights based on legal significance are then applied to further adjust the ranking. Finally, the system inputs specific fields of reranked legal cases into the large model for comprehensive analysis and synthesis, generating final responses tailored to user queries. It is worth mentioning that the generation process incorporates built-in reference technology to ensure that every answer is verifiable and traceable, greatly enhancing the credibility and persuasiveness of the answer.

Experiments demonstrate that MLCQA substantially improves accuracy, terminology usage, and clarity compared with general-purpose models, highlighting the effectiveness of integrating structured legal knowledge in handling Macao's complex judicial documents. The key contributions of this paper include (1) as the first legal Q&A platform specifically designed for Macao's legal system, it deeply integrates LLMs and RAG technologies to provide robust intelligent support for Macao's legal community; and (2) the system proposes a structured prompt method simulating legal professionals' thinking patterns, along with a multi-stage field information utilization strategy, significantly optimizing the RAG framework to achieve dual improvements in precision retrieval and high-quality responses. It also features built-in citation functionality to enhance answer traceability; and (3) through rigorous testing, our approach demonstrates outstanding performance in three critical dimensions: accuracy, terminology, and clarity. Since its launch, it has attracted 882 visits and over 7042 calls, fully demonstrating its practical value and widespread recognition.

2 Overview

Fig. 2 illustrates the development process of MLCQA, which consists of three primary modules designed to achieve three key objectives: 1) Data Munging: Expert-Level Legal Information Extraction, 2) RAG system: Ensuring Relevance and Authority in Case Retrieval, 3) Answer and Source: Interpretable Answer Generation with Traceability.

2.1 Expert-Level Legal Information Extraction

We develop a legal information extraction method that identifies and classifies fine-grained legal elements from court decisions. Inspired by how legal professionals analyze judgments, we design a prompt-guided extraction procedure combining rule-based signals with LLMs reasoning, ensuring that the extracted content captures both explicit textual markers and implicit legal logic. Let a raw legal document be denoted as D. We predefine an extraction field set:

$$\mathcal{F} = \{\text{factual circumstances, judge explanation, legal provisions, } \ldots\}.$$

(1)

Regex-Based Extraction. For structured or well-patterned fields $f \in \mathcal{F}$, we apply a domain-specific regex

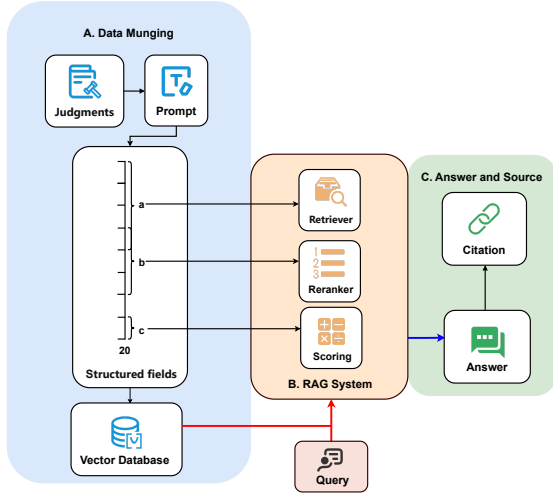Figure 2: Development process of MLCQA

pattern $\mathcal{R}_f$ to obtain candidate spans:

$$x_f = \{\, s \subset D \mid s \text{ matches } \mathcal{R}_f \,\}. \tag{2}$$

Example for legal provisions:

$$\mathcal{R}_{\text{law\_citation}} = \texttt{Article \textbackslash d+(?:\textbackslash(\textbackslash d+\textbackslash))?} \\ \texttt{of [A-Za-z]+ Law.} \tag{3}$$

LLM-Guided Extraction. For dispersed or reasoning-intensive fields $f \in \mathcal{F}$, we also apply an LLM with a structured prompt $\mathcal{P}_f$ to extract content that requires reasoning beyond regex patterns:

$$x_f = \text{LLM}(D, \mathcal{P}_f), \tag{4}$$

For example, for the `verdict` field, the LLM prompt instructs extraction in a syllogistic structure: *Prompt (verdict): summarize the judge's reasoning using three steps: (1) factual circumstances (minor premise) from verified facts; (2) legal basis (major premise) from relevant statutes; (3) conclusion combining facts and law to derive the judgment ...*

Using this pipeline, we process nearly 16,000 Macao's public court judgments into a structured database containing 20 expert-designed fields, enabling fine-grained downstream retrieval and reasoning.

2.2 Ensuring Relevance and Authority in Case Retrieval

Given a user query q, our retrieval system performs multi-stage filtering using the structured fields extracted in Section 2.1.

Vector Retrieval with Retrieval Fields. We construct the retrieval representation of each document by concatenating its selected retrieval fields:

$$r_d = \text{concat}(x_f : f \in \mathcal{F}_{\text{retrieval}}). \tag{5}$$

A bi-encoder $E(\cdot)$ encodes both the query and document, then similarity is computed using cosine similarity:

$$\mathbf{h}_q = E(q), \mathbf{h}_d = E(r_d), s_{\text{vec}}(q,d) = \frac{\mathbf{h}_q \cdot \mathbf{h}_d}{\|\mathbf{h}_q\| \, \|\mathbf{h}_d\|}. \tag{6}$$

The top-K candidate documents $\mathcal{C}_K$ with high similarity are selected.

Cross-Encoder Reranking with Rerank Fields. For each candidate $d \in \mathcal{C}_K$, we construct a reranking representation:

$$c_d = \text{concat}(x_f : f \in \mathcal{F}_{\text{rerank}}). \tag{7}$$

A cross-encoder f_θ takes (q, c_d) as input and compute reranking score:

$$s_{\text{ce}}(q,d) = f_\theta(q, c_d). \tag{8}$$

Rule-Driven Importance Scoring. Macao's judicial practice includes explicit importance rules (court hierarchy, publication type, sentencing implications, etc.). We encode these using indicator or numeric features $\phi_i(d)$:

$$\text{imp}(d) = \sum_i w_i \cdot \phi_i(d), \tag{9}$$

where w_i are weights representing legal-domain priorities. Final retrieval score can be obtained as follows:

$$s_{\text{final}}(q,d) = \alpha \, s_{\text{ce}}(q,d) + (1-\alpha) \, \text{imp}(d). \tag{10}$$

We return the reranked documents $\tilde{\mathcal{C}}_K$.

2.3 Interpretable Answer Generation with Traceability

We only expose the *generation fields* of reranked cases to the LLM during the answer generation stage:

$$g_d = \text{concat}(x_f : f \in \mathcal{F}_{\text{generation}}), \quad d \in \tilde{\mathcal{C}}_K. \tag{11}$$

Then we employ the legal expert LLM to perform case-based legal reasoning and output the answer:

$$a = \text{LLM}(q, \{g_d\}_{d \in \tilde{\mathcal{C}}_K}). \tag{12}$$

The citation mechanism provides traceability from the generated text back to the underlying statutes, judicial reasoning, and extracted fields, ensuring professional-grade reliability in legal contexts.

3 Demonstration

This demonstration describes a case retrieval-based question-answering system. It utilizes the techniques from Section 2. to ensure that every answer is linked to the original source material, with citations provided for verification. The three main functions are illustrated in Fig. 3.

Legal Understanding and Analysis: Our system enables precise identification of user intent and in-depth legal analysis. When a user inputs a question, the system first employs natural language processing technology to interpret both surface-level semantics and underlying legal intent. For instance, when handling extreme questions like "How can I kill someone and receive the minimum sentence?", the system not only identifies key legal elements such as "intentional homicide" and "sentencing factors," but also proactively assesses potential legal violations and ethical risks, providing a warning prompt: "Your question involves intentional homicide and how to seek the minimum penalty. Such
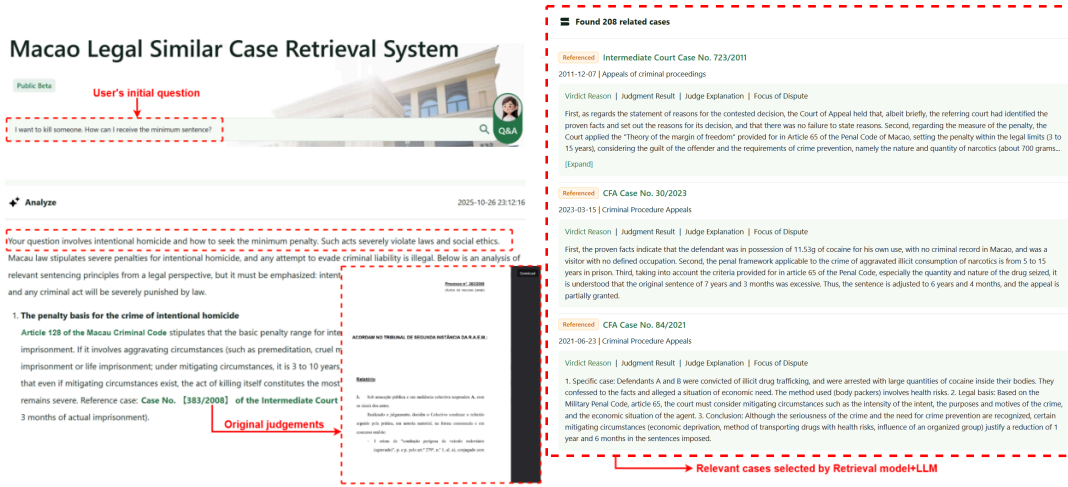
Figure 3: System Demo

Model	Accuracy (S^L/S^H)	Term precision (S^L/S^H)	Clarity (S^L/S^H)	Alignment Ratio
MLCQA	3.92 / 3.4	3.97 / 3.9	3.76 / 3.7	94.6%
GPT4	3.80 / 3.1	3.88 / 3.2	3.75 / 3.8	88.1%
Qwen-turbo	3.73 / 3.0	3.73 / 3.1	3.58 / 3.3	86.3%

Table 1: Results of evaluation

acts severely violate laws and social ethics." Additionally, considering that public users often use non-professional expressions, the system employs a semantic model enhanced with legal knowledge to deduce core legal issues from everyday, vague descriptions. For example, when a user asks, "My uncle gave me an old house before he passed away, and I've lived in it for thirty years. Is it mine?" The system identifies the legal need as "how to legally obtain property rights," thereby transforming "everyday questions" into "legal needs." After clarifying the problem, the system will intelligently associate and cite relevant authoritative precedents and legal provisions, conduct preliminary analysis and reasoning, and provide users with professional and educational reference answers.

Legal References and Citation Integration: To enhance the reliability and transparency of responses, the system incorporates interactive citation links. Users can directly click on referenced legal provisions or case names in the answers, which will instantly display the full text of the cited provisions or redirect to the original judicial documents. This feature ensures that every response in the system is verifiable and legally substantiated.

Case Recommendation: After addressing users' specific queries, the system identifies their underlying information needs and proactively suggests relevant public cases. These recommendations aren't just random listings—they're carefully curated, with each case highlighting key dimensions users care about, presented in a structured format. Current categories include Verdict Reason, Judgment Results, Judge Explanation, and Focus of Dispute.

4 Evaluation

We compiled 40 real exam questions and answers from the Law School of the University of Macau, employing Farui Plus, the legal expert LLM, as the evaluation model and conducting human evaluation to comprehensively compare the performance of MLCQA, Qwen, and Deepseek on these legal questions. The evaluation was conducted across three dimensions: accuracy, coherence, and clarity. "Accuracy" measures the degree to which the system provides correct or valid responses to legal questions; "term precision" assesses the relevance and professionalism of legal terminology used by the system; and "clarity" evaluates the logical consistency and readability of the system's output. The models were scored on a 1-5 scale (1 being the lowest and 5 the highest) across these dimensions. We measure human–LLM score alignment as the normalized mean absolute deviation between human scores and LLM-generated scores across three evaluation dimensions.

$$\text{Alignment} = 1 - \frac{1}{3}\sum_{i=1}^{3}\frac{|S_i^L - S_i^H|}{4}. \qquad (13)$$

S_i^L and S_i^H denote the human and LLM-generated scores for the i-th evaluation dimension, with the denominator 4 normalizing the 1-5 Likert scale difference. Detailed survey results are presented in Table 1, which shows that our system significantly outperformed the other comparison models across all evaluation dimensions.

5   Conclusion

This paper presents a legal Q&A system developed using structured databases and RAG technology. By leveraging our data processing methodology, we effectively utilize the strong logical reasoning inherent in Macao's judicial documents to extract high-quality legal knowledge that significantly enhances the model's response capabilities. Through simulating legal professionals' thought processes to guide RAG invocation logic, we improve the model's matching accuracy. Our research demonstrates that the model trained on 16,000 Macao judicial documents outperforms general-purpose models in specific legal queries. Moving forward, we plan to integrate additional legal databases and refine the model's performance by constructing precise knowledge graphs.

6   Acknowledgements

References

Ashley, K. D. 2017. *COMPUTATIONAL MODELS OF LEGAL REASONING*, 1–2. Cambridge University Press.

Bench-Capon, T.; Araszkiewicz, M.; Ashley, K.; Atkinson, K.; Bex, F.; Borges, F.; Bourcier, D.; Bourgine, P.; Conrad, J. G.; Francesconi, E.; et al. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*, 20(3): 215–319.

Dahl, M.; Magesh, V.; Suzgun, M.; and Ho, D. E. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1): 64–93.

Hindi, M.; Mohammed, L.; Maaz, O.; and Alwarafy, A. 2025. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access*, 13: 46171–46189.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Wiratunga, N.; Abeyratne, R.; Jayawardena, L.; Martin, K.; Massie, S.; Nkisi-Orji, I.; Weerasinghe, R.; Liret, A.; and Fleisch, B. 2024. CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In *International Conference on Case-Based Reasoning*, 445–460. Springer.

Zhang, Y.; Li, D.; Peng, G.; Guo, S.; Dou, Y.; and Yi, R. 2024. A Dynamic Retrieval-Augmented Generation Framework for Border Inspection Legal Question Answering. In *2024 International Conference on Asian Language Processing (IALP)*, 372–376.