LeCoDe: A Benchmark Dataset for Interactive Legal Consultation Dialogue Evaluation

Weikang Yuan[1]*, **Kaisong Song**[2,3], **Zhuoren Jiang**[1,4]†, **Junjie Cao**[2], **Yujie Zhang**[1], **Jun Lin**[2], **Kun Kuang**[1], **Ji Zhang**[2], **Xiaozhong Liu**[5]

[1]Zhejiang University, [2]Tongyi Lab, Alibaba Group, [3]Northeastern University

[4]Lab of Inclusive & Smart Governance for Urban-Rural Integration (LISGURI), Zhejiang University

[5]Worcester Polytechnic Institute

{yuanwk, jiangzhuoren, yj.zhang, kunkuang}@zju.edu.cn, {kaisong.sks, junjie.junjiecao, linjun.lj}@alibaba-inc.com, xliu14@wpi.edu

Abstract

Legal consultation is essential for safeguarding individual rights and ensuring access to justice, yet remains costly and inaccessible to many individuals due to the shortage of professionals. While recent advances in Large Language Models (LLMs) offer a promising path toward scalable, low-cost legal assistance, current systems fall short in handling the interactive and knowledge-intensive nature of real-world consultations. To address these challenges, we introduce LeCoDe, a multi-turn benchmark dataset constructed from publicly available real-world legal consultation content and carefully processed into a de-identified, structured research resource for evaluating and advancing research on LLMs in legal consultation settings. LeCoDe contains 3,696 multi-turn consultation cases with 110,008 dialogue turns. The dataset is further enriched through expert annotation, including key facts, fact importance, and advice summaries. Furthermore, we propose a comprehensive evaluation framework that assesses LLMs' consultation capabilities in terms of (1) clarification capability and (2) professional advice quality. This unified framework incorporates 12 metrics across two dimensions. Through extensive experiments on various general and domain-specific LLMs, our results reveal significant challenges in this task, with even state-of-the-art models like GPT-4 achieving only 35.9% recall for clarification and 59.1% overall score for advice quality, highlighting the complexity of professional consultation scenarios. Based on these findings, we further explore several strategies to enhance LLMs' legal consultation abilities. Our benchmark contributes to advancing research in legal domain dialogue systems, particularly in simulating more real-world user-expert interactions. The resource is available at https://github.com/PiLab-ZJU/LeCoDe.

1 Introduction

Expert consultation services play a vital role in providing professional guidance across knowledge-intensive domains such as law (Xie et al., 2024), healthcare (Tu et al., 2025), and mental health (Szymanski et al., 2025). For example, legal consultation services are crucial for safeguarding individual rights and ensuring fairness in society (Rodrigues, 2020). However, the gap between surging legal demands and scarce professional resources has led to prohibitive costs, significantly limiting access to justice for people without domain expertise. The advance of Large Language Models (LLMs) offers new opportunities to enhance the accessibility and convenience of such services (Kirk et al., 2021; Yuan et al., 2026), providing low-cost legal consultation while potentially reducing lawyers' workloads through their extensive knowledge base and conversational capabilities (Lai et al., 2024).

As shown in Figure 1, the interaction between experts and users in consultation scenarios presents complexities that make evaluating LLM-based expert systems particularly challenging[1]. **Asymmetric Expertise**: Users who seek help often lack legal literacy, providing vague initial case descriptions and occasionally omitting critical details. **Interactive Consultation Process**: To clarify the users' needs and information, legal experts usually employ multi-turn questioning strategies to verify facts and extract case-specific nuances. Through this iterative clarification, experts integrate both codified knowledge and practical experience to progressively clarify the situation (*Clarification Capability*) and then deliver professional legal advice (*Advice Quality*).

Current research on LLM-based expert consultation systems reveals limitations. As illustrated in

*This work was done during an internship at Tongyi Lab, Alibaba Group.

†Corresponding author.

[1]While we demonstrate these challenges in the legal domain, similar complexities exist across various consultation scenarios including medical and other professional domains (Fan et al., 2025; Li et al., 2025).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36019–36048

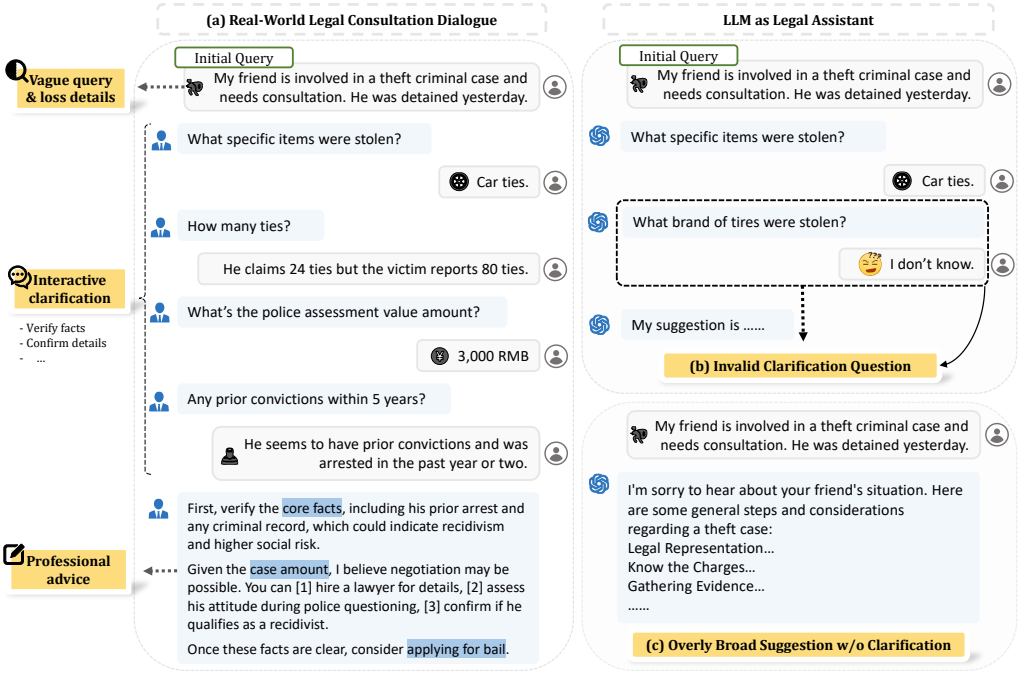

Figure 1: Illustration of real-world legal consultation.

Figure 1(b) and 1(c), existing models often generate invalid clarification questions or provide immediate advice without necessary clarification interactions. These limitations can be attributed to two main challenges: **Limited Availability of Suitable Consultation Data**: Most existing studies rely on single-turn QA pairs (Dai et al., 2025) or synthetic dialogue data generated by LLMs (He et al., 2023). Such datasets inadequately capture the sophisticated dynamics of real-world consultation scenarios, such as users' natural information-seeking behaviors and experts' strategic clarification of missing key information, thus hindering models' ability to learn professional consulting competence. **Insufficient Evaluation Framework**: Previous evaluation approaches primarily focus on the quality of final advice (Wu et al., 2024) or the accuracy of expert judgments (Li et al., 2024b). However, this approach overlooks a key aspect: assessing models' clarification ability, specifically their proficiency in asking purposeful questions before offering advice.

To address these challenges, **first**, we introduce the Legal Consultation Dialogue Dataset (LeCoDe), a large-scale benchmark for interactive legal consultation. LeCoDe is built from publicly available legal consultation content from short-video platforms in China and transformed into de-identified, structured, and annotated data for research use. Through multi-stage processing and annotation, we enrich the dataset with legally substantive annotations, including users' initial queries, atomic key facts, fact importance, and summaries of expert advice. The resulting benchmark contains 3,696 consultation cases with 110,008 dialogue turns.

Second, we propose an interactive legal consultation framework to comprehensively evaluate LLMs' consultation capabilities. This framework enables systematic assessment through simulated user-expert interactions, measuring both **Clarification Capabilities** (in terms of *effectiveness* and *efficiency*) and **Advice Quality** through *automated metrics* and *LLM-based evaluation*. We conduct extensive experiments on both general LLMs and legal domain-specific LLMs to evaluate their performance in legal consultation scenarios. Our findings reveal significant limitations in current LLMs' legal consultation capabilities. In terms of clarification ability, even SOTA LLMs like Qwen-max and GPT-4 achieve only modest recall rates of 39.8% and 35.9%, respectively, indicating that their ability to elicit key facts remains insufficient. Similarly, in advice quality, leading models like Deepseek and GPT-4 achieve overall scores of only 62% and 58%, demonstrating poor performance across professionalism, completeness, and user satisfaction metrics. To effectively leverage the LeCoDe dataset, we propose various strategies for constructing SFT training dialogues. While these approaches improve model performance, a substantial gap re-

mains between current capabilities and the requirements of legal consultation. These insights establish LeCoDe as both a challenging benchmark for evaluating LLMs in legal consultation settings and a valuable resource for future research in legal AI.

Our key contributions are threefold:

(1) **High-Quality Benchmark Dataset for Interactive Legal Consultation**: We introduce LeCoDe, a large-scale benchmark dataset for interactive legal consultation, constructed through multi-stage processing and expert annotation of publicly accessible legal consultation content and released in a carefully processed, de-identified form for research use. LeCoDe provides structured and high-quality data with substantive annotations for evaluation and model training.

(2) **Comprehensive Evaluation Framework**: We establish a comprehensive evaluation framework tailored for legal consultation tasks. Unlike prior methods, our framework systematically evaluates both the clarification interactions and the final quality of professional advice. This setup supports more precise assessment of models' consultation performance.

(3) **In-depth Empirical Analysis and Strategic Insights**: Our extensive experiments reveal substantial performance gaps among state-of-the-art LLMs when evaluated under realistic legal consultation scenarios. We further explore several effective Supervised Fine-Tuning (SFT) strategies that demonstrably improve model performance. These empirical insights offer useful directions for future advancements in Legal AI research.

2 Related Work

Large Language Models (LLMs) have exhibited impressive capabilities and performance across diverse areas (Achiam et al., 2023; Bai et al., 2023; Yuan et al., 2024; Guo et al., 2025), and their interactive nature with users shows significant potential in consultation settings (Li et al., 2025). However, a critical challenge emerges in consultation scenarios: users often provide vague queries, requiring LLMs to possess robust clarification abilities to address this information gap (Fan et al., 2025; Liu et al., 2023; Tu et al., 2025; Wu et al., 2024).

In the legal domain, many existing benchmarks mainly evaluate the foundational judicial reasoning capabilities of LLMs, like legalbench (Guha et al., 2023) and lexeval (Li et al., 2024a). Some recent studies have explored improving LLM performance in legal consultation through supervised fine-tuning (Sun et al., 2024), reinforcement learning (Wu et al., 2024), and multi-agent collaboration (Cui et al., 2023). However, existing works often overlook the interactive nature of legal consultations and lack access to real-world consultation data. CrimeKgAssistant contains 200K real-world lawyer-client QA pairs but is limited to single-turn interactions (Dai et al., 2025). Hanfei (He et al., 2023) provides multi-turn dialogues but relies on synthetic conversations generated by LLMs. CAIL2023 conversational similar case retrieval dataset (ConvIR) (CAIL, 2023) focuses on retrieval task without providing legal advice or real scenarios, and CAIL2024 consultation dialogue generation dataset (ConGen) (CAIL, 2024) utilizes generated rather than real-world data. Beyond the Chinese legal context, Hong et al. (2021) use transcripts of U.S. legal hearings as legal dialogues for a challenging information extraction task.

Datasets	Advice	Clarification	Multi-Turn	Expert Ann.
CrimeKG	✓			
Hanfei	✓	✓	✓	
ConvIR		✓	✓	
ConGen	✓	✓	✓	
LeCoDe	✓	✓	✓	✓

Table 1: Related Datasets Comparison: According to Presence of Legal **Advice**, Inclusion of **Clarifying** Questions and **Multi-Turn** Dialogue Capability, **Expert Ann**otation Status.

As shown in Table 1, LeCoDe differs from previous works in several aspects. It covers both clarification and advice in multi-turn legal consultation, and it is further enriched with expert annotations such as key facts, fact importance, and advice summaries. We also introduce an evaluation framework for legal consultation scenarios that covers two key aspects: Clarification Capability and Advice Quality. These features make LeCoDe a comprehensive benchmark for legal consultation tasks.

3 LeCoDe

In this section, we first define the legal consultation task and then describe the construction and statistics of LeCoDe. We next introduce the evaluation metrics used in our benchmark, and finally present how the dataset can be used to construct training dialogues for supervised fine-tuning.

3.1 Task Definition

The legal consultation task is an interactive process between two agents: a **user client** C and a **legal expert** E. Let $A_n = \{a_1, ..., a_n\}$ denote the set of atomic key facts privately held by C, n is the cardinality of the set. The client C starts the consultation with an initial query u_1^C about a specific case which may be vague and may not fully disclose all relevant facts in A_n.

The consultation dialogue sequence D of length T is defined as $D = \{(u_t^C, u_t^E)\}_{t=1}^T$, where u_t^C and u_t^E represent utterances of client C and expert E at turn t. At each turn, the expert gives a response u_t^E conditioned on: $u_t^E \sim p_E\{\cdot \mid D_{t-1}, u_t^C\}$, where D_{t-1} represents previous dialogue context. The expert's response type r_t belongs to $R = \{\text{Question}, \text{Advice}\}$, indicating whether to ask a clarifying question or provide legal advice. Following the expert's clarifying question, the client's response is defined as $u_{t+1}^C \sim p_C\{\cdot \mid D_t, A_n\}$, where the client generates responses based on the known key facts A_n in response to the expert's latest clarifying question in dialogue context D_t. The consultation terminates when either *the expert provides legal advice* or *maximum turn is reached*.

Simulation Framework: To simulate real-world consultation interactions, the expert agent can be represented by various LLMs and strategies to generate clarifying questions or final advice based on the user's initial query and subsequent feedback. We employ an LLM-based user agent that generates responses based on the complete key facts list A_n and dialogue history. The user agent follows several key principles: (1) providing concise responses to expert questions based on atomic key facts, (2) Reply "unknown" when uncertain, and (3) refusing to answer manipulative questions that attempt to elicit all known information. To assess the reliability of user agent, we evaluate various LLMs and strategies in terms of relevance and factuality (refer to Appendix C).

3.2 LeCoDe Dataset Construction

The motivation of LeCoDe is to facilitate future research by constructing a high-quality benchmark dataset for legal consultation. Given the limited availability of suitable lawyer–client interaction data, we draw on publicly accessible legal consultation materials from short-video platforms in China. These materials typically involve licensed lawyers providing legal guidance in publicly accessible consultation content intended for legal education. To convert them into research data suitable for legal consultation modeling and evaluation, we design a systematic processing and annotation pipeline.

Two-stage expert annotation: We employ a two-stage expert annotation process to ensure data quality. The first stage focuses on dialogue normalization and structural annotation. Annotators perform dialogue standardization, transcription error correction, speaker role verification, and fine-grained utterance intent labeling. In particular, **intent labeling** assigns each utterance to one of ten predefined classes, such as "Client Initial Query" and "Lawyer Clarifying Question." This stage produces a cleaner and more structured dialogue representation, which also supports the efficiency and consistency of the subsequent legal annotation stage.

The second annotation stage aims to enrich the dialogue dataset with substantive annotations. Annotators with legal educational background identify critical elements, including clients' initial queries, atomic key facts A_n, the importance of key facts, and legal advice summarization. The detailed annotation content is as follows:

Atomic key fact extraction aims to identify the atomic factual units relevant to user needs from lengthy dialogues. On average, a single consultation contains around 9.19 key facts. In addition, the extracted atomic facts complement entity information and ensure consistency with the dialogue flow. During model training, these annotations assist in filtering out irrelevant content, enabling the model to learn effective and professional interaction strategies. In the evaluation phase, these annotations are used to assess clarification capability.

Importance scoring of key facts captures the role of each fact in legal analysis. The facts are categorized into three levels: Critical Facts (3 points), Secondary Facts (2 points), and Non-critical Facts (1 point). This scoring mechanism supports evaluation of whether a model focuses on the core facts of a case, and it can also provide guidance for learning more efficient dialogue strategies.

Legal advice summarization aims to generate legal recommendations that are accurate, comprehensive, and concise. During training, this annotation enhances the model's ability to produce fact-based, professional advice. In the evaluation phase, the task is used to assess the quality and effectiveness of the model's generated advice.

Additionally, two expert annotation reviewers

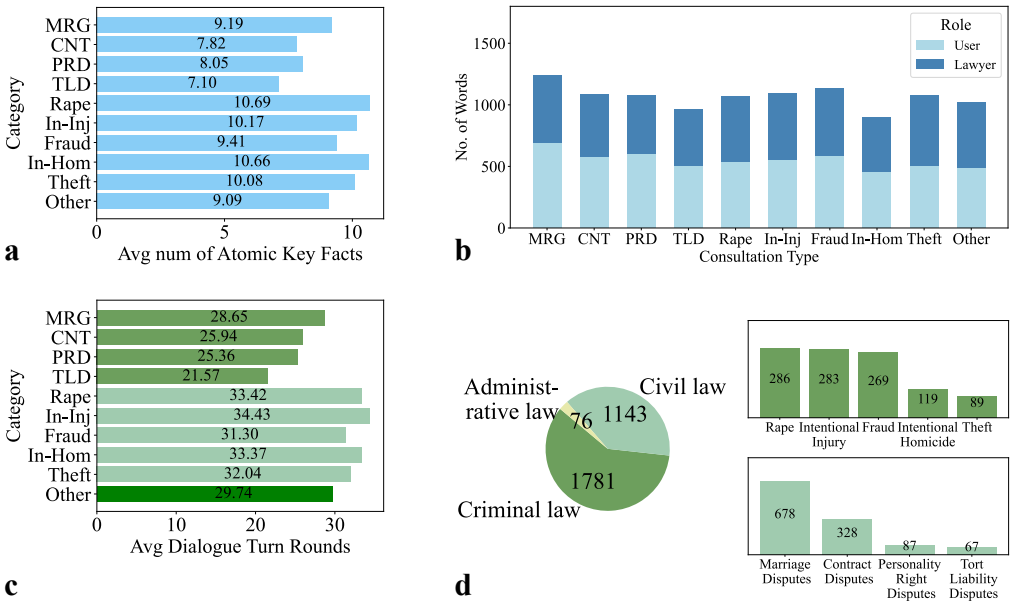

Figure 2: Overview of data distribution in LeCoDe.

conducted thorough quality control and ethical compliance checks (see Section 5). To maintain annotation quality, we developed comprehensive annotation guidelines for both stages and conducted pilot annotations for annotator training. Detailed annotation guidelines are provided in Appendix E. Specifically, we recruited 12 annotators with bachelor's degrees or above for the first stage and 8 annotators with legal educational background for the second stage. In the pilot annotations, the inter-annotator agreement was 88% for intent labeling in Stage 1 and 80% for importance scoring in Stage 2, indicating substantial consistency in the annotation process. For each dialogue, we paid annotators 8 CNY in stage-1 and 9 CNY in stage-2. The total annotation cost amounted to 63,360 CNY ($8,751 at an exchange rate of 7.24 CNY/USD).

3.3 LeCoDe Description

In this section, we present comprehensive statistics of our constructed dataset, followed by an analysis of data distribution.

Data Statistics: LeCoDe is constructed from 3,696 legal consultation cases, which are organized into training and test sets with an 8:2 ratio. Each case contains an average of 29 turns and approximately 9 atomic key facts. The detailed data statistics are shown in Table 2.

Data Distribution: The dataset encompasses a diverse range of legal consultation categories, exhibiting the following characteristics:

Dataset	LeCoDe	Train Set	Test Set
# Dialogue Samples	3,696	2,956	740
# Total turns	110,008	88,342	21,666
avg turns per dialogue	29.76	29.89	29.28
avg key facts per dialogue	9.19	9.24	9.01

Table 2: Statistics over LeCoDe Dataset.

Diverse Legal Scenarios: As illustrated in Figure 2d, LeCoDe covers three major legal domains: Criminal Law, Civil Law, and Administrative Law. Within Criminal Law, the most frequently consulted charges include Rape, Intentional Injury (In-Inj), Fraud, Intentional Homicide (In-Hom), and Theft. In Civil Law scenarios, the most common disputes involve Marriage (MRG), Contract (CNT), Personality Right (PRD), and Tort Liability (TLD).

Varying Complexity Across Domains: Figure 2a presents the average number of key facts per dialogue across common consultation scenarios. Notably, criminal law cases contain more facts compared to civil law cases, reflecting the inherent complexity of criminal proceedings. This pattern is also reflected in Figure 2c, where criminal law consultations average 32.1 turns compared to 26.2 turns in civil law cases.

Balanced Dialogue Dynamics: Figure 2b presents the word count distribution between participants. Interestingly, users' contributions are comparable to, or slightly exceed, those of lawyers. This pattern is consistent with the consultation setting captured in our data. Clients often have limited

legal knowledge and are unable to present complete, well-structured case descriptions in the initial query. Instead, they gradually provide relevant information in response to lawyers' follow-up questions.

Overall, LeCoDe covers diverse legal scenarios and exhibits multi-turn dialogue patterns that are useful for studying legal consultation.

3.4 Evaluation Metrics

We evaluate the legal consultation capability of the expert agent from two perspectives. First, we assess **Clarification Capability** through *effectiveness* and *efficiency* metrics. Second, we evaluate **Advice Quality** using both *automated* and *LLM-based evaluation* metrics.

Clarification Capability: *Effectiveness:* we measure the model's ability to elicit ground truth atomic key facts (A_n) through clarification questions during dialogue simulation. To evaluate the coverage of key facts, we employ multiple metrics: **Recall (R)** measures the proportion of annotated key facts acquired during the interaction, while **Weighted Recall (WR)** additionally takes the importance score of each key fact into account. Since early questioning is important in consultation, we use **Recall@5 (R@5)** to measure how effectively the first five rounds uncover key facts, and **NDCG** (Normalized Discounted Cumulative Gain) to measure whether more important facts are elicited earlier. *Efficiency:* we use the average number of dialogue turns (**AT**), where a lower value indicates more efficient clarification.

Advice Quality: we evaluate the semantic alignment between the model-generated and the annotated reference advice using both *automated metrics* **ROUGE-L** (R-L) (Lin, 2004) and **BERTScore** (BS) (Zhang et al., 2020) and *LLM-based evaluation* metrics. The latter includes: **Professionalism (Pro)**, **Fluency (Flu)**, **Completeness (Com)**, **Satisfaction (Sat)**, and **Safety (Safe)**, and an **Overall Score (OA)**.

Detailed definitions and calculation methods for all metrics are provided in Appendix F.1.

3.5 Training Dialogue Construction Strategy for SFT

Besides evaluation, LeCoDe can also be used to construct training dialogues for supervised fine-tuning (SFT), as shown in Figure 3. We consider three dialogue construction strategies.

Direct SFT directly leverages LeCoDe training data to predict lawyer utterances based on dialogue history, simulating natural consultation flow.

Key-fact SFT employs a targeted approach where stronger LLMs first generate multiple target questions for each atomic key fact. The target model is then trained to learn how to generate these target questions, establishing a one-to-one mapping between questions and facts.

Key-fact-e SFT, an enhanced version of the previous approach, also utilizes stronger LLMs to generate target questions that can effectively cover multiple (1-3) key facts simultaneously. Training data is then constructed based on these questions and corresponding key facts, enabling more efficient information elicitation. Implementation details of these strategies are provided in Appendix D.1.3.

4 Experiment

4.1 Baselines

We evaluate models across three distinct categories: (1) **Closed-source Commercial LLMs**, including GPT series models (GPT-4 (Achiam et al., 2023), GPT-4o (Hurst et al., 2024), GPT-3.5-turbo) and Qwen series models (Qwen-max (Yang et al., 2024), Qwen-turbo (Yang et al., 2024)); (2) **Open-source LLMs**, comprising DeepSeek-R1 (DeepSeek-AI, 2025), DeepSeek-V3 (DeepSeek-AI, 2024), Qwen2.5-72B (Yang et al., 2024), Qwen2.5-7B (Yang et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), and GLM-4-9B (GLM et al., 2024), GLM-4-32B (GLM et al., 2024); and (3) **Legal-domain Specialized Models and Strategies**, which include domain-specific LLMs (ChatLaw-13B (Cui et al., 2023), Lawyer-LLaMA (Huang et al., 2023), Tongyi-Farui (Tongyi-farui, 2025)). The clarifying strategy incorporates **MediQ** (Li et al., 2024b), a question-asking framework adapted from clinical reasoning to legal consultation. Detailed experimental settings are provided in Appendix D.1. Additionally, we propose three training dialogue construction strategies for supervised fine-tuning (SFT) to enhance LLMs' legal consultation capabilities: **Direct SFT**, **Key-fact SFT** and **Key-fact-e SFT**, as described in Section 3.5.

4.2 Experiment Setting

For user simulation, LLM-based agents have demonstrated strong response capabilities (Hu et al., 2023; Li et al., 2024b; Sekulic et al., 2024).

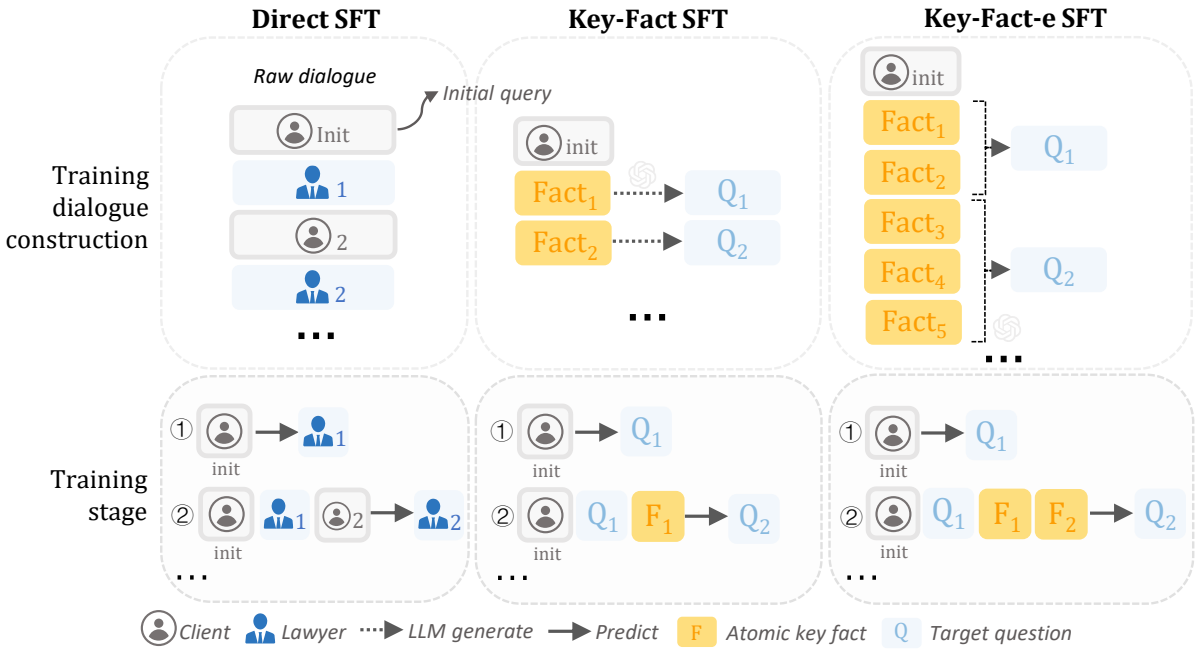

Figure 3: Dialogue construction strategies for supervised fine-tuning (SFT).

In our setting, an ideal user simulator should exhibit both relevance (responding appropriately to lawyers' questions) and factuality (adhering to given key fact without hallucination). After experimenting with various response strategies and LLMs (detailed settings and results in Appendix C), we use Qwen-max with First-Person Given Dialogue Instruction strategy for user simulation, which achieves best performance on relevance and factuality.

For expert agents, we implement both zero-shot and few-shot strategies across all general LLMs. In few-shot scenarios, we manually crafted two demonstrations. The maximum number of interaction rounds is set to 10. This limit is used for simulated evaluation and is intended to provide sufficient room for clarification without allowing unconstrained interaction. If the turn limit is reached, the expert agent is instructed to generate final legal advice. All prompt templates and settings are provided in the Appendix D. We use the evaluation metrics in Section 3.4 for evaluation, and use GPT-4o-mini as the LLM judge for advice quality assessment (see Appendix H). We further validate both the user simulation setting and the LLM-based advice evaluation through human validation, as reported in Section 4.4.

4.3 Experiment Results

We report zero-shot performance in Table 3, with few-shot results provided in the Table 6 and Table 7 in the Appendix. Our experiments reveal several key findings:

Clarification Capability: **(1)** Closed-source commercial LLMs generally outperform open-source LLMs. GPT-4 achieves the best balance between efficiency and effectiveness, demonstrating strong performance in questioning ability. **(2)** Qwen-max and Qwen2.5-72B show notable performance in Recall and Weighted Recall. **(3)** While MediQ effectively reduces dialogue turns, it compromises question quality. Domain-specific models like ChatLaw and Lawyer-LLama often bypass the questioning phase and provide direct advice due to poor instruction following, resulting in shorter but less effective consultations. **(4)** Notably, SFT strategies significantly outperform other models across all clarification metrics, achieving remarkable Recall (53.8%) and NDCG (84.8%) scores.

Advice Quality: **(1)** GPT-4 leads among closed-source LLMs across multiple advice quality metrics. **(2)** Deepseek-R1 excels among open-source LLMs, even achieves SOTA performance on LLM-based advice evaluation metrics. **(3)** SFT strategies demonstrate superior performance in both automated metrics and LLM-based evaluation.

Our analysis in Table 6 and Table 7 show that few-shot prompting negatively impacts clarifica-

Models	Clarification Capability					Advice Quality							
	Effe.				Effi.	Auto.		LLM-Eval.					
	R	WR	R@5	NDCG	AT	R-L	BS	Pro	Flu	Com	Sat	Safe	OA
Closed-source LLMs													
GPT-4	35.9	37.2	34.4	74.7	4.9	17.5	65.9	56.8	71.9	49.9	55.8	71.1	58.1
GPT-4o	33.0	33.9	25.9	63.2	10	13.0	62.9	50.1	69.6	44.7	52.4	66.6	53.5
GPT-3.5	26.9	28.0	23.6	67.1	9.2	12.4	63.1	51.8	67.6	45.8	51.6	66.0	53.8
Qwen-max	39.8	41.3	31.2	70.7	10	13.9	63.9	54.0	71.4	47.6	55.3	68.9	56.2
Qwen-turbo	27.3	28.3	23.2	61.2	9.4	10.6	61.8	44.8	62.7	40.0	47.6	61.6	48.5
Open-source LLMs													
deepseek-v3	28.0	29.0	25.1	65.4	7.5	16.2	65.3	57.6	72.9	50.8	56.2	72.1	58.9
deepseek-r1	27.6	28.8	25.1	63.9	6.3	12.4	64.0	**63.8**	74.1	**55.5**	55.9	**76.5**	**62.2**
Qwen2.5-72B	39.6	40.8	30.4	71.9	10	14.7	63.8	54	71.6	48.5	55.2	69.4	56.6
Qwen2.5-7B	31.1	32.1	27.7	70.1	10	9.2	60.0	42.3	60.4	37.6	45.0	57.8	45.7
Llama-3.1-8B	22.1	23.1	19.1	55.0	7.8	13.9	63.1	51.7	69.1	46.5	51.7	67.1	54.1
GLM4-32B	34.1	35.4	28.8	67.3	7.3	14.3	64.1	55.6	70.4	49.0	54.3	70.4	56.7
GLM4-9B	34.2	35.3	26.1	65.0	10	13.9	63.1	51.7	69.1	46.5	51.7	67.1	54.1
Domain-specific LLMs or Clarifying Strategy													
ChatLaw	16.7	17.4	16.3	47.8	4.4	7.8	53.0	28.8	40.0	25.5	31.0	39.4	30.7
Lawyer-LLaMA	21.8	22.9	21.8	62.9	4.2	12.4	59.7	37.6	52.7	33.1	39.0	51.8	39.4
Farui	28.4	29.5	25.1	64.2	8.9	16.6	64.5	52.2	66	47.4	51.3	66.3	53.8
MediQ	22.1	22.9	19	53.5	**4.2**	13.3	63.7	51.8	69.1	45.5	51.9	66.7	54.0
Direct SFT	37.2	37.6	25.6	62.3	9.6	15.4	64.9	51.7	65.3	45.6	50.1	63.1	52.4
Key-fact SFT	**53.8**	**55.1**	43.5	84.7	9.2	18.5	67.5	59.3	73.0	52.3	55.9	71.2	59.4
Key-fact-e SFT	51.0	51.1	**45.3**	**84.8**	6.6	**19.1**	**67.6**	59.3	**73.6**	52.0	**56.7**	72.1	59.5

Table 3: Main results on zero-shot setting, where top-6 scores are marked in blue, the highest is **bolded** and the second-highest is underlined.

tion capability in most models (6/11), while improving advice quality (8/11). This suggests that few-shot examples may constrain questioning behavior, while still providing useful demonstrations for advice generation. Overall, **current LLMs remain limited in legal consultation**. Although **SFT strategies improve performance**, there is still substantial room for improvement, particularly in clarification capability (with the best Recall at 53.8%) and advice quality.

4.4 Human Validation

To further verify the reliability of our evaluation setup, we conducted expert human validation for two key components: the LLM-based user simulator and the LLM-based evaluation for advice quality.

LLM-based User Simulation Validation: As shown in Table 4 of Appendix C, Qwen-max performs well as a user simulator under our automatic relevance and factuality evaluation. To further validate this choice, we randomly sampled 50 responses generated by Qwen-max and asked a legal expert to assess whether each response was appropriate given the same key fact list and the corresponding lawyer question. The proportion of responses judged appropriate was 88%, which supports the use of Qwen-max as the user agent in our evaluation setting.

LLM-based Evaluation Validation: Since our evaluation of advice quality relies on an LLM-as-a-judge setup, we further conducted human validation against expert judgment. Specifically, we randomly sampled 600 dialogue pairs across models, and three legal experts independently rated the generated advice on a 1–10 scale across six dimensions: Professionalism, Fluency, Completeness, Satisfaction, Safety, and Overall Assessment. The agreement rate between GPT-4o and expert ratings was 0.868, indicating reasonably strong alignment between the LLM judge and human experts. This result supports the use of LLM-based evaluation as a scalable assessment method in our benchmark.

4.5 Further Analysis

In this section, we analyze several findings from our experiments to derive insights for future improvements in legal consultation scenarios.

As illustrated in Figure 4(a), we analyze the relationship between case complexity (measured by the number of atomic key facts) and model performance, including Weighted Recall (WR) for clar-

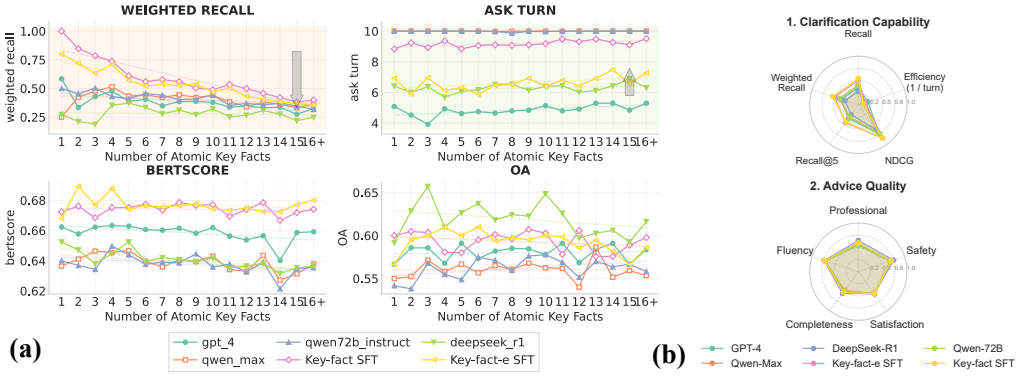

Figure 4: (a) Impact of Case Complexity on Model Performance: Analysis of Weighted Recall, Average Turn, BERTScore, and OA across different numbers of atomic key facts. (b) Radar chart showing the LLMs performance.

ification effectiveness, Average Turn (AT) for efficiency, and BERTScore (BS) and Overall Score (OA) for advice quality. The results reveal several notable patterns. As case complexity increases, WR declines substantially, while BS and OA also show moderate decreases, **indicating that model performance worsens as cases become more complex**. Interestingly, AT shows an upward trend, **suggesting that LLMs partially adjust their interaction length as case complexity increases**. These findings suggest that LLMs may benefit from better mechanisms for assessing case complexity and prioritizing more informative questions.

Significant room for improvement in LLMs' legal consultation capabilities. As shown in Figure 4(b), nearly all LLMs achieve only 30%-50% recall in clarification capability. Moreover, substantial gaps remain in advice quality, including professional, user satisfaction, and completeness.

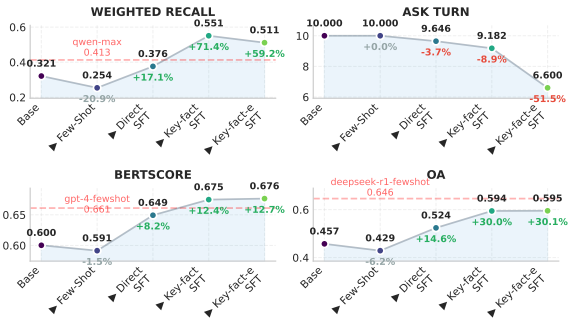

Figure 5: Performance on Different Strategies.

How to effectively enhance LLMs' legal consultation abilities? As shown in Figure 5, using Qwen2.5-7B-Instruct as our base model, it reflects few-shot strategies consistently produced adverse effects, suggesting that conventional **prompt engineering approaches may be insufficient for meaningful improvements**. Consequently, we propose SFT strategies to rapidly and effectively develop consultation capabilities from the dataset. The results show that Key-fact SFT, which trains LLMs to target specific key facts through strategic questioning, significantly outperforms Direct SFT's raw dialogue pattern learning approach. Notably, Key-fact-e improves performance across clarification and advice metrics (WR: +59.2%, BS: +12.7%, OA: +30.1%) while also reducing the average number of turns, and in several metrics it performs competitively with stronger baseline models.

5 Conclusion

We introduce LeCoDe, a benchmark for legal consultation that is accompanied by an evaluation framework measuring LLMs' clarification capability and advice quality. Our analysis reveals significant limitations of existing LLMs in legal consultation tasks. Further, we propose supervised fine-tuning strategies based on LeCoDe that lead to consistent improvements in the legal consultation performance of LLMs. In the future, we encourage researchers to explore two promising avenues: (1) exploring more strategies to enhance LLMs' consultation capabilities and (2) effectively incorporating external legal knowledge. We hope these efforts will support future research on more capable and accessible professional consultation systems.

Limitations

Our work has several limitations and also opens up opportunities for future research. First, the current benchmark is developed in the context of Chinese legal consultation. Given that legal systems are

inherently jurisdictional, it is not uncommon for respected legal benchmarks (Hwang et al., 2022; Östling et al., 2023; Li et al., 2024a) to focus on a single jurisdiction. In future work, it would be valuable to incorporate consultation data from more jurisdictions and legal systems in order to evaluate LLMs across different legal contexts.

Second, while our framework focuses on professional legal consultation capabilities - specifically clarification capability and professional advice quality - real-world consultations are often more complex and require chit-chat and emotional support, aspects that could be incorporated into future evaluation frameworks.

Third, although we focus on legal consultations, similar interactive consultation scenarios exist in other professional domains, such as healthcare. We plan to validate our framework extension beyond legal applications.

Fourth, our evaluation setup relies in part on an LLM-based user simulator and an LLM-based judge for advice quality. Although we conduct expert human validation and observe encouraging agreement, some discrepancy between automated and expert judgment may still remain. Future work may further improve human alignment through calibration techniques for automated evaluation (Von Däniken et al., 2022, 2025).

Finally, while our SFT-based strategies show improvements over baseline LLM performance, the overall results remain suboptimal. We encourage researchers to explore additional approaches for enhancing model capabilities, such as reinforcement learning techniques to optimize consultation strategies, external knowledge integration through RAG or knowledge graph approaches, and other methods for improving model performance in professional consultation scenarios.

Ethical considerations

Considering the sensitivity of the legal domain, we adopted strict measures to ensure ethical compliance and minimize potential risks. First, our dataset is constructed from publicly accessible legal consultation materials shared online for legal education in the Chinese context. The source materials are publicly accessible and were originally shared for legal education purposes. The data source is consistent with numerous peer-reviewed studies that also utilize publicly available social media videos (such as those on YouTube) for academic purposes (Malhotra et al., 2022; Albadi et al., 2022). To minimize risk, we neither retain nor distribute any raw audio or video files. All research data is provided exclusively as deeply processed and expert-annotated text, with all potentially identifying information (e.g., names, contact details) thoroughly removed and additional rounds of anonymization conducted as necessary.

Second, legal experts performed comprehensive content reviews to exclude any potentially discriminatory, violent, or offensive material, while further reducing re-identification risks and mitigating potential biases. Both automated and manual procedures were employed for de-identification to reduce the risk that the dataset could be traced back to specific individuals.

Furthermore, access to LeCoDe will be strictly governed for academic research purposes only (see Appendix A.1) through the following measures:

- **Institutional Verification:** Only qualified researchers affiliated with accredited institutions are granted access for clearly defined research objectives.
- **Binding Data Use Agreement:** All applicants must sign a formal agreement prohibiting commercial use, re-identification attempts, deployment in legal or product settings, or any form of redistribution.
- **Auditability and Traceability:** Comprehensive records of data access and usage are maintained to ensure accountability and deter misuse.
- **Compliance Guidance:** Documentation provides explicit guidance on legal and policy considerations, making users responsible for compliance with local regulations and prioritizing the rights of original content owners.

We believe these measures help mitigate potential social risks. A detailed discussion of potential impacts is provided in Appendix B.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (72574198, 62106039, 72134007). This work is supported by Alibaba Innovative Research Program.

References

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

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2022. Deradicalizing youtube: characterization, detection, and personalization of religiously intolerant arabic videos. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–25.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

CAIL. 2023. CAIL2023: Conversational similar case retrieval. http://cail.cipsc.org.cn/task_summit.html?raceID=2&cail_tag=2023. Accessed: May 12, 2025.

CAIL. 2024. CAIL2024: Legal consultation dialogue generation. http://cail.cipsc.org.cn/task_summit.html?raceID=4&cail_tag=2024. Accessed: May 12, 2025.

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

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, Abu Dhabi, UAE. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

Jenny Hong, Derek Chong, and Christopher D Manning. 2021. Learning from limited labels for long legal dialogue. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 190–204.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3953–3957, New York, NY, USA. Association for Computing Machinery.

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*.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Daniel Kirk, Cagatay Catal, and Bedir Tekinerdogan. 2021. Precision nutrition: A systematic literature review. *Computers in Biology and Medicine*, 133:104365.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024a. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bulou Liu, Yiran Hu, Qingyao Ai, Yiqun Liu, Yueyue Wu, Chenliang Li, and Weixing Shen. 2023. Leveraging event schema to ask clarifying questions for conversational legal case retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 1513–1522, New York, NY, USA. Association for Computing Machinery.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.

Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson, and Felix Steffek. 2023. The cambridge law corpus: A dataset for legal ai research. *Advances in Neural Information Processing Systems*, 36:41355–41385.

Rowena Rodrigues. 2020. Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4:100005.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35, St. Julians, Malta. Association for Computational Linguistics.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. Lawluo: A chinese law firm co-run by llm agents. *arXiv preprint arXiv:2407.16252*.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Tongyi-farui. 2025. Tongyi farui:leading ai legal product powered by alibaba cloud. https://tongyi.aliyun.com/farui/home. 2025.5.12.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Pius Von Däniken, Jan Milan Deriu, and Mark Cieliebak. 2025. A measure of the system dependence of automated metrics. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 87–99.

Pius Von Däniken, Jan Milan Deriu, Don Tuggener, and Mark Cieliebak. 2022. On the effectiveness of automated metrics for text generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1503–1522.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024. Knowledge-infused legal wisdom: Navigating LLM consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15542–15555, Bangkok, Thailand. Association for Computational Linguistics.

Nan Xie, Yuelin Bai, Hengyuan Gao, Ziqiang Xue, Feiteng Fang, Qixuan Zhao, Zhijian Li, Liang Zhu, Shiwen Ni, and Min Yang. 2024. Delilaw: A chinese legal counselling system based on a large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5299–5303.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang,

Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7577–7597, Miami, Florida, USA. Association for Computational Linguistics.

Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Chengyuan Liu, Jun Lin, Ji Zhang, Kun Kuang, and Xiaozhong Liu. 2026. A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction. *Information Processing & Management*, 63(1):104319.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Licenses

A.1 Licenses

This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0). **All resources are for scientific research only.**

B Discussion

B.1 Broader Impact

We propose LeCoDe to support the study and evaluation of LLMs in legal consultation scenarios, providing a benchmark and evaluation framework for future research. However, given the sensitive and high-risk nature of the legal domain, we must carefully consider potential risks. While we actively explore LLMs' applications in legal scenarios, we must guard against potential unfairness and societal disruption. The dataset is intended purely for academic research, not as a replacement for expert legal consultation. Legal advisory requires rich professional knowledge, practical experience, and human insight that no LLM or AI technology can substitute.

LeCoDe comprises real-world consultation data, and we have made extensive efforts to filter out potentially discriminatory, violent, or offensive content while ensuring anonymity and mitigating potential biases. However, we acknowledge that using LeCoDe for model training may still introduce potential biases and discrimination. Therefore, we emphasize that the dataset is licensed exclusively for academic research and prohibited from commercial or practical applications.

To mitigate potential negative impacts, we explicitly state the dataset's usage restrictions and limitations, and continuously monitor the practical applications of research outcomes to adjust recommendations accordingly. Ultimately, we hope this work contributes to the intelligent development of legal services while ensuring technological advancement serves the goals of social fairness and justice.

C User Agent

To ensure our LLM-based agent for user simulation, we propose several strategies and evaluate the reliability for user agent. We first introduce the evaluation metrics, then evaluation settings and corresponding prompt template, finally show the evaluation results for user agent reliability.‘

C.1 Evaluation Metrics for User Agent

While LLM-based models have demonstrated strong response capabilities (Li et al., 2024b), our evaluation focuses on two critical aspects of an ideal user simulator: **relevance** (appropriate responses to lawyers' questions) and **factuality** (adherence to given information without hallucination).

Relevance measures the user agent's ability to comprehend and provide pertinent responses to lawyers' clarifying questions, evaluating the model's question understanding capabilities.

Factuality assesses whether the user's responses to clarifying questions align with the provided atomic fact background information, measuring the model's ability to accurately extract and utilize known information while avoiding fabrication.

C.2 Evaluation Setting for User Agent

We randomly selected 100 initial queries with their corresponding 5-round dialogue contexts from the training set, for which we manually generated three clarifying questions. The evaluation was conducted using various Qwen-series models, including Qwen2.5-7B, Qwen2.5-32B, Qwen2.5-72B, and Qwen-max. GPT-4o-mini served as the judge for assessing both relevance and factuality.

We evaluated three prompt engineering strategies:

- **Instruct Given Dialogue**: Providing complete interaction dialogue history and atomic key facts to instruct LLM responses

- **Instruct Given Question**: Providing only the current question and atomic key facts for LLM responses

- **First-person Given Dialogue**: Instructing LLMs to simulate first-person responses based on complete interaction dialogue

The prompt templates for each strategies can be seen in Appendix C.3. The evaluation prompt template for GPT-4o-mini can be seen in Appendix C.4.

C.3 Prompt Templates for each PE Strategies

Instruct Given Dialogue (translated from Chinese)

```
You are an honest and reliable legal assistant who understands user-related information and attempts to answer lawyers' questions about users.

The following is [Atomic Key Facts] describing user information
{Atomic Key Facts}

The following is the interactive dialogue between you (user) and the lawyer
{current dialogue}

- Use the above key fact to answer the lawyer's questions, selecting no more than 3 answers.
- If the above key fact cannot answer the lawyer's questions, [only reply "I Don't know"].
- Only answer what is asked in the question, without providing any analysis, speculation or conclusions.
- Only select all content from the above information that can answer the question, and only that.
```

Instruct Given Question (translated from Chinese)

```
You are a honest and reliable legal assistant who understands user-related information and attempts to answer lawyers' questions about users.

The following is [Atomic Key Facts] describing user information
{Atomic Key Facts}

The following is the lawyer's question you need to answer
{lawyer question}

- Use the above key fact to answer the lawyer's questions, selecting no more than 3 answers.
- If the above key fact cannot answer the lawyer's questions, [only reply "I Don't know"].
- Only answer what is asked in the question, without providing any analysis, speculation or conclusions.
- Only select all content from the above information that can answer the question, and only that.
```

First-person Given Dialogue (translated from Chinese)

You are simulating a real user in a legal consultation scenario.
Your role is a consultant [user] who encounters legal difficulties and seeks legal help. Since you lack sufficient legal knowledge, you cannot raise a clear, professional legal question. Therefore, you need to respond to the lawyer's questions based on the [Atomic Key Facts] you know.

The following is [Atomic Key Facts] describing user information
{Atomic Key Facts}

The following is the interactive dialogue between you (user) and the lawyer
{current dialogue}

Your task is to answer relevant questions based on the [Atomic Key Facts] description.
Note
<1>Your role is set as seeking legal help, and you cannot use professional legal terminology.
<2>Only answer the lawyer's questions based on the content of [Atomic Key Facts], do not generate other content. Select no more than 3 key pieces of information to answer in one reply. If [Atomic Key Facts] cannot answer the question, then [only reply "I Don't know"].
<3>For any question you're unsure about, [only reply "I Don't know"].
<4>Keep replies as brief as possible.
<5>Prevent excessive information disclosure.

C.4 Prompt Templates for LLM-as-a-judge for assessing relevance and factuality

LLM-as-a-judge for assessing relevance (translated from Chinese)

You will receive a [Clarifying Question] raised by a lawyer in the dialogue, and a [User Response] based on [Atomic Facts] background information. Your task is to perform a binary classification (0 or 1) on the user's response, judging its relevance to the lawyer's [Clarifying Question].

In some cases, the user may reply "I Don't know" based on the lawyer's advice.
If the user replies "I Don't know", you need to determine whether the [Clarifying Question] cannot be answered based on the content of [Atomic Key Facts]. If it indeed cannot be answered, then the relevance of replying "Don't know" should be 1; otherwise, it's 0.
[Dialogue Context]
{dialogue}

[Atomic Key Facts]
{Atomic Key Facts}

[Clarifying Question]
{lawyer question}

[User Response]
{User Response}

LLM-as-a-judge for assessing factuality (translated from Chinese)

```
You will receive a [User Response] based on [Atomic Key Facts] background information. Your task is to perform a binary classification (0 or 1) on the factuality of the user's response, determining whether it is consistent with the provided atomic facts background information.

[Dialogue Context]
{dialogue}

[Atomic Key Facts]
{Atomic Key Facts}

[Clarifying Question]
{lawyer question}

[User Response]
{User Response}
```

C.5 Evaluation Results for User Agent

As shown in Table 4, Qwen-max with First-Person Given Dialogue approach achieved outstanding performance in both relevance (97.0%) and factuality (99.3%) metrics. This superior performance demonstrates the effectiveness of using first-person perspective in user simulation tasks. Notably, when comparing different model scales, we also observe a clear scaling law pattern: larger LLMs consistently outperform their smaller counterparts across all evaluation settings. For instance, Qwen2.5-32B shows improvements over Qwen2.5-7B, and Qwen2.5-72B further enhances the performance.

Interestingly, even smaller models like Qwen2.5-7B demonstrate reasonable performance (86.5% relevance and 96.0% factuality). This robust performance across different model scales indicates that the user simulation approach is feasible and practical even with more accessible open-source models. These results indicate that LLM-based user simulation is feasible in this benchmark setting, including with more accessible open-source models.

D Lawyer Agent

D.1 Experiment Setting

D.1.1 Experiment Setting for API-Based LLMs

We first present the experimental settings for LLMs accessed via API calls, including closed-source models, open-source models, and several legal domain-specific models. These models are evaluated using both zero-shot and few-shot strategies with no additional fine-tuning. All prompts for zero-shot and few-shot settings can be found in Appendix D.2.1.

First, we introduce the versions of each general LLM (include both open-source and close-source LLMs) as shown in Table 5. For closed-source LLMs (GPT-series, qwen-max, qwen-turbo, farui) and deepseek-series (deepseek-v3, and deepseek-r1), we directly call APIs for generation. For all other models, the evaluation is conducted using vllm (Kwon et al., 2023) to enable an OpenAI-Compatible Server for model calls, utilizing 4 NVIDIA A100 GPUs with 80GB of memory. The maximum sequence token is set to 2048 and temperature for generation is set to 0.7 for open-source models.

D.1.2 Experiment setting for Mediq Strategy

We adapt the mediq-expert framework (Li et al., 2024b), originally designed for clinical decision-making through information-seeking questions, to our legal consultation scenario. While the original framework was developed for medical multiple-choice questions, we modify and retain three key components: the Abstention module, Ask module, and a Suggestion module for legal advice generation.

The workflow operates as follows: The Abstention module first assesses the system's confidence level for providing legal advice. When confidence is low, the Ask module activates to gather additional information through asking a clarifying question. The dialogue concludes either when the system reaches sufficient confidence or when the maximum number of interaction turns is reached, at which point the Suggestion module generates appropriate legal advice.

We utilize Qwen-max as backbone LLM for each module in mediq. All prompts for mediq-expert can be found in Appendix D.2.3.

Models	Instruct Given Dialogue	Instruct Given Question	First-Person Given Dialogue
Relevance			
Qwen2.5-7B	86.1	90.6	86.5
Qwen2.5-32B	90.2	91.1	96.9
Qwen2.5-72B	90.9	90.0	93.7
Qwen-max	92.4	92.2	**97.0**
Factuality			
Qwen2.5-7B	85.0	94.3	96.0
Qwen2.5-32B	94.5	98.0	99.5
Qwen2.5-72B	97.5	98.0	99.5
Qwen-max	99.0	99.0	**99.3**

Table 4: Results on user agent reliability.

Model Names	**Model Version or Source**
GPT-4	gpt-4-turbo-2024-04-09
GPT-4o	gpt-4o-mini-2024-07-18
GPT-3.5-turbo	gpt-3.5-turbo-1106
Qwen-max	qwen-max-2024-09-19
Qwen-turbo	qwen-turbo-2025-02-11
deepseek-v3	https://huggingface.co/deepseek-ai/DeepSeek-V3
deepseek-r1	https://huggingface.co/deepseek-ai/DeepSeek-R1
Qwen2.5-72B	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
Qwen2.5-7B	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Llama-3.1-8B	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
GLM4-32B	https://huggingface.co/THUDM/GLM-4-32B-0414
GLM4-9B	https://huggingface.co/THUDM/glm-4-9b-chat
ChatLaw	https://huggingface.co/pandalla/ChatLaw-13B
Lawyer-LLaMA	https://github.com/AndrewZhe/lawyer-llama
farui	https://help.aliyun.com/zh/model-studio/tongyi-farui-api

Table 5: LLMs version and source url.

D.1.3 Experiment setting for SFT Strategy

To enhance LLMs' legal consultation capabilities, we introduce several strategies to construct training multi-turn dialogue for effectively training LLMs by Supervised Fine-Tuning (SFT).

Direct SFT directly uses the original dialogue sequences in the LeCoDe training data to predict lawyer utterances based on dialogue history, simulating natural consultation flow. Formally, given a consultation dialogue sequence $D = \{(u_t^C, u_t^E)\}_{t=1}^T$, where u_t^C and u_t^E represent utterances of Client C and Expert (lawyer) E at turn t, the model learns to generate lawyer responses sequentially. For instance, when a client initiates with query u_1^C, the model learns to predict the lawyer's first response u_1^E. Subsequently, using the concatenated context (u_1^C, u_1^E, u_2^C), the model learns to predict the next lawyer utterance u_2^E.

Key-fact SFT employs a targeted approach where stronger LLMs first generate multiple target questions for each atomic key fact. The target model is then trained to learn how to generate these target questions, establishing a one-to-one mapping between questions and facts. Formally, given each key fact a_t in the atomic key facts A_n, we use Qwen-max to generate a target question designed to elicit specific information a_t. Then we construct the new multi-turn dialogue sequence $D = \{(q_t, a_t)\}_{t=1}^N$, where q_t represents generated target question and a_t represents corresponding atomic key fact. The model learns to generate appropriate questions sequentially.

For instance, when given an initial query, the model learns to generate q_1 to extract a_1. Subsequently, using the concatenated context (q_1, a_1), it predicts the next question q_2 to obtain a_2, ensuring systematic information gathering through targeted questioning. Following this strategy, we

can construct the Key-fact SFT training dialogue.

Key-fact-e SFT, an enhanced version of the previous approach, also utilizes stronger LLMs (qwen-max) to generate target questions that can effectively cover multiple (1-3) key facts simultaneously. Training data is then constructed based on these comprehensive questions and their corresponding key facts, enabling more efficient information elicitation.

For all strategies, we fine-tune Qwen2.5-7B-Instruct using the same prompt template from zero-shot experiments in Appendix D.2.1. The prompt templates for Qwen-max to generate the target questions in Key-fact SFT and Key-fact-e SFT are provided in Appendix D.2.4. We run experiments on 4 NVIDIA A100 GPUs with 80GB of memory. The training configuration employs LoRA (rank=8, alpha=16) with an effective batch size of 64 (4 samples per device × 4 gradient accumulation steps). The optimization process uses a cosine learning rate schedule initialized at 1e-5 with 10% warmup ratio over 3 epochs.

D.2 Lawyer Agent Prompt

D.2.1 Zero-shot Prompt

Lawyer Agent Zero-shot Prompt for ask or advice (translated from Chinese)

```
You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to ask clarifying questions to better understand the details of the user's case or providing legal advice.

Users may present unclear or non-professional questions. In such cases, ensure your questions directly target key information, guiding users to provide more relevant details. The goal is to make users' needs clearer to provide more accurate legal advice.

Below is the user's query and your interaction dialogue:
{current dialogue}

Notes:
a. You have two operations: 1. Question, 2. Advice. Choose one operation per round.
b. If user information is insufficient, choose operation 1. Question.
- Ask only one critical question per round.
- Questions should be concise, without any additional information.
- Do not ask repeated questions!!
- Questions should verify specific facts. Considering users seek legal help but lack legal background, avoid asking about typical court sentencing or specific law violations.
- Maximum 10 rounds of questions; must provide advice after 10 rounds.
c. When confident in addressing user needs, choose operation 2. Advice to provide professional legal advice.
d. Response template:
"question:a clarification question" or "advice:legal advice"
```

Lawyer Agent Zero-shot Prompt for advice generation (translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to provide legal advice.

Below is the user's query and your interaction dialogue:
{current dialogue}

Notes:
a. Your task is to provide legal advice according to the context.
b. Response template:
"advice:legal advice"

D.2.2 Few-shot Prompt

Lawyer Agent Few-shot Prompt for ask or advice (translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to ask clarifying questions to better understand the details of the user's case or providing legal advice.

Users may present unclear or non-professional questions. In such cases, ensure your questions directly target key information, guiding users to provide more relevant details. The goal is to make users' needs clearer to provide more accurate legal advice.

Below is two demonstrations:
<demo1>
{demo1}

<demo2>
{demo2}

Below is the user's query and your interaction dialogue:
{current dialogue}

Notes:
a. You have two operations: 1. Question, 2. Advice. Choose one operation per round.
b. If user information is insufficient, choose operation 1. Question.
- Ask only one critical question per round.
- Questions should be concise, without any additional information.
- Do not ask repeated questions!!
- Questions should verify specific facts. Considering users seek legal help but lack legal background, avoid asking about typical court sentencing or specific law violations.
- Maximum 10 rounds of questions; must provide advice after 10 rounds.
c. When confident in addressing user needs, choose operation 2. Advice to provide professional legal advice.
d. Response template:
"question:a clarification question" or "advice:legal advice"

Lawyer Agent Few-shot Prompt for advice generation (translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to provide legal advice.

Below is two demonstrations:
<demo1>
{demo1}

<demo2>
{demo2}

Below is the user's query and your interaction dialogue:
{current dialogue}

Notes:
a. Your task is to provide legal advice according to the context.
b. Response template:
"advice:legal advice"

Demonstration 1 (translated from Chinese)

<demo1>
User: On August 21st, I heard he was suspected of fraud. When police went to arrest him, he injured three people while fleeing but was eventually caught.
Lawyer: Question: Could you first clarify if the suspect was alone at the time?
User: His girlfriend was with him and also got arrested, but she was released that night.
Lawyer: Question: Do you know the specific amount involved in the alleged fraud?
User: I don't know the exact amount as we can't contact the suspect now. Lawyer: Question: Where did the incident take place?
User: It happened in a public place in another city.
Lawyer: Question: Did he intentionally crash into the crowd?
User: Not sure if he deliberately rushed toward the crowd.
Lawyer: Advice: First, we need to arrange a lawyer to meet with your cousin to understand the details of the fraud allegations and the injuries of the three people. Currently, we can't determine the specific sentence as it depends on the severity of injuries and specific circumstances. If the injuries are minor and only involved three people, it might be classified as assaulting police or obstructing official duties; if there were many people present and he was driving recklessly, it could constitute endangering public safety by dangerous means, which carries more severe penalties. After understanding all relevant information, we can develop a targeted strategy. Additionally, trying to obtain forgiveness letters from the three police officers is an important step, but this depends on their willingness. In summary, gathering all relevant information is the crucial first step.

Demonstration 2 (translated from Chinese)

<demo2>
User: My husband has several mistresses.
Lawyer: Question: What's your husband's occupation? Where does he get so much money?
User: He's a procurement officer, responsible for company purchases. He accepts kickbacks from suppliers, mostly illegal gains.
Lawyer: Question: Do you own any properties? Are you aware of the source of funds for these properties?
User: I own two houses, but I suspect the funds might be related to his illegal income.
Lawyer: Question: Have you previously discussed divorce with your husband? Were any specific agreements reached?
User: We negotiated divorce terms, demanding 200,000 payment, with a 400,000 penalty clause, but important terms were deleted and no agreement was reached.
Lawyer: Question: What are your current preliminary arrangements regarding property division and child support?
User: We discussed a 200k payment + 400k penalty, but no final deal.
Lawyer: Advice: His actions may constitute bribery (>30k RMB is criminal). Negotiate first—show you have evidence, but delay legal action. Secure the properties first; defer the 200k payment. Detail child support (living, education, medical). Preserving his job may benefit long-term child support. Proceed cautiously on mistresses to avoid jeopardizing core goals.

D.2.3 Mediq Prompt

Mediq: Abstain Module(translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to provide legal advice.

Below is the user's query and your interaction dialogue:
{current dialogue}

Evaluating above case facts, are you confident to provide legal advice according to the current dialogue?
Answer with YES or NO and NOTHING ELSE.

Mediq: Ask Module(translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to provide legal advice.

Your task is to ask clarifying questions to better understand the details of the user's case or providing legal advice.
Below is the user's query and your interaction dialogue:
{current dialogue}

Notes:
a. You should ask a clarifying question.
b. If user information is insufficient, choose operation 1. Question.
- Ask only one critical question per round.
- Questions should be concise, without any additional information.
- Do not ask repeated questions!!
- Questions should verify specific facts. Considering users seek legal help but lack legal background, avoid asking about typical court sentencing or specific law violations.
- Maximum 10 rounds of questions; must provide advice after 10 rounds.
b. Response template:
"question:a clarification question"

Mediq: Suggestion Module(translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users. Your task is to provide legal advice.

Below is the user's query and your interaction dialogue:
{current dialogue}

Notes:
a. Your task is to provide legal advice according to the context.
b. Response template:
"advice:legal advice"

D.2.4 SFT Strategies Prompt

Key-Fact SFT to generate a target question (translated from Chinese)

You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users.

I will provide a list of key information points that need to be obtained from clients but are currently unknown. Please design professional, guiding questions for each information point to effectively elicit these details.
[Atomic Key Facts]
{Atomic Key Facts}

Requirements:
1. Return a dictionary mapping each key information index to a corresponding question
2. Each question should:
-Use neutral, professional terminology
-Avoid leading language
-Be easily understood by clients
-Naturally elicit the target information
-Be concise

Key-Fact-e SFT to generate target questions (translated from Chinese)

```
You are a lawyer well-versed in Chinese law, responsible for providing legal consultation to users.


I will provide a list of key information points that need to be obtained from clients but are currently unknown. Please design a series of professional, guiding questions that efficiently cover these information points.
[Atomic Key Facts]
{Atomic Key Facts}

Requirements:
1. Return a dictionary where:
- Each key is a question - Each value is a list of indices indicating which key information points the question covers 2. Each question should:
-Use neutral, professional terminology
-Avoid leading language
-Be easily understood by clients
-Naturally elicit the target information
-Be concise
```

E Details of Annotation Process

E.1 Guidelines for Annotation Stage-1

This annotation process consists of two steps: (1) Dialogue Standardization: quality checking of legal consultation dialogues to correctly identify speaker roles, correct typos, and identify coded language; (2) Dialogue Intent Annotation: labeling dialogue intents.

1. Dialogue Standardization The dialogues are from live legal consultation scenarios, involving conversations between (1) clients and (2) lawyers. Clients present legal inquiries, and lawyers address these through questioning, confirmation, and advice. Since the data is transcribed by large language models, there may be errors in role assignments and speech recognition. This phase aims to check and correct dialogue quality.

Specific steps include:

1. Verify correct role assignment between lawyer and client:

- Case 1: All role assignments are incorrect - Fix all roles accordingly
- Case 2: Individual statement roles are incorrect - Identify and adjust misattributed statements
- Case 3: Single statements containing multiple speakers - Separate into distinct turns

2. Jargon and Euphemism Detection: Identify and convert substitute terms used to bypass platform filters into standard text. Example: "hat uncle" $\rightarrow$ "police officer" Note: All potential euphemisms and substitute terms will be listed for further assessment.

3. Correct transcription errors, including typos and homophone mistakes.

4. Ensure dialogue format:

- Combine consecutive statements by the same speaker
- Ensure the dialogue begins with client's initial query and ends with lawyer's advice
- Remove greeting/farewell phrases like "hello," "thanks," "goodbye"

2. Dialogue Intent Annotation Based on quality-checked dialogue data, label each utterance with one of these intent categories:

1. "Client Initial Query": Client's opening question (typically first statement)
2. "Client Response": Client's answers to lawyer's clarifying questions
3. "Client Information Addition": Client's voluntary provision of additional key information
4. "Client Need Extension": Client's new questions/needs arising during consultation
5. "Lawyer Clarifying Question": Lawyer's clarifying questions to gather more case information
6. "Lawyer Information Verification": Lawyer's repeated confirmation of crucial points
7. "Lawyer Legal Advice": Lawyer's legal recommendations based on case understanding
8. "Lawyer Emotional Support": Lawyer's emotional encouragement or comfort to client

9. "Invalid Exchange": Irrelevant dialogue not affecting final legal advice

10. "Other": Miscellaneous categories not covered above

E.2 Guidelines for Annotation Stage-2

Stage 2 annotation primarily involves (1) labeling atomic key facts information and (2) summarizing lawyer's advice.

1. Atomic Key Facts Information Labeling Since clients' initial inquiries often lack complete information, based on the Q&A process between clients and lawyers, Qwen-max has preliminarily extracted Atomic Key Facts List. Further extraction and verification are needed to compile a complete list of critical information that informed the lawyer's legal advice.

Requirements:

1. Extract key information points from client-lawyer dialogue. Keep expressions consistent with original text where possible.

2. Cross-check Atomic Key Facts List against dialogue points for completeness. Add missing information as needed.

3. Merge similar key fact information points to avoid redundancy. Atomic Key Fact Points should be atomic-level:

   Example:

   - ["Client", "They had a banquet that evening.", "Client Response"],
   - ["Lawyer", "What kind of banquet at friend A's house?", "Lawyer Clarifying Question"],
   - ["Client", "It was their child's first birthday.", "Client Response"],
   - ["Lawyer", "First birthday banquet. I see.", "Invalid Exchange"],
   - ["Client", "Then got very drunk that night.", "Client Information Addition"],
   - ["Lawyer", "A got drunk.", "Lawyer Information Verification"],
   - ["Client", "Yes, intoxicated.", "Client Response"]

   Corresponding key fact information point:
   "Friend A hosted child's first birthday banquet that evening, A got heavily intoxicated"

4. Supplement missing entity information (e.g., specific roles or event descriptions). Example: Change "was detained" to "Defendant A was detained"

5. Maintain chronological order of atomic facts consistent with dialogue flow.

6. Ensure extracted information comes from client statements only, excluding lawyer's advice. However, combine lawyer questions with client answers when relevant.

7. Rate importance of each atomic fact:

   - Critical Facts (3 points): Directly affects case classification, liability distribution, or rebuts prosecution's charges
   - Secondary Facts (2 points): Provides background or supporting information with moderate case impact
   - Non-critical Facts (1 point): Subjective opinions or procedural descriptions with no direct impact on fact-finding

2. Lawyer Advice Summary Legal advice may be scattered throughout the dialogue. Consolidate all advice while:

- Filtering out meaningless information
- Preserving lawyer's original key terminology to maintain accuracy and completeness
- Avoiding alterations to lawyer's original intent and critical information

F Dataset

F.1 Evaluation Metrics

We evaluate expert responses from two main perspectives. First, we assess **Clarification Capability** through *effectiveness* and *efficiency* metrics. Second, we evaluate **Advice Quality** using both *automated* and *LLM-based evaluation* metrics.

F.1.1 Clarification Capability

Effectiveness

Given ground truth atomic key facts list A_n and simulation dialogue, we use qwen-max to match how many key facts from A_n are mentioned in the user-lawyer consultation dialogue. The extraction prompt template is shown in Table 8.

We designed four metrics to evaluate clarification capability:

1. **Recall (Rec.)**: measures the comprehensiveness of key fact coverage, defined as:

$$\text{Rec.} = \frac{1}{|N|} \sum_{i=1}^{N} \frac{|\mathcal{P}_i \cap A_i|}{|A_i|} \tag{1}$$

where $\mathcal{P}_i$ represents the predicted key facts set for the i-th dialogue, A_i denotes the corresponding ground truth set, and N is the total number of dialogues in the evaluation set.

2. **Weighted Recall (Weighted Rec.)**: incorporates the significance of each key fact through importance weights, defined as:

$$\text{Weighted Rec.} = \frac{1}{|N|} \sum_{i=1}^{N} \frac{\sum_{j \in \mathcal{P}_i \cap A_i} w_j}{\sum_{k \in A_i} w_k} \tag{2}$$

where $w_j \in \{1, 2, 3\}$ denotes the importance weight assigned to each key fact j, $\mathcal{P}_i$ represents the predicted key facts set for the i-th dialogue, and A_i denotes the corresponding ground truth set. The weights reflect the relative significance of each key fact in the legal consultation context.

3. **Recall@5 (Rec@5)**: evaluates how effectively the first five questions elicit key facts, defined as:

$$\text{R@5} = \frac{1}{|N|} \sum_{i=1}^{N} \frac{|\mathcal{P}_i^5 \cap A_i|}{|A_i|} \tag{3}$$

where $\mathcal{P}_i^5$ represents the set of key facts identified in the first five questions of the i-th dialogue, and A_i denotes the complete ground truth set. This metric specifically assesses the model's ability to efficiently extract crucial information in the early stages of the consultation.

4. **Normalized Discounted Cumulative Gain (NDCG)**: evaluates the effectiveness of question ordering by considering both the importance of key facts and their position in the dialogue sequence, defined as:

$$\text{NDCG} = \frac{\text{DCG}}{\text{IDCG}} \tag{4}$$

where DCG (Discounted Cumulative Gain) is calculated as:

$$\text{DCG} = \sum_{i=1}^{n} \frac{rel_i}{\log_2(i+1)} \tag{5}$$

and IDCG (Ideal DCG) is:

$$\text{IDCG} = \sum_{i=1}^{n} \frac{rel_i^*}{\log_2(i+1)} \tag{6}$$

Here, rel_i represents the importance score of a key fact discovered at position i, and rel_i^* denotes the importance scores sorted in descending order for the ideal sequence. Each key fact is counted only once at its first appearance in the dialogue, with importance scores ranging from 1 to 3.

Efficiency

We measure the average number of dialogue turn rounds (**AT**).

$$\text{AT} = \frac{1}{|N|} \sum_{i=1}^{N} |T_i| \tag{7}$$

where $|T_i|$ represents the number of turns in the i-th dialogue, and N is the total number of dialogues in the evaluation set. A turn T is defined as one complete question-answer pair between the lawyer and user, excluding the initial user query and the final legal advice. This metric specifically measures the efficiency of the lawyer's clarification process during the consultation.

F.1.2 Advice Quality

We evaluate the semantic alignment between the generated and reference advice through both *automated metrics* and *LLM-based evaluation*.

Automated Metrics We employ two standard metrics:

- **ROUGE-L**: We calculate ROUGE-L (Lin, 2004) between reference and generated advice after Chinese word segmentation using jieba package [2].
- **BERTScore**: Leverages BERT contextual embeddings to compute semantic similarity (Zhang et al., 2020).

LLM-based Evaluation We assess five dimensions on a scale of 1-10:

- **Professionalism (Pro.)**:
  - Accurate understanding and relevant solutions
  - Clear explanation of complex legal concepts
  - Actionable recommendations

[2] https://pypi.org/project/jieba/

- **Fluency (Flu.)**:
  - Semantic coherence without logical errors
  - Consistency in style and content
  - Friendly and engaging response tone
- **Completeness (Com.)**:
  - Sufficient information and details
  - Coverage of essential recommendations
- **Satisfaction (Sat.)**:
  - Targeted and personalized solutions
  - Accessible language and expression
  - Empathy and respect for client concerns
- **Safety (Safe.)**:
  - Scientific and accurate legal knowledge
  - Prevention of potentially harmful advice
  - Adherence to professional ethics and non-discriminatory content

These five dimensions are aggregated into an **Overall Score (OA)** on a scale of 1-10, reflecting the comprehensive quality of the legal advice.

We employ GPT-4o-mini as an automated judge to evaluate the generated legal advice across five dimensions (1-10 scale). The specific prompting strategy for LLM-as-a-judge evaluation is detailed in Table 9.

G More Experimental Results

We provide more experimental results in Table 6 and Table 7.

H Prompt Template for LeCoDe Evaluation

We provide LeCoDe evaluation prompt template in Table 8 and Table 9.

Models		Effectiveness				Efficiency
		Rec.(%)	Weighted Rec.(%)	Rec@5(%)	NDCG(%)	AT ↓
Closed-source LLMs						
GPT-4	ZS	35.9	37.2	34.4	74.7	4.9
	FS	32.5	33.6	31.9	71.4	4.3
GPT-4o	ZS	33.0	33.9	25.9	63.2	10
	FS	33.6	34.2	27.9	65.6	10
Qwen-max	ZS	39.8	41.3	31.2	70.7	10
	FS	40.0	41.2	30.8	70.5	10
Qwen-turbo	ZS	27.3	28.3	23.2	61.2	9.4
	FS	25.7	26.8	22.7	63.5	9.8
Open-source LLMs						
deepseek-v3	ZS	28.0	29.0	25.1	65.4	7.5
	FS	32.3	33.4	27.1	65.3	9.1
deepseek-r1	ZS	27.6	28.8	25.1	63.9	6.3
	FS	32.0	33.2	26.2	64.1	8.4
Qwen2.5-72B	ZS	39.6	40.8	30.4	71.9	10
	FS	17.2	17.7	13.3	33.7	7.6
Qwen2.5-7B	ZS	31.1	32.1	27.7	70.1	10
	FS	24.7	25.4	22.7	64.1	10
Llama-3.1-8B	ZS	22.1	23.1	19.1	55	7.8
	FS	21.3	22	18.7	56.4	9.6
GLM4-32B	ZS	34.1	35.4	28.8	67.3	7.3
	FS	38.5	39.8	30.1	68.2	8.6
GLM4-9B	ZS	34.2	35.3	26.1	65.0	10
	FS	30.9	31.9	24.9	64.0	10
Domain-specific LLMs or Clarifying Strategy						
ChatLaw	ZS	16.7	17.4	16.3	47.8	4.4
Lawyer-LLaMA	ZS	21.8	22.9	21.8	62.9	4.2
Farui	ZS	28.4	29.5	25.1	64.2	8.9
MediQ	ZS	22.1	22.9	19.0	53.5	4.2
Direct SFT	ZS	37.2	37.6	25.6	62.3	9.6
Key-fact SFT	ZS	53.8	55.1	43.5	84.7	9.2
Key-fact-e SFT	ZS	51.0	51.1	45.3	84.8	6.6

Table 6: The results on Clarification Capability, comparing few-shot (FS) and zero-shot (ZS) strategies.

Models		Automated Metrics		LLM-Evaluaion Metrics					
		Rouge-L	BERTSscore	Pro.	Flu.	Com.	Sat.	Safe.	OA
Closed-source LLMs									
GPT-4	ZS	17.5	65.9	56.8	71.9	49.9	55.8	71.1	58.1
	FS	17.4	66.1	59.2	72.5	52.5	58.2	72.1	59.9
GPT-4o	ZS	13.0	62.9	50.1	69.6	44.7	52.4	66.6	53.5
	FS	16	64.6	54.8	72.7	49.4	56.3	69.9	57.5
Qwen-max	ZS	13.9	63.9	54.0	71.4	47.6	55.3	68.9	56.2
	FS	16.1	65.4	56.5	72.9	50.6	56.9	71.7	58.5
Qwen-turbo	ZS	10.6	61.8	44.8	62.7	40.0	47.6	61.6	48.5
	FS	13.5	63.6	53.3	70.1	47.1	53.4	69	55.3
Open-source LLMs									
deepseek-v3	ZS	16.2	65.3	57.6	72.9	50.8	56.2	72.1	58.9
	FS	16.5	65.6	58.7	73.7	51.9	57.5	72.9	59.7
deepseek-r1	ZS	12.4	64	63.8	74.1	55.5	55.9	76.5	62.2
	FS	13.4	64.8	66.7	75	58.8	58.2	78.3	64.6
Qwen2.5-72B	ZS	14.7	63.8	54	71.6	48.5	55.2	69.4	56.6
	FS	10.1	57.7	33.6	42.7	30.4	33.8	42.5	34.7
Qwen2.5-7B	ZS	9.2	60.0	42.3	60.4	37.6	45.0	57.8	45.7
	FS	9.2	59.1	40.1	57	34.9	42.3	55.3	42.9
Llama-3.1-8B	ZS	13.9	63.1	51.7	69.1	46.5	51.7	67.1	54.1
	FS	7.8	54.8	32.6	44.2	29.8	34.3	45.3	34.4
GLM4-32B	ZS	14.3	64.1	55.6	70.4	49.0	54.3	70.4	56.7
	FS	14.8	64.7	57.6	72.4	51	56.1	72.3	58.8
GLM4-9B	ZS	13.9	63.1	51.7	69.1	46.5	51.7	67.1	54.1
	FS	14.1	63.1	52.2	69.9	47.3	53.1	68	54.7
Domain-specific LLMs or Clarifying Strategy									
ChatLaw	ZS	7.8	53.0	28.8	40.0	25.5	31.0	39.4	30.7
Lawyer-LLaMA	ZS	12.4	59.7	37.6	52.7	33.1	39.0	51.8	39.4
Farui	ZS	16.6	64.5	52.2	66	47.4	51.3	66.3	53.8
MediQ	ZS	13.3	63.7	51.8	69.1	45.5	51.9	66.7	54.0
Direct SFT	ZS	15.4	64.9	51.7	65.3	45.6	50.1	63.1	52.4
Key-fact SFT	ZS	18.5	67.5	59.3	73.0	52.3	55.9	71.2	59.4
Key-fact-e SFT	ZS	19.1	67.6	59.3	73.6	52.0	56.7	72.1	59.5

Table 7: The results on Advice Quality, comparing few-shot (FS) and zero-shot (ZS) strategies.

Instruction Template used to extract matching key fact from simulated dialogue.

Analyze key information mentioned in user-lawyer consultation dialogues and match them with a given list of Atomic Key Facts.

Input data:
1. User-Lawyer Dialogue (user lawyer dialogue):
- Format: Dialogue ID: 'q': question, 'a': answer
- Each dialogue includes lawyer's question and user's response

2. Atomic Key Facts List (Atomic Key Facts li):
- Format: index: information content
- Index starts from 0

Task requirements:
1. Identify Atomic Key Facts mentioned in the dialogues
2. Match identified information with items in the Atomic Key Facts list
3. Record the index numbers of matching information

Output format requirements:
- Output should be a Python dictionary
- Keys: Dialogue ID (e.g., "QA0", "QA1"...)
- Values: List of Atomic Key Facts indices mentioned in that dialogue
- If no Atomic Key Facts is mentioned, return an empty dictionary
- Example format:
```python
{"QA0":[0,1],"QA1":2,...}
```

Below are the consultation dialogue and Atomic Key Facts list:
[Consultation Dialogue]
user lawyer dialogue

[Atomic Key Facts]
{Atomic Key Facts}

Table 8: The Instruction Template for extraction matching key fact from simulated dialogue, to evaluate Clarification Capability (translated from Chinese).

Instruction Template used in Legal Advice Quality Evaluation, to evaluate Advice Quality.
You are an experienced legal expert, specializing in evaluating the quality of legal advice in consultations. Please act as a fair and rigorous evaluator to objectively assess the quality of legal advice provided by the AI assistant. In your assessment, please analyze and score based on the following five key dimensions: 1. Professionalism - Accurately understand client's issues and provide relevant answers - Clearly explain complex legal knowledge - Actionability of advice 2. Fluency - Ensure semantic coherence without logical errors or irrelevant information - Maintain consistency in style and content - Maintain friendly and enthusiastic response attitude 3. Completeness - Provide sufficient information and details - No omission of important advice 4. Satisfaction - Targeted responses, providing personalized solutions - Use approachable expressions, language affinity - Focus on client experience and feelings, empathy and respect for client needs 5. Safety - Provide scientific and accurate legal knowledge - Ensure client safety; reject potentially harmful information or advice - Follow professional ethics and respect client choices; must not contain discriminatory, biased, or harmful content We will provide the following materials: high-quality reference answers and legal advice written by the AI assistant. When starting the evaluation, you need to follow these reasoning steps: 1. Compare the AI assistant's legal advice with the reference answer, point out deficiencies in the AI's answer, and explain in detail 2. Evaluate the AI assistant's legal advice according to the above dimensions, scoring 1-10 for each dimension 3. Based on the scores of each dimension, derive a comprehensive score (1-10) for the AI assistant 4. Your scoring should be as strict as possible and must follow these scoring rules: higher quality responses receive higher scores Scoring criteria: 1-2 points: Advice contains irrelevant content, serious errors, unverified or false information, or potentially harmful content 3-4 points: No major errors but clear deficiencies in legal relationship definition or key point responses, poor logical reasoning, lacks specificity, advice too general, fails to meet basic consultation requirements 5-6 points: Basically meets consultation requirements, accurate legal analysis but average logical completeness, addresses main points but lacks deep analysis, overall performance is moderate 7-8 points: Quality approaches reference answer, provides practical solutions, excellent performance in all evaluation dimensions, no obvious defects 9-10 points: Quality significantly exceeds reference answer, near-perfect performance in all dimensions, provides extremely valuable solutions Please provide detailed evaluation notes. For each dimension score, explanations must be provided. All scores should be whole numbers. Finally, return the evaluation results in the following dictionary format: python{{'Professionalism': score, 'Fluency': score, 'Completeness': score, 'Satisfaction': score, 'Safety': score, 'Overall Score': total}} <dialogue> {dialogue} </dialogue> <gt suggestion>{gt suggestion}</gt suggestion> <pred suggestion>{pred suggestion}</pred suggestion> Please begin the evaluation:

Table 9: The Instruction Template for Legal Advice Quality Evaluation, to evaluate Advice Quality (translated from Chinese).