The Missing Premise: Evaluating and Guiding Legal Consultation Agents When Client Narratives Are Incomplete

Anonymous ACL submission

Abstract

Legal consultation is a challenging legal task that requires eliciting material facts from incomplete client narratives before advice can be given. Existing legal benchmarks typically assume that the facts required for reasoning are available, while overlooking fact elicitation in real-world consultation. To address this gap, we introduce FILC-Bench, a dynamic benchmark designed to reflect the incomplete and evolving nature of legal consultation. Built from real cases, FILC-Bench contains 486 consultation instances and 1,311 interactions across low, medium, and high levels of narrative incompleteness, allowing evaluation of both consultation behavior and advice quality. Extensive experiments across various LLMs and agent frameworks reveal that current methods struggle to assess factual sufficiency and recover missing facts, resulting in legal advice without sufficient factual support. Following lawyers' practice, we propose the Sufficiency-Guided Consultation Agent (SGCA). It tracks the consultation state, clarifies material unresolved facts, and reviews drafts from the opposing counsel perspective to produce advice supported by the record. Across three incompleteness levels, SGCA with Qwen3 32B improves the average score from 2.12 to 4.07, demonstrating the effectiveness of sufficiency-guided consultation under incomplete legal narratives.

1 Introduction

As Large Language Models (LLMs) are increasingly integrated into legal service workflows, legal consultation has emerged as an important application scenario for LLM-based legal agents (Li et al., 2025c; Wu et al., 2024; Sun et al., 2024; Fei et al., 2025; Lai et al., 2024). Typically, in legal consultation, clients first describe their disputes or personal situations, and agents are then expected to assess the initially disclosed facts, identify potential legal issues, and provide preliminary legal advice. Through this process, clients can better understand their legal position, thereby identifying legally appropriate ways to protect their interests.

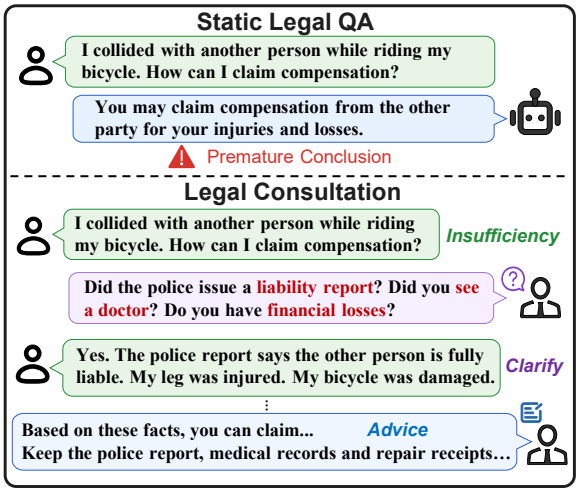

Figure 1: Static Legal QA versus legal consultation with incomplete client narratives. Static Legal QA may produce premature conclusions, whereas legal consultation requires clarification of missing material facts.

As one of the central scenarios for legal assistance, legal consultation has been adopted by many benchmarks to evaluate the domain capabilities of LLM-based agents (Guha et al., 2023; Fei et al., 2024; Li et al., 2024, 2025b; Jia et al., 2026). Early benchmarks, as shown in Figure 1, often follow a static Legal QA format, which overlooks the dynamic process of consultation (Guha et al., 2023; Fei et al., 2024; Li et al., 2024). Recent works move toward more realistic legal assistance by introducing multi-turn interaction between clients and agents over a provided factual record (Li et al., 2025b; Jia et al., 2026). In practice, however, clients often begin with incomplete narratives, as limited legal expertise and their personal view of the dispute may lead them to omit material facts (Yuan et al., 2025; Liu et al., 2023). Existing benchmarks, as summarized in Table 1,

Bench.	MT	Dyn.	Elicit.	Suff.	Eval.
LegalBench	×	×	×	×	Out.
LawBench	×	×	×	×	Out.
LexEval	×	×	×	×	Out.
LexRAG	✓	×	×	×	Out.
J1-LC	✓	✓	×	○	Out.
FILC-Bench	✓	✓	✓	✓	**Proc.& Out.**

Table 1: Comparison with representative legal benchmarks. MT/Dyn./Elicit./Suff. denote multi-turn, dynamic interaction, information elicitation, and fact sufficiency management; Eval. denotes outcome only (Out.) or process and outcome (Proc.&Out.) evaluation; ✓ / ○ / × indicate full/partial/no support.

give limited attention to eliciting missing material facts from clients and assessing whether agents can judge factual sufficiency before giving advice. Moreover, their evaluation is often centered on final outcomes, making it difficult to assess how agents manage the consultation process.

To address this gap, we introduce the **F**actual **I**ncompleteness in **L**egal **C**onsultation Benchmark (**FILC-Bench**), a dynamic benchmark designed to evaluate whether legal agents can work from incomplete client narratives. Built from real, publicly available civil cases, FILC-Bench controls initially visible facts and facts revealed during consultation. It contains 486 consultation instances and 1,311 interactions under low, medium, and high narrative incompleteness. To evaluate whether agents can manage missing facts while preserving legal reliability, FILC-Bench uses five metrics: Factual Grounding, Sufficiency Management, Legal Soundness, Issue Handling and Client Utility. The first two metrics focus on the consultation process, while the latter three evaluate the quality and practical value of the final advice. This design allows us to evaluate how agents manage the consultation process, rather than only judging a final answer produced from a complete input. As a result, FILC-Bench evaluates agents under conditions that are closer to real-world legal consultation.

Building on this consultation setting, we further propose the **S**ufficiency-**G**uided **C**onsultation **A**gent (**SGCA**). It combines explicit consultation state tracking with Sufficiency-Guided Inquiry (SGI) and Adversarial Counsel Review (ACR). Inspired by how lawyers keep track of a client's evolving factual account, SGCA maintains a consultation state that records known facts, unresolved facts, and confirmed absences throughout the dialogue. For each request, SGCA checks this state to decide whether unresolved facts may affect the legal conclusion. When such facts remain material, Sufficiency-Guided Inquiry (SGI) asks targeted clarification questions and updates the state with the client's response. Considering the adversarial nature of legal disputes, SGCA further introduces Adversarial Counsel Review (ACR). It reviews the draft from the opposing counsel perspective and identifies unsupported facts, evidence gaps, or defense risks. SGCA then revises the draft into advice grounded in the consultation record, while stating any remaining uncertainty clearly.

Extensive experiments on FILC-Bench show that legal consultation remains challenging when agents answer from the initial narrative without explicit mechanisms for managing missing facts. Even GPT 5.4 receives only 2.78 for handling issues under high narrative incompleteness. With Qwen3 32B as the backbone, SGCA raises the average scores from 2.12 to 4.07 across all three incompleteness levels. These findings show that FILC-Bench exposes a capability gap in current legal agents, while SGCA demonstrates the value of explicit missing fact management.

Our contributions are summarized as follows: (1) We introduce FILC-Bench, a benchmark for legal consultation under incomplete client narratives that controls fact omission and disclosure during consultation using real cases. (2) We propose SGCA, a consultation agent that combines explicit fact state tracking, sufficiency guided clarification, and review from the opposing counsel perspective. (3) Extensive experiments on FILC-Bench show that existing models and agentic methods still struggle under incomplete client narratives, while SGCA demonstrates the effectiveness of designing consultation agents around fact sufficiency.

2 Related Work

Legal Benchmarks. Legal NLP benchmarks have long provided standardized settings for evaluating legal understanding and assistance capabilities (Guha et al., 2023; Fei et al., 2024; Li et al., 2024; Hwang et al., 2022; Chalkidis et al., 2022; Niklaus et al., 2023). For example, LegalBench (Guha et al., 2023) collects expert designed reasoning problems, while LawBench (Fei et al., 2024) evaluates legal knowledge and application across a broader set of tasks. LexEval (Li et al., 2024) further organizes legal tasks along increasing levels of difficulty from basic knowledge to

more complex reasoning. Recent benchmarks have also moved closer to agent settings and practical legal assistance scenarios, with LexRAG (Li et al., 2025c) evaluating retrieval augmented generation in multi-turn legal consultation and Ready Jurist One (Jia et al., 2026) introducing dynamic environments for procedural legal tasks. Despite this progress, these benchmarks generally assume that the material facts required for legal reasoning are already fully available, giving limited attention to legal consultation in practice, where agents must work from incomplete client narratives. FILC-Bench addresses this gap by controlling the information available in the initial client narrative and the facts revealed during consultation, thereby simulating how lay clients may enter consultation before all material facts are disclosed.

LLM-Based Legal Agents. Agentic methods have been used to improve LLM performance on complex tasks (Wei et al., 2022; Wang et al., 2023; Shinn et al., 2023). Some studies improve agent capability through interactive agent design (Mou et al., 2025; Yue et al., 2025b; Li et al., 2023). In the legal domain, agents have been applied to a range of practice scenarios (Zhang et al., 2025; Liao et al., 2026; Yang et al., 2026; Sun et al., 2024; Yue et al., 2025a). For example, LawThinker (Yang et al., 2026) integrates multiple legal tools and verifies results to improve agent performance across dynamic legal scenarios. LawLuo (Sun et al., 2024), specially designed for legal consultation, uses multiple agents to simulate different roles within a law firm. These systems show the potential of agentic methods for complex legal assistance, yet they often assume that the material facts required for legal reasoning are already available. In contrast, our work focuses on the earlier consultation stage, where the facts needed for legal advice must be established before advice generation.

3 FILC-Bench

In this section, we formalize legal consultation with incomplete client narratives, describes the construction of FILC-Bench from civil judgments, and defines an evaluation protocol that measures both the consultation process and the final advice.

3.1 Task Formulation

We formulate legal consultation under incomplete client narratives as a dialogue over multiple turns grounded in a structured case record. Let $c \in \mathcal{C}$ denote a structured case record that contains the full case facts and the reference judgment. Each record c is associated with an ordered sequence of consultation topics $\mathcal{Q}_c = (q_1, \ldots, q_n)$, with q_j denoting a client's request that the lawyer agent must handle within the same consultation.

For an incompleteness level ℓ, the client facts in c are partitioned into visible facts V_c^ℓ and omitted facts O_c^ℓ. The opening narrative is generated from V_c^ℓ, so the lawyer agent begins with only a partial factual description. Given the current topic q_j and the dialogue history, the agent must reason over the evolving consultation record and provide a legal response that addresses the topic in light of the facts currently available in the dialogue.

An evaluation instance is indexed by (c, ℓ). The output is the completed consultation history

$$\mathcal{H}_c^\ell = (H_1, \ldots, H_n),$$

where H_j denotes the interaction history for topic q_j. Evaluation is conducted over $\mathcal{H}_c^\ell$.

3.2 Benchmark Construction

To instantiate this task, we construct FILC-Bench from civil judgments. Figure 2 provides a summary of the construction pipeline. We first collect civil judgments from China Judgments Online, a widely used public source of Chinese court judgments (Zhong et al., 2020). For each case, we use LLMs to extract case facts, client requests, and judgment text. The extracted fields are manually checked, and cases with unclear facts or limited suitability for consultation are removed.

For a given structured record c, we construct instances under different incompleteness levels ℓ. The client facts are split into visible facts V_c^ℓ and omitted facts O_c^ℓ, with low, medium, and high narrative levels varying in the amount of legally relevant information withheld. An LLM then rewrites V_c^ℓ into the opening client narrative N_c^ℓ, while O_c^ℓ is only kept for possible disclosure during the dialogue. The client requests are rewritten into ordered consultation topics $\mathcal{Q}_c$, and the judgment text is condensed into the reference judgment J_c.

Finally, we manually review all instances to ensure that they remain faithful to the source case and plausible as legal consultations. Each instance is represented as $I_c^\ell = (N_c^\ell, O_c^\ell, \mathcal{Q}_c, J_c)$.

3.3 Interaction Protocol

Given an instance $I_c^\ell = (N_c^\ell, O_c^\ell, \mathcal{Q}_c, J_c)$, we simulate a multi-turn consultation between a lawyer

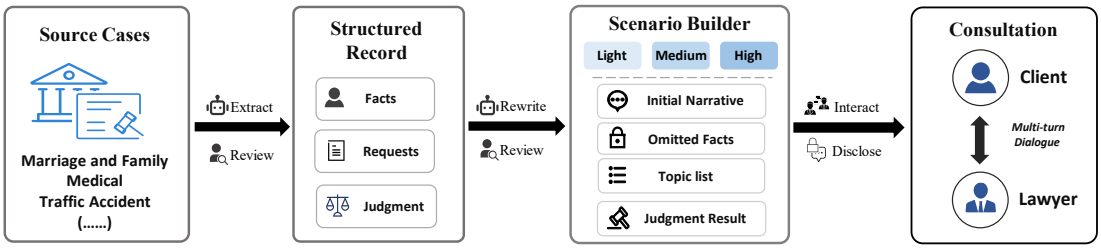

Figure 2: Construction pipeline of FILC-Bench. Civil judgments are converted into structured records, from which consultation instances are created with different incompleteness levels. Each instance is used to run a multi-turn consultation with controlled disclosure of omitted facts.

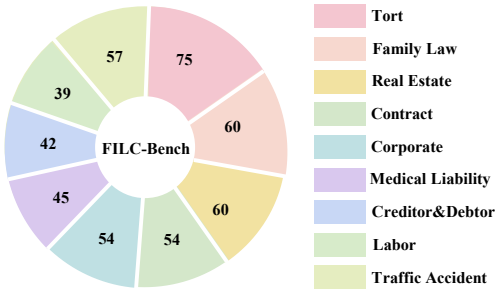

Figure 3: Category distribution of FILC-Bench instances across three levels of narrative incompleteness.

agent and a client agent. The client agent is driven by an LLM and follows the predefined topic order and the fact disclosure rule. It first presents the opening client narrative N_c^ℓ, and then raises the consultation topics in $\mathcal{Q}_c = (q_1, \ldots, q_n)$ in order. The client holds the omitted facts O_c^ℓ, but does not reveal them proactively. During topic q_j, let $l_{j,t}$ be the lawyer utterance at turn t, and $R_{j,t}$ denote the omitted facts already revealed before this turn. The client agent then checks whether $l_{j,t}$ asks about any unrevealed fact in $O_c^\ell \setminus R_{j,t}$.The recoverable omitted facts at this turn are

$$M_{j,t} \subseteq O_c^\ell \setminus R_{j,t}.$$

The next client response is determined as

$$u_{j,t+1} = \begin{cases} \text{Reveal}(M_{j,t}), & M_{j,t} \neq \emptyset, \\ \text{Respond}(H_{j,t}), & \text{otherwise}, \end{cases}$$

where $H_{j,t}$ denotes the dialogue history before the response. If $M_{j,t}$ is nonempty, the client agent reveals the corresponding facts. Otherwise, it responds based on $H_{j,t}$ without adding omitted facts. After topic q_j is completed, the client agent moves to the next topic when available. The consultation continues until all topics in $\mathcal{Q}_c$ are handled.

3.4 Evaluation Protocol

FILC-Bench evaluates each instance over both the completed consultation history $\mathcal{H}_c^\ell$ and the final advice. This design captures how the lawyer agent manages the consultation process, since the final advice may still be unreliable when important material facts were never confirmed. To make this process measurable, we use five scoring metrics, with three focused on the final advice and two focused on the consultation process.

The final advice is evaluated by **Legal Soundness (LS)**, **Issue Handling (IH)**, and **Client Utility (CU)**. LS measures whether the advice applies relevant legal rules correctly and reaches legally reliable conclusions. IH measures whether the advice addresses the relevant disputes, claims, defenses, and risks. CU measures whether the advice helps the client better understand the situation and possible next steps. LS and IH are assessed with reference to the judgment J_c, while CU focuses on the usefulness of the advice to the client.

The consultation process is evaluated by **Factual Grounding (FG)** and **Sufficiency Management (SM)**. FG measures whether the agent stays within the facts available in the dialogue and avoids unsupported assumptions. SM measures whether the agent identifies missing material facts, asks necessary clarification questions, and limits conclusions when the factual record remains incomplete.

Each metric is scored on a 0–5 scale using automatic LLM scoring methods, with reference to J_c, V_c^ℓ, and O_c^ℓ. Section 5.5 reports the validation results of the automatic scoring, and Appendix D provides the full scoring instructions.

3.5 Benchmark Statistics

FILC-Bench contains 486 dynamic consultation instances and 1,311 topic interactions across low, medium, and high levels of narrative incompleteness. As shown in Figure 3, these cases cover 9 representative legal categories and encompass common disputes in daily legal practice. Across incompleteness levels, instances are constructed under the same topic criteria. The opening narrative and

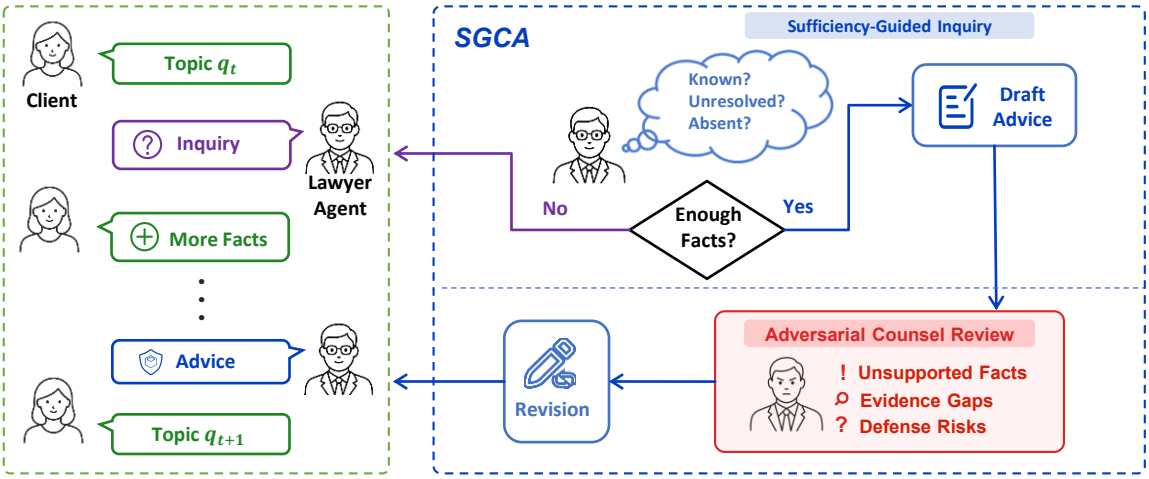

Figure 4: Overview of SGCA. For each consultation topic, SGCA tracks the consultation state, checks factual sufficiency, uses SGI to clarify material gaps, and applies ACR to review the draft from an opposing counsel perspective. The final advice is grounded in the consultation record and makes remaining uncertainty explicit.

omitted facts vary according to the assigned level. The benchmark includes 1,169 omitted facts in total, with more material facts withheld as narrative incompleteness increases. Appendix B.2 reports further details on dataset statistics and distributions.

4 Sufficiency-Guided Consultation Agent

Inspired by how lawyers handle cases in real-world legal practice, we design SGCA, a lawyer agent framework for legal consultation under incomplete client narratives. For each consultation topic q_t, it first determines whether the disclosed facts are sufficient for a bounded legal response before drafting advice. As shown in Figure 4, SGCA first uses the current fact state to steer the consultation. When the record is still insufficient, SGI turns unresolved material facts into targeted clarification questions; once the available facts can support a qualified analysis, SGCA drafts advice and then uses ACR to revise claims that are unsupported or overextended. We describe each step below.

4.1 Sufficiency-Guided Inquiry

SGI is the core component of SGCA. It maintains a fact state $S = (K, U, A)$ across the ordered topics to decide what should be asked before advice is drafted. Here, K, U, and A stand for Known facts stated by the client, Unresolved facts that may affect the current topic, and facts confirmed as Absent. By carrying the updated state across topics, SGI keeps advice grounded in facts confirmed during the consultation. At the start of the consultation, K is initialized from the opening client narrative N_c^{ℓ}, while U and A are empty. This state separates what is already known from what needs confirmation, allowing clarification to focus on facts that may change the legal assessment.

For a topic q_j, SGI first checks the current state and adds to U the missing facts that may affect the answer. It then asks targeted clarification questions about the facts in U. After each client response, SGI updates the state by adding newly disclosed facts to K, recording confirmed absences in A, and keeping unanswered or unclear facts in U. This inquiry loop continues until U contains no unresolved facts for the current topic. SGCA then drafts a preliminary answer based on the updated state, which is passed to ACR for further examination.

4.2 Adversarial Counsel Review

Legal disputes are shaped by adversarial positions (Prakken, 2004). Advice drafted from the client's narrative may still reflect a one-sided view of the case, leading the agent to overlook potential risks. ACR is therefore designed to challenge the draft from an opposing perspective.

After SGI completes the clarification loop, SGCA drafts preliminary legal advice from the updated fact state. ACR then reviews the draft from the opposing counsel perspective and produces a structured rebuttal on behalf of the other party. The rebuttal identifies weak factual support, evidentiary gaps, and requests that may be difficult to sustain. SGCA uses it to narrow unsupported conclusions and mark specific legal risks, keeping the final advice within safer boundaries.

Methods	Low						Medium						High					
	LS	IH	FG	SM	CU	Avg.	LS	IH	FG	SM	CU	Avg.	LS	IH	FG	SM	CU	Avg.
Direct Reasoning																		
Large-scale LLMs																		
GPT-5.4	4.52	2.78	4.65	3.69	4.00	3.93	4.51	2.80	4.48	3.65	3.81	3.85	4.62	2.78	4.14	3.66	3.62	3.76
DeepSeek-V4	4.58	2.41	4.37	3.57	3.51	3.69	4.55	2.04	3.94	3.52	3.29	3.47	4.47	2.05	3.31	3.55	3.10	3.30
Legal-specific LLMs																		
ChatLaw2-7B	2.09	0.09	1.93	1.91	1.80	1.56	2.01	0.03	1.81	1.97	1.74	1.51	2.04	0.01	1.90	1.94	1.51	1.48
LawLLM-7B	1.76	0.00	1.78	1.89	1.23	1.33	1.86	0.01	1.70	1.92	0.98	1.29	1.83	0.00	1.74	1.94	1.06	1.31
General LLMs																		
GLM-4-9B	2.22	0.02	1.99	1.93	1.75	1.58	2.17	0.02	2.01	1.91	1.73	1.57	2.35	0.02	1.98	1.98	1.60	1.59
Ministral-8B	2.10	0.09	1.96	1.84	2.41	1.68	2.08	0.04	1.86	2.00	2.17	1.63	2.17	0.01	1.80	1.95	2.17	1.62
Qwen3-14B	2.82	0.34	3.03	1.03	2.40	1.92	2.77	0.35	2.70	1.10	2.39	1.86	2.67	0.27	2.49	1.30	2.11	1.77
Qwen3-32B	3.21	0.54	3.50	1.26	2.64	2.23	3.19	0.51	3.12	1.28	2.44	2.11	3.19	0.40	2.74	1.50	2.33	2.03
LLM-based Agent (with Qwen3-32B)																		
Reflection	3.42	0.63	3.56	1.34	2.98	2.39	3.31	0.56	3.46	1.38	2.87	2.32	3.24	0.46	2.85	1.72	2.62	2.18
Plan-and-Solve	3.20	0.62	3.41	1.57	2.71	2.30	3.25	0.56	3.29	1.62	2.66	2.28	3.18	0.48	2.75	1.72	2.47	2.12
LawLuo	3.81	1.54	3.58	3.54	3.10	3.12	3.76	1.44	3.52	3.37	3.04	3.03	3.71	1.50	3.60	3.48	2.95	3.05
SGI	3.31	1.28	4.38	3.43	3.25	3.13	3.20	1.48	4.31	3.71	3.21	3.18	3.24	1.37	4.51	3.90	3.17	3.24
SGCA	**3.98**	**3.33**	**4.65**	**3.98**	**4.41**	**4.07**	**3.93**	**3.22**	**4.70**	**4.11**	**4.28**	**4.05**	**3.74**	**3.28**	**4.74**	**4.49**	**4.17**	**4.08**

Table 2: Main results by narrative incompleteness level and evaluation metric. LS, IH, FG, SM, and CU denote Legal Soundness, Issue Handling, Factual Grounding, Sufficiency Management, and Client Utility, respectively. Each metric is scored on a 0–5 scale, where 5 indicates the strongest performance.

5 Experiments

5.1 Experimental Setup

We evaluate all methods on FILC-Bench, a controlled benchmark for legal consultation with incomplete client narratives. The direct reasoning experiments examine whether current LLMs can provide reliable legal advice when material facts are omitted from the opening client narrative, while the agentic method experiments test whether consultation strategies can improve quality under the same disclosure protocol. All methods are evaluated across low, medium, and high levels of narrative incompleteness. The Client Agent is driven by Qwen3 32B and kept fixed across all experiments, and all compared systems act only as Lawyer Agents in the same interaction environment.

For direct reasoning, we use the same consultation prompt for large-scale general models, open general models, and legal-domain models, including GPT 5.4 (OpenAI, 2026), DeepSeek V4 (DeepSeek-AI, 2026), GLM4 9B (GLM Team, 2024), Ministral 8B (Mistral AI, 2024), Qwen3 14B/32B (Yang et al., 2025), ChatLaw2 7B (Cui et al., 2023), and LawLLM 7B (Shu et al., 2024). For agentic methods, we use Qwen3 32B as the shared backbone, including Reflection (Shinn et al., 2023), Plan-and-Solve (Wang et al., 2023), LawLuo (Sun et al., 2024), SGI, and SGCA. Large-scale proprietary models are accessed through official APIs, while the remaining models are served with vLLM using two NVIDIA A100 GPUs in our experiments. More detailed experimental settings are provided in Appendix C.5.

We evaluate completed consultation histories rather than isolated final answers. Following Section 3.4, each history is scored on five 0–5 evaluation metrics: Legal Soundness (LS), Issue Handling (IH), Factual Grounding (FG), Sufficiency Management (SM), and Client Utility (CU). DeepSeek V4 is used as the automatic judge for the main evaluation, and Section 5.5 validates its agreement with human scores.

5.2 Main results

Table 2 shows that FILC-Bench exposes a gap between producing legally plausible answers and managing unresolved facts during consultation. In the direct reasoning setting, even strong models show uneven performance across legal quality and process-sensitive dimensions. Under high narrative incompleteness, GPT 5.4 reaches 4.62 on Legal Soundness but only 2.78 on Issue Handling. This gap suggests that a model can analyze the disclosed factual record plausibly while still failing to address issues that depend on missing or unresolved facts. A similar pattern appears for Qwen3 32B: its average score decreases from 2.23 in the low setting to 2.03 in the high setting, while Issue Handling falls from 0.54 to 0.40. These results indicate that direct

Methods	IH	SM	CU	Avg.
Direct Reasoning	0.48	1.35	2.47	2.12
+ Suff.	1.76	2.65	2.99	2.87
+ Suff. + SGI	1.38	3.68	3.21	3.18
+ Suff. + ACR	2.49	3.63	3.37	3.49
⊕ SGCA	**3.28**	**4.19**	**4.29**	**4.07**

Table 3: Ablation results averaged over all levels of narrative incompleteness. Suff. denotes the sufficiency check over the current consultation state. SGI adds fact state guided clarification, and ACR adds review from the opposing counsel perspective. ⊕ SGCA denotes the full configuration. Avg. is computed over all five evaluation metrics.

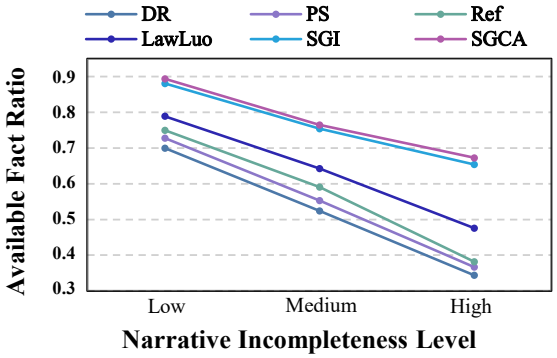

Figure 5: Final available fact ratio across levels of narrative incompleteness. The ratio measures the share of client side facts available to the agent after consultation, including facts visible in the opening narrative and omitted facts disclosed through clarification. DR denotes Direct Reasoning; PS, Plan and Solve; Ref, Reflection.

reasoning tends to rely on the initially disclosed facts and may leave material gaps unhandled when the client narrative is incomplete.

Current agentic methods only partly address this limitation. Reflection and Plan-and-Solve add internal reasoning steps, but their gains remain limited because they do not explicitly recover omitted material facts or maintain factual sufficiency. LawLuo performs better, reaching about 3.0 on average, which suggests that legal role-based interaction can help consultation quality. However, it still lags behind SGCA, indicating that multi-agent role simulation alone is not sufficient for managing missing facts and factual boundaries.

SGI improves factual grounding and sufficiency management through targeted clarification, and the full SGCA achieves the best performance at every incompleteness level. Under high narrative incompleteness, SGCA increases Issue Handling from 0.40 to 3.28 and Sufficiency Management from 1.50 to 4.49 compared with the Qwen3 32B backbone. These gains indicate stronger handling of unresolved issues and clearer control over factual boundaries, which are central to reliable legal consultation under incomplete client narratives.

5.3 Ablation Study

Table 3 reports how each component contributes to SGCA under incomplete client narratives. Suff. denotes the configuration that only applies the sufficiency check over the current consultation state. Suff. improves the average score from 2.12 to 2.87 and Sufficiency Management from 1.35 to 2.65, indicating that explicitly checking factual sufficiency already helps the agent avoid treating an incomplete narrative as fully answerable. Adding SGI further raises the average score to 3.18, with larger gains on Sufficiency Management and Client Utility. The improvement is consistent with SGI's role in maintaining an explicit fact state and asking targeted clarification questions over missing material fact dimensions. However, SGI alone still has limited Issue Handling, suggesting that recovering facts does not by itself address disputed issues, weak claims, or litigation risks.

The ACR variant complements SGI by improving Issue Handling from 1.38 to 2.49 and the average score from 3.18 to 3.49, suggesting that opposing counsel review helps expose contested issues and weak factual support. The full SGCA achieves the best results on all reported metrics. These results show that clarification and adversarial review address complementary parts of consultation with incomplete facts. Complete ablation results by level are reported in Appendix C.

5.4 Clarification Efficiency

Clarification efficiency examines whether additional questions make more client side facts available for advice, while keeping the visible interaction burden reasonably controlled for clients. Figure 5 reports the final available fact ratio, defined as the proportion of client side facts available to the agent by the end of the consultation. Methods without targeted clarification mainly rely on the opening narrative, so their available factual basis decreases as more facts are omitted. For example, Direct Reasoning falls from 0.700 in the low setting to 0.344 in the high setting. In contrast, SGI and SGCA make more omitted facts available through targeted clarification, raising the ratio to 0.654–0.673 in the high setting and to nearly 0.9

Method	Turns	Calls	Score
Direct Reasoning	7.84	3.42	2.12
Reflection	7.84	8.26	2.30
Plan-and-Solve	7.84	5.84	2.23
LawLuo	12.17	7.67	3.05
SGI	12.77	9.58	3.18
SGCA	12.77	14.97	4.07

Table 4: Interaction cost and overall performance. Turns denote dialogue length visible to the client, Calls denote LLM invocations per case, and Score is the average over five evaluation metrics.

Metric	IAA	ρ	MAE
LS	0.837	0.883	0.365
IH	0.861	0.887	0.378
FG	0.840	0.834	0.344
SM	0.901	0.909	0.325
CU	0.723	0.776	0.533
Avg.	0.832	0.858	0.429

Table 5: Validation of DeepSeek V4 judge scores against human scores. IAA denotes Krippendorff's alpha after adjudication, ρ denotes Spearman correlation with the human reference scores, and MAE denotes mean absolute error.

in the low setting. This pattern helps explain the main results, since stronger consultation quality depends on a larger factual basis for advice and clear boundaries around unresolved facts.

Table 4 reports client visible turns, LLM calls per case, and overall average score. Turns indicate the interaction burden perceived by client, while Calls reflect the computational cost of running the agent. General agentic methods all keep the same visible dialogue length at 7.84 turns, since they do not ask targeted clarification questions. However, Reflection and Plan and Solve increase LLM calls by 4.84 and 2.42 over Direct Reasoning, while improving the score by only 0.18 and 0.11 points. This suggests that additional internal reasoning alone brings limited benefit in this setting.

SGI increases visible turns by 4.93 over Direct Reasoning and improves the average score by 1.06 points. SGCA keeps nearly the same visible dialogue length as SGI, but further raises the average score by 0.89. Its higher system cost comes from additional LLM calls, increasing from 9.58 to 14.97, because ACR reviews the draft before final advice is produced. These results show that SGCA does not improve performance by placing more interaction burden on the client. Instead, it uses the clarified facts more effectively through internal review, with additional computation on the system side. Detailed costs and revealed omitted facts by incompleteness level are reported in Appendix C.

5.5 LLM Judge Validation

The main evaluation uses DeepSeek V4 as the LLM judge across the five metrics. LLM based evaluation is scalable, but it may introduce judge preferences (Li et al., 2025a). We validate its alignment with human scores on a stratified subset of 120 histories, covering different methods and narrative incompleteness levels. Three annotators independently score the histories using the rubric in Section 3.4. They are blind to method identity, model scores, and aggregate results. When the maximum disagreement exceeds two points, adjudication is applied before forming the human reference score.

Table 5 reports the alignment between DeepSeek V4 and human reference scores. IAA measures agreement among human annotators, Spearman's ρ measures rank correlation with the human reference, and MAE measures absolute scoring error. The results show that the rubric yields stable human references after adjudication and that the LLM judge provides reliable relative comparisons across system outputs. Client Utility shows lower alignment because the greater subjectivity of helpfulness judgments, its scores remain sufficiently consistent to support comparative evaluation. Appendix D provides the full rubric and validation protocol, while Appendix F lists the judge prompts.

6 Conclusion

This work studies legal consultation at the stage where the factual basis for advice is being formed. We introduce FILC-Bench, a dynamic benchmark that evaluates consultation with incomplete client narratives through controlled fact disclosure. Extensive experiments reveal limitations of existing methods in this setting, as they often produce plausible legal analysis while still failing to handle unresolved facts. To address this problem, we propose SGCA, which maintains an evolving fact state, uses SGI for targeted clarification, and applies ACR before final advice. Experiments show that SGCA consistently improves consultation quality across all levels of narrative incompleteness, especially in issue handling and sufficiency management. Future work may further enhance LLM consultation capabilities and integrate them into broader legal agents equipped with more specialized task tools.

Limitations

This work studies legal consultation under incomplete client narratives, with the following limitations. First, FILC-Bench is derived from Chinese civil judgment documents, so it reflects Chinese civil consultation scenarios more directly than other jurisdictions and live intake conversations. Second, the Client Agent follows a controlled disclosure rule for omitted facts. This design supports consistent evaluation, but it simplifies the variability of real clients, such as vague expression, memory gaps and emotional narration. Third, SGCA focuses on fact sufficiency management under incomplete narratives without using retrieval, external legal tools, or broader procedural functions. Finally, the evaluation relies mainly on automatic scoring based on LLMs. Human validation supports comparative reliability, but model specific scoring preferences may still remain.

Ethical Considerations

FILC-Bench is constructed from publicly available civil judgment documents. Such documents may contain sensitive personal information. We remove or anonymize personally identifiable information during processing and release only the benchmark fields needed for research. The released fields support reproducible evaluation rather than reconstruction of original case records. Users of the benchmark should follow applicable data use requirements and avoid attempts to identify litigants or other individuals mentioned in the source cases.

This work studies legal consultation agents for research and evaluation. SGCA is designed to produce advice grounded in available facts and attentive to possible disputes, while use in real cases should remain under the supervision of licensed legal professionals. The system should not be treated as a substitute for professional legal judgment.

Human validation was conducted by graduate annotators with legal training for research evaluation. Annotators were informed of the evaluation purpose, followed a structured scoring protocol, and were compensated for their time at a rate consistent with local research assistant work. The evaluation used processed benchmark records and model outputs, and no additional personal data was collected or stored during the process.

ChatGPT was used solely to assist with language polishing. All claims, analyses, and conclusions remain the responsibility of the authors.

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *Preprint*, arXiv:2306.16092.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

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

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Jidong Ge, and Vincent Ng. 2025. InternLM-law: An open-sourced Chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392, Abu Dhabi, UAE. Association for Computational Linguistics.

GLM Team. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

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

Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yidong Liu, Zejun Li, Yun Song, and Zhongyu Wei. 2026. Ready jurist one: Benchmarking language agents for legal intelligence in dynamic environments. *Preprint*, arXiv:2507.04037.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025b. LegalAgentBench: Evaluating LLM agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria. Association for Computational Linguistics.

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025c. LexRAG: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3606–3615, New York, NY, USA. Association for Computing Machinery.

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

Hui Liao, Chuan Qin, Yongwen Ren, Hao Li, Zhenya Huang, Yanyong Zhang, and Chao Wang. 2026. VERDICT: Verifiable evolving reasoning with directive-informed collegial teams for legal judgment prediction. *Preprint*, arXiv:2603.19306.

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.

Mistral AI. 2024. Ministral 8b.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2025. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4975–5001, Albuquerque, New Mexico. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

OpenAI. 2026. Introducing gpt-5.4.

Henry Prakken. 2004. Analysing reasoning about evidence with formal models of argumentation. *Law, Probability and Risk*, 3(1):33–50.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Dilin Shu, Haoran Zhao, Haoyue Wang, Siyu Ma, Ge Qu, Wenguang Chen, and Xingze Wu. 2024. Lawllm: Law large language model for the chinese legal domain and a legal llm benchmark. *Preprint*, arXiv:2407.21065.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. Lawluo: A multi-agent collaborative framework for multi-round chinese legal consultation. *Preprint*, arXiv:2407.16252.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yiquan Wu, Kun Kuang, Weiming Lu, Beining Hu, and Fei Wu. 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 15464–15483, Bangkok, Thailand. Association for Computational Linguistics.

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

Xinyu Yang, Chenlong Deng, Tongyu Wen, Binyu Xie, and Zhicheng Dou. 2026. Lawthinker: A deep research legal agent in dynamic environments. *Preprint*, arXiv:2602.12056.

Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Jun Lin, Kun Kuang, Ji Zhang, and Xiaozhong Liu. 2025. Lecode: A benchmark dataset for interactive legal consultation dialogue evaluation. *Preprint*, arXiv:2505.19667.

Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuanjing Huang, and Zhongyu Wei. 2025a. Multi-agent simulator drives language models for legal intensive interaction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6552–6585, Albuquerque, New Mexico. Association for Computational Linguistics.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2025b. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25796–25804.

Kaiyuan Zhang, Jiaqi Li, Yueyue Wu, Haitao Li, Cheng Luo, Shaokun Zou, Yujia Zhou, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025. Chinese court simulation with llm-based agent system. *Preprint*, arXiv:2508.17322.

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

A Discussion

Incomplete client narratives are a common starting point of legal consultation. Clients often describe disputes from their own perspective and may leave out facts that are legally relevant but not salient to them. FILC-Bench makes this problem observable in a controlled evaluation setting. Civil judgment documents provide structured facts, claims, defenses, and outcomes, which allow us to construct consultation instances with controlled levels of omitted information. The benchmark therefore focuses on a narrower but central capability of legal consultation agents, namely whether they can operate when the facts needed for advice are incomplete and still developing. At the same time, effective consultation should avoid excessive questioning. A useful agent should obtain material missing facts while keeping the interaction concise and actionable.

SGCA should be understood as an upstream analysis component for legal consultation. Its role is to organize the facts of a dispute before downstream legal tasks are performed. In this sense, SGCA complements legal retrieval, statutory reasoning, case based reasoning, document drafting, litigation strategy generation, and other modules in legal agent systems. For example, before retrieving relevant cases or drafting a legal document, an agent may first use SGCA to identify what is known, what remains unresolved, and what has been ruled out. The agent can then decide whether the available facts are sufficient for the downstream task. This makes SGCA compatible with existing legal agents, where it can serve as a preliminary consultation layer that improves the factual grounding of later reasoning and generation steps.

The goal of SGCA is to improve the factual grounding and process quality of legal consultation within a bounded advisory setting. Its outputs should be treated as auxiliary legal information rather than a substitute for professional legal judgment. This boundary is especially important when the initial narrative is incomplete, because some facts may remain unknown, disputed, or difficult to verify even after clarification. For this reason, SGCA frames its final advice around the facts available during consultation and explicitly marks remaining uncertainty. Such advice helps clients understand possible legal issues and risks while avoiding conclusions that exceed the current factual record.

B Additional Details on FILC-Bench

This section supplements the benchmark description in the main paper. It explains how source judgments are converted into dynamic consultation instances, reports statistics by legal category, and clarifies how structured fields are used for dialogue construction and evaluation.

B.1 Data Construction Details

Each dynamic consultation instance is derived from a civil judgment through a structured conversion pipeline. We collected 397 raw civil judgments and manually screened 353 candidate cases for benchmark construction. Cases were retained only when the judgment contained separable plaintiff allegations, requested remedies, defendant responses, court reasoning, and a final holding. We excluded 57 cases because of missing party statements, inseparable factual descriptions, incomplete reasoning, duplicated disputes, or insufficient factual detail. The retained cases were then parsed with LLM assistance and manually checked for field boundaries, factual consistency, and separation among client available information, opposing party information, and court reference information.

Each source case is organized into three layers. The lay narrative layer contains facts and requests that a non expert client could plausibly provide during consultation. The dispute layer contains defendant responses, opposing arguments, and evidentiary challenges. The oracle layer contains court found facts, court reasoning, and the final holding, which are reserved for reference based evaluation. Based on this structure, dynamic instances are created by masking part of the lay narrative layer before the first client turn. Visible facts form the opening narrative, while masked client side facts become omitted facts that may later become available through the client simulator. The low, medium, and high levels are constructed from the same retained source case pool by masking different proportions of the lay narrative layer. The omitted fact ratio is below 30% for low, between 30% and 50% for medium, and greater than 50% for high. This process produces 486 dynamic consultation instances and 1,311 topic interactions.

To support controlled interaction, each omitted fact is assigned a material fact dimension, such as time, identity, agreement terms, performance, notice, injury, causation, medical records, payment records, or prior negotiation. During dialogue,

Level	Initially visible facts	Omitted client side facts
Low	PF1 ceremony/cohabitation; PF2 marriage registration; PF3 two children; PF4 frequent conflicts; PF5 alleged lack of childcare by the defendant.	PF6 complete breakdown of the marital relationship.
Medium	PF1 ceremony/cohabitation; PF2 marriage registration; PF3 two children.	PF4 frequent conflicts; PF5 alleged lack of childcare by the defendant; PF6 complete breakdown of the marital relationship.
High	PF1 ceremony/cohabitation; PF3 two children.	PF2 marriage registration; PF4 frequent conflicts; PF5 alleged lack of childcare by the defendant; PF6 complete breakdown of the marital relationship.

Table 6: Example of fact masking across incompleteness levels in a marriage and family case. Omitted client side facts become available only when the lawyer asks about the corresponding material fact dimension.

Level	Cases	Visible facts	Omitted facts
Low	162	555	237
Medium	162	398	394
High	162	254	538
Total	486	1,207	1,169

Table 7: Benchmark statistics by narrative incompleteness level. Visible and omitted facts are counted from structured plaintiff fact identifiers used by the client agent.

Category	Cause types	Cases	Topics	Omitted	Share
Tort	12	75	213	154	15.4
Marriage and Family	7	60	144	147	12.3
Real Estate	7	60	168	155	12.3
Traffic Accident	4	57	69	69	11.7
Contract	6	54	147	153	11.1
Corporate	7	54	135	167	11.1
Medical	3	45	93	151	9.3
Creditor and Debtor	3	42	87	82	8.6
Labor	2	39	255	91	8.0
Total	51	486	1,311	1,169	100.0

Table 8: Benchmark statistics by legal category. Cause types count distinct original `cause` values in each category. Shares are percentages of the 486 constructed consultation instances.

the client simulator reveals an omitted fact only when the lawyer asks about the corresponding dimension. Manual verification then checks whether the opening narrative contains only visible facts, whether omitted facts are client side and legally relevant, whether topic lists match client requests, and whether oracle information is absent from the dialogue input. Table 6 provides a concrete masking example, where PF denotes plaintiff fact identifiers.

B.2 Dataset Statistics

This section provides additional statistics for FILC-Bench. Table 7 reports the number of constructed consultation instances, visible facts, and omitted facts under each level of narrative incompleteness. Table 8 further reports statistics by legal category. The source cases are collected from nine categories that frequently appear in legal consultation, covering marriage and family, traffic accident, contract, corporate, creditor and debtor, labor, medical, real estate, and tort. The reported counts refer to constructed consultation instances rather than raw judgment documents.

B.3 Dataset Fields

Table 9 lists the structured field groups extracted from each civil judgment. The schema separates information used to construct consultation inputs from court information used for evaluation. This separation ensures that the consultation agent only receives client available information, while court found facts, reasoning, and judgment results are retained for scoring legal correctness and issue coverage.

The field groups are used for different stages of benchmark construction and evaluation. Facts and requests are used to construct the opening narrative, omitted facts, and consultation topics. Court found facts, court reasoning, and judgment results are retained as references for evaluating legal correctness and issue coverage. This design keeps the dialogue grounded in information that a client could plausibly provide, while preserving court materials for evaluation. They are not provided to the consultation agent during dialogue generation.

C Additional Details on Experiments

This section complements the main experimental results by unpacking four questions: which SGCA components drive the gains, whether the improvements are statistically reliable, how much interaction they require, and whether the gains hold across legal categories.

Field group	Description
Facts	Case events, relationships, losses, performance, notice, negotiation, evidence, and other factual information relevant to the consultation.
Requests	Consultation topics, requested remedies, compensation items, or other outcomes sought by the client.
Court found facts	Facts accepted by the court after reviewing the evidence and party statements.
Court reasoning	Legal reasoning, rule application, and treatment of disputed issues.
Judgment result	Final holding, remedy allocation, liability allocation, or dismissal of claims.

Table 9: Structured field groups extracted from civil judgments.

Audit dimension	Avg. score	Acceptance rate
Relevance	1.85	88.8%
Client naturalness	1.87	90.8%

Table 10: Human audit results for generated consultation topics. Scores are averaged over two annotators on a three point scale, where 0 denotes invalid, 1 denotes partially acceptable, and 2 denotes acceptable.

C.1 Client Agent Audit

We conduct a human audit to examine whether the consultation topics generated by the client simulator are reasonable. We sample 120 generated topics across legal categories and narrative incompleteness levels. Two graduate annotators with legal training independently score each topic for relevance and client naturalness on a three point scale. Relevance measures whether the topic is connected to the source case and the client request, while client naturalness measures whether it sounds like a plausible question from a lay client. For both dimensions, 0 denotes an invalid topic, 1 denotes a partially acceptable topic with clear issues, and 2 denotes an acceptable topic. Table 10 reports the averaged audit results.

The results indicate that the generated topics are generally relevant to the source cases and are usually expressed in a way that is plausible for a lay client. This supports the use of the client simulator for constructing the consultation flow in FILC-Bench.

C.2 Ablation Results

Table 11 decomposes SGCA into the sufficiency check, SGI, and ACR components and reports results for all five evaluation metrics across incompleteness levels. The + Suff. variant improves Sufficiency Management without user clarification, indicating that state based sufficiency checking already helps avoid premature advice. SGI adds targeted fact recovery, while ACR improves dispute related dimensions without asking clarification questions. SGCA combines both sources of improvement and remains the strongest variant across all incompleteness levels.

C.3 Robustness Analysis

Table 12 examines whether the gains in the ablation study are stable across individual cases after aggregating all levels of narrative incompleteness. For each comparison, Δ is computed from paired case score differences. Confidence intervals are estimated with paired bootstrap, and p values are obtained with paired sign permutation tests using 10,000 resamples. We report the dimensions most directly related to consultation quality, together with the overall score.

The results show that SGCA yields reliable gains over Direct Reasoning and SGI on Issue Handling, Sufficiency Management, Client Utility, and the overall score. SGI also outperforms Direct Reasoning, indicating that identifying missing facts and asking targeted clarification questions already improves consultation quality. The additional gains of SGCA over SGI show that adversarial review further strengthens issue handling and the usefulness of final advice.

C.4 Results by Legal Category

Table 13 reports results by legal category. It includes the average scores of Direct Reasoning, SGI, and SGCA, the improvement of SGCA over Direct Reasoning, and additional statistics on consultation topics and clarification behavior. SGCA improves over Direct Reasoning in every retained legal category, with an average gain of 2.02 points. This suggests that the improvement is not driven by a single case type, but appears across different legal consultation scenarios.

C.5 Documentation of LLMs

We use external model artifacts only for research evaluation. Proprietary models, including GPT 5.4 and DeepSeek V4, are accessed through their official APIs, and we do not redistribute their model weights or derived proprietary assets. For open-weight models, we follow the corresponding model licenses and usage terms. GLM-4 9B is used under its custom GLM-4 9B License, and Ministral 8B

Variant	Low						Medium						High					
	LS	IH	FG	SM	CU	Avg.	LS	IH	FG	SM	CU	Avg.	LS	IH	FG	SM	CU	Avg.
Direct Reasoning	3.21	0.54	3.50	1.26	2.64	2.23	3.19	0.51	3.12	1.28	2.44	2.11	3.19	0.40	2.74	1.50	2.33	2.03
+ Suff.	3.28	1.86	3.85	2.65	3.26	2.98	3.24	1.81	3.50	2.67	3.03	2.85	3.16	1.61	3.83	2.64	2.67	2.78
+ Suff. + SGI	3.31	1.28	4.38	3.43	3.25	3.13	3.20	1.48	4.31	3.71	3.21	3.18	3.24	1.37	4.51	3.90	3.17	3.24
+ Suff. + ACR	4.26	2.64	4.24	3.62	3.59	3.67	4.20	2.57	4.01	3.65	3.40	3.57	4.10	2.27	3.55	3.61	3.12	3.33
⊕ SGCA	**3.98**	**3.33**	**4.65**	**3.98**	**4.41**	**4.07**	**3.93**	**3.22**	**4.70**	**4.11**	**4.28**	**4.05**	**3.74**	**3.28**	**4.74**	**4.49**	**4.17**	**4.08**

Table 11: Full ablation results by narrative incompleteness level. Each level contains 162 evaluated case histories. LS, IH, FG, SM, and CU denote Legal Soundness, Issue Handling, Factual Grounding, Sufficiency Management, and Client Utility, respectively.

Metric	Comparison	Δ [95% CI]	p
IH	SGCA vs Direct	2.80 [2.68, 2.92]	$< .001$
IH	SGI vs Direct	0.90 [0.79, 1.00]	$< .001$
IH	SGCA vs SGI	1.90 [1.77, 2.04]	$< .001$
SM	SGCA vs Direct	2.84 [2.67, 3.01]	$< .001$
SM	SGI vs Direct	2.33 [2.16, 2.51]	$< .001$
SM	SGCA vs SGI	0.51 [0.38, 0.64]	$< .001$
CU	SGCA vs Direct	1.81 [1.70, 1.93]	$< .001$
CU	SGI vs Direct	0.74 [0.64, 0.84]	$< .001$
CU	SGCA vs SGI	1.08 [0.96, 1.19]	$< .001$
Avg.	SGCA vs Direct	1.94 [1.86, 2.01]	$< .001$
Avg.	SGI vs Direct	1.06 [0.98, 1.13]	$< .001$
Avg.	SGCA vs SGI	0.88 [0.81, 0.95]	$< .001$

Table 12: Paired tests for method gains. Confidence intervals are estimated by paired bootstrap over case score differences, and p values are computed by paired sign permutation tests. Positive Δ indicates that the first method scores higher.

is used under the Mistral Research License. Other open-weight legal and general models are used according to the licenses specified by their official model releases. All model outputs are used only for benchmark evaluation and analysis in this paper.

For inference, proprietary models are queried through official APIs, and open models are served with vLLM `0.18.1` on NVIDIA A100 GPUs. We set the temperature to 0.7 for all generation calls. The maximum output length is 2,048 tokens for client agent and lawyer agent calls, and 32,768 tokens for evaluation calls. All methods use the same case input, consultation topic, and prompt templates under each narrative incompleteness level. For Qwen3 based agent experiments, the same non thinking chat template is used across all agent methods to ensure the same backbone behavior.

D Evaluation Rubrics and Judge Validation Details

This section provides additional details on the evaluation protocol. It describes the five scoring dimensions, the score anchors used by the judge, and the human validation procedure behind the validation results reported in the main paper.

D.1 Metric Definitions

Table 14 summarizes the five evaluation dimensions and their score anchors. The rubric is designed to score the completed dialogue history and the final advice together, so it captures both legal reliability and how the agent handles incomplete facts during consultation. Each metric uses a 0–5 scale, where higher scores indicate stronger performance under the facts available in the dialogue. The low, middle, and high anchors in the table provide condensed guidance for assigning scores consistently across cases.

D.2 Judge Validation Protocol

The validation set contains 120 completed dialogue histories. We sample 40 histories from each narrative incompleteness level. Within each level, eight histories are sampled for each compared method, namely Direct Reasoning, Reflection, Plan-and-Solve, SGI, and SGCA. Legal domains are balanced within each level and method group when enough candidate histories are available. Table 15 summarizes the validation set and adjudication process.

The three annotators are students with legal training and law firm internship experience. Before formal annotation, all annotators receive unified training on the five dimensional rubric and complete a pilot calibration round. The calibration focuses on adjacent score boundaries and on distinctions that are easy to confuse, especially Factual Grounding and Sufficiency Management. Pilot examples are used only for training and are excluded from the validation set.

During formal annotation, annotators work independently. They receive anonymized sample identifiers, completed dialogue histories, and the scoring rubric. Method names, narrative incompleteness levels, LLM judge scores, system identities, and aggregate experimental results are hidden. This

Category	Direct	SGI	SGCA	Δ	Topics	Omit./T.	Clar.	Yield	N
Tort	2.06	3.40	4.35	2.29	81	1.01	2.68	0.26	84
Traffic Accident	2.14	3.24	4.28	2.14	23	1.05	1.72	0.39	36
Contract	2.14	3.40	4.22	2.08	49	1.06	2.44	0.24	54
Creditor and Debtor	2.33	3.23	4.34	2.02	29	1.03	1.43	0.43	42
Labor	2.01	2.81	3.99	1.98	85	1.18	2.38	0.22	36
Medical	2.27	3.43	4.23	1.96	31	2.03	2.60	0.29	45
Real Estate	2.24	3.27	4.18	1.94	61	1.16	2.30	0.25	69
Corporate	2.22	3.31	4.14	1.93	35	1.24	1.76	0.32	45
Marriage and Family	2.26	3.32	4.11	1.85	43	1.13	2.61	0.29	51
Mean	2.19	3.27	4.20	2.02	49	1.21	2.21	0.30	51

Table 13: Results by legal category. Categories are sorted by the improvement of SGCA over Direct Reasoning. Δ denotes the improvement of SGCA over Direct Reasoning. Topics denotes retained consultation topics, Omit./T. denotes omitted facts per topic, Clar. denotes clarification questions, Yield denotes omitted facts made available per clarification question, and N denotes evaluated case histories.

Metric	Scoring focus	Low anchors	Middle anchors	High anchors
Legal Soundness	Whether the final advice applies relevant legal rules correctly, reaches reliable conclusions, and avoids materially wrong legal guidance with reference to court judgment information.	0–1: materially wrong or mostly incorrect legal advice.	2–3: partially relevant or generally correct, but with important missing conditions.	4–5: correct legal advice with minor omissions or accurate application of the relevant standard.
Issue Handling	Whether the consultation addresses central disputes, possible defenses, evidentiary weaknesses, procedural risks, and litigation strategy.	0–1: no recognition or only superficial recognition of the legal issue.	2–3: partial coverage of disputes, defenses, evidence needs, or risks.	4–5: strong or comprehensive handling of claims, defenses, evidentiary burdens, and procedural risks.
Factual Grounding	Whether the advice is supported by facts available in the dialogue, avoids unsupported factual claims, and marks uncertainty when facts remain unresolved.	0–1: fabricated, unsupported, or weakly connected factual assertions.	2–3: partial or generally faithful use of dialogue facts, with some unsupported leaps.	4–5: strong or precise use of material facts, with unresolved uncertainty clearly marked.
Sufficiency Management	Whether the agent identifies missing material facts, asks necessary clarification questions, uses newly provided facts, and limits conclusions when the record remains incomplete.	0–1: no awareness of missing facts or only generic caveats.	2–3: some awareness of missing facts, with limited clarification or conditioning.	4–5: targeted diagnosis, clarification, use of new facts, and controlled conclusions under uncertainty.
Client Utility	Whether the advice is clear and useful for the client, including risk explanation, practical next steps, and guidance for further evidence or action.	0–1: unusable response or little practical help.	2–3: limited general guidance or useful but incomplete advice.	4–5: actionable or highly useful consultation support tailored to the client's facts and likely choices.

Table 14: Evaluation dimensions and score anchors. Each metric is scored on a 0–5 scale. Low, middle, and high anchors summarize how scores are assigned for each dimension.

Item	Value
Dialogue histories	120
Incompleteness levels	3
Histories per level	40
Compared methods	5
Histories per method per level	8
Evaluation dimensions	5
Sample metric pairs	600
Adjudicated pairs	53
Adjudication rate	8.8%
Annotators	3

Table 15: Summary of the human validation set and adjudication process.

blind setting reduces the risk that annotators favor a particular method or adjust their scores toward the LLM judge.

Human reference scores are computed from the three annotator scores for each metric. Before adjudication, inter annotator agreement is computed using Krippendorff's alpha. If the maximum and minimum scores among the three annotators differ by more than two points for a sample and metric pair, that pair is flagged for adjudication. The annotators discuss each flagged pair according to the rubric and assign a consensus score. This consensus score is used as the human reference for flagged pairs, while non flagged pairs use the mean of the three independent scores.

Agreement between the LLM judge and the human reference scores is measured with rank correlation and mean absolute error. Spearman's ρ and Kendall's τ evaluate whether the LLM judge preserves the relative ordering of outputs, while mean absolute error measures the absolute gap between judge scores and human reference scores. The compact validation results are reported in the main paper.

E Case Study

The case study provides a qualitative view of the mechanism behind the quantitative gains. Figures 6 and 7 compare two consultation trajectories for the same high incompleteness tort dispute instance. The baseline answers from the sparse initial narrative, whereas SGCA first asks about missing material facts and then produces a bounded answer that addresses disputes, evidentiary risks, and remaining uncertainty. The figures highlight key facts and legal judgments in the dialogue so that the differences in fact use and advice scope can be inspected directly. They also provide scoring rationales for all five evaluation dimensions, showing why the two trajectories receive different scores on legal soundness, issue handling, factual grounding, sufficiency management, and client utility.

F Prompts

This section documents the prompt families used for the client agent, SGCA and LLM judge scoring. The original experiments were conducted with Chinese prompts. For readability, Figures 8–19 present English versions that preserve the structure and operative instructions of the experimental prompts.

The prompt set covers the main stages of the pipeline. SGCA prompts perform intake, extract initial facts, diagnose whether the current record is sufficient, update the fact state, and generate the final answer. Client agent prompts generate user responses while exposing omitted facts only when the lawyer asks about a matching material fact dimension. Adversarial review prompts examine draft advice from an opposing side and identify unsupported assumptions, overlooked defenses, and litigation risks before final revision. Judge prompts score completed dialogue histories under Legal Soundness, Issue Handling, Factual Grounding, Sufficiency Management, and Client Utility.

Figures 8–19 provide the English prompt cards in pipeline order. They first show state construction and sufficiency control, then client simulation and adversarial review, followed by the judging prompts for each evaluation dimension. This presentation makes the prompt components explicit without relying on hidden prompt logic.

Direct Reasoning (Qwen3-32B)	
Client Topic	Compensation claim arising from injury and medical and transportation expenses after a neighbor assault
Full Dialogue	**Turn 0 | General Public** Hello, I would like to consult about a dispute involving the right to life, bodily rights, and health rights. The general situation is that the other party and I are neighbors. I would like to ask a few specific questions below. **Turn 1 | Legal Trainee** Hello, thank you for your consultation. Based on what you described, you and the other party are neighbors, and the matter involves a dispute over the right to life, bodily rights, and health rights. I have understood this background information. I will answer your specific questions one by one based on the relevant legal provisions. Please continue with your question. **Turn 2 | General Public** I was **injured** after my **neighbor hit me**, and I incurred **medical and transportation expenses** Can I ask him to compensate me? **Turn 3 | Legal Trainee** Yes, you may **request compensation**. Under the relevant provisions of the Civil Code, a person who infringes upon another person's civil rights and interests through **fault** and causes damage shall bear tort liability. If you were injured because of your neighbor's infringing act and therefore incurred medical and transportation expenses, these are direct losses, and you have the right to ask the other party for compensation. You should keep the relevant **medical invoices**, **transportation receipts**, and other evidence so that they can support your compensation claim. If negotiation fails, you may protect your lawful rights and interests through reporting to the police, mediation, or litigation. **Turn 4 | General Public** Thank you for your answer. These are all the questions I wanted to ask for now.
Judge Reasons	• **LS:** It gives a definite conclusion without incorporating **fault allocation**, omits **responsibility apportionment**, and does not state the **evidentiary conditions** needed for the claim. • **IH:** The answer merely restates the right to compensation and fails to identify core disputes such as **tort elements** and **causation**. • **FG:** It does not fabricate facts, but it does not condition statements such as negotiation failure on disclosed facts. • **SM:** It does not identify any **missing factual gaps**, gives a conclusion directly, and only gives generic advice to preserve receipts. • **CU:** The answer provides only **general legal knowledge**, does not use **case-specific facts**, and gives broad and non-specific action and evidence advice.

Figure 6: Case study for Direct Reasoning (Qwen3-32B). The model gives answer from the sparse initial narrative and receives weaker judge comments.

Sufficiency-Guided Consultation Agent (SGCA)

Full Dialogue

Turn 0 | Client

Hello, I would like to consult about a dispute involving the right to life, bodily rights, and health rights. **The general situation** is that **the other party and I are neighbors**. I would like to ask a few specific **questions**

Turn 1 | Lawyer

Hello, thank you for your consultation. Based on what you described, you and the other party are **neighbors**, and the matter involves a dispute over the **right to life, bodily rights, and health rights**. To provide an **accurate legal analysis**, I will ask necessary **clarification questions** about the **incident process**, **damages**, and **responsibility** before answering each question. Please continue.

Turn 2 | Client

I was **injured** after my **neighbor hit me**, and I incurred **medical and transportation expenses**. Can I ask him to compensate me?

Turn 3 | Lawyer

Could you describe the specific **process** of how you were injured? Did the other party commit any clearly **wrongful act**?

Turn 4 | Client

That day, we had a dispute over some everyday trivial matter. Later, he hit me. My **eyes, nose, feet**, and several other body parts were injured.

Turn 5 | Lawyer

When exactly did this happen? Did anyone else witness it? Did you call **the police** or keep any **related evidence**? Also, how much did you spend on **medical and transportation expenses**, and do you have **receipts**?

Turn 6 | Client

The incident happened on **January 17, 2015**. When we were arguing, no one nearby saw it. After I was hit, I called **the police**, and the police came. As for the **medical and transportation expenses**, I remember that the medical expenses were more than **RMB 9,491**, and the transportation expenses were about **RMB 500**. I have

Turn 7 | Lawyer

Hello, thank you for the additional information. Based on the facts currently **confirmed** in your statement, you have the **right to claim compensation** for the **medical and transportation expenses** caused by your **neighbor's assault**. However, because the case may involve **disputed issues** and **evidentiary risks**, the following is a comprehensive legal analysis and recommendation.
According to your statement, on **January 17, 2015**, you were assaulted after a dispute with your neighbor, suffered injuries to multiple body parts, incurred **RMB 9,491** in medical expenses and **RMB 500** in transportation expenses, reported the incident to the police, and retained the relevant **receipts**. Under **Article 1165 of the Civil Code**, which regulates liability for infringing upon another person's civil rights and interests through **fault** and causes damage shall bear tort liability. Therefore, you have the **right to claim compensation** for medical and transportation expenses from the other party
Evidence and response strategy. You should prepare **police records**, police notes, and **dispatch records** to prove the occurrence of the incident and the other party's conduct...You should also preserve **medical records**, **diagnostic certificates**, **transportation receipts** …
The exact **incident process** and **responsibility allocation** remain uncertain; the **causal relationship** between the injuries and the medical expenses needs support from medical records or injury appraisal; the **statute-of-limitations issue** requires confirmation; and the **completeness and probative value** of the evidence still need to be evaluated in litigation.

Turn 8 | Client

Thank you for your answer. These are all the questions I wanted to ask for now. I will end the consultation here.

Judge Reasons

- **LS:** The conclusion is correct, absorbs key disclosed facts, states the required premises, defenses, and risks, and covers **medical expenses**, transportation expenses, and **limitation-period issues**.
- **IH:** It fully identifies **disputes**, defenses, and **evidentiary risks**, gives **concrete strategies**, and covers the client's **request**.
- **FG:** The lawyer strictly relies on dialogue facts, does not fabricate facts or turn uncertain facts into established ones, and uses **conditional language** without overclaiming.
- **SM:** It accurately identifies **gaps** such as **limitation period** and **fault allocation**, asks follow-up questions, absorbs the disclosed facts, and gives a **closed-loop analysis**.
- **CU:** The answer is clear, **fact-specific**, and actionable, with concrete steps and an **evidence checklist**.

Figure 7: Case study for SGCA on the same consultation instance. SGCA elicits additional facts before answering and produces fact-grounded advice with stronger judge comments across all five evaluation dimensions.

Intake and Initial Fact Extraction Prompt

You are acting as a fact-driven legal consultation lawyer.

Current stage: factual intake after the consultation begins:
1. Do not answer specific legal questions.
2. Use professional language appropriate for legal consultation to briefly restate the factual background you have understood.
3. Clearly state that subsequent responses will first conduct necessary clarification around key facts, and then answer the client's questions one by one.
4. Do not ask fact follow-up questions in this stage.
5. Do not fabricate information that the client has not stated.
6. Respond naturally, clearly, and concisely.

Please extract the ordinary client's first-turn statement into professional legal facts.

Client's first-turn statement: **{first_message}**

Output JSON format:
{
  "confirmed_facts": ["concise, professional legal facts"]
}

Figure 8: Prompt for intake and initial fact extraction, used to extract the initial facts and requests from the client narrative.

Sufficiency Diagnosis Prompt

You are a fact-driven legal consultation lawyer.

Your task is to decide, based on the current topic and fact state, whether the lawyer should ask for facts in this round or directly answer.

Rules:
1. Judge factual sufficiency only around the current topic.
2. Follow-up questions must serve legal judgment, not idle curiosity.
3. Do not repeat questions about facts that have already been confirmed, denied, or clarified.
4. If the known facts are sufficient to give a bounded legal opinion, stop asking and answer directly.
5. If key facts are still missing and may affect liability, remedy, evidence, limitation period, amount, or litigation risk, ask one concise follow-up question.
6. Each follow-up should preferably cover 1-3 key factual points, not a long questionnaire.
7. Do not reveal or assume latent facts.
8. Output must be JSON only.

Please decide whether, under the current topic, the lawyer should ask for key facts in this round or proceed to answer.

Current fact state. You must first read the confirmed, denied, and uncertain information to avoid repeated follow-up questions: **{fact_state}**

Client's current question: **{user_message}**

Output JSON format:
{
  "need_followup": true/false,
  "reason": "why a follow-up is or is not needed",
  "missing_facts": ["fact gaps that affect the current judgment but have not yet been obtained; write only fact gaps, not information that has already been confirmed or denied"],
  "followup_question": "if need_followup=true, write the lawyer's natural follow-up question to the client here; if need_followup=false, write an empty string"
}

Figure 9: Prompt for sufficiency diagnosis, used to determine whether the current facts are sufficient for the consultation topic.

Fact-State Update Prompt

You are a fact-state maintainer.

You need to update the current legal consultation fact state according to the lawyer's follow-up question and the client's answer.

Rules:
1. Record only facts explicitly expressed by the client or reasonable restatements of them.
2. Distinguish confirmed_facts, denied_facts, uncertain_facts, and missing_facts.
3. Do not write the lawyer's inference as a confirmed fact.
4. Do not fabricate facts.
5. Output must be JSON only.

Please update the fact state according to the follow-up question and answer.

Fact state before update: **{fact_state}**

Lawyer's follow-up question: **{lawyer_question}**

Client's answer: **{general_public_answer}**

Output JSON format:
```
{
  "confirmed_facts": ["facts that have been confirmed to exist"],
  "denied_facts": ["facts that have been confirmed not to exist"],
   "missing_facts": ["facts that are still missing and may affect the judgment"]
}
```

Figure 10: Prompt for fact state update, used to update known facts, unresolved facts, and confirmed absences after each client response.

User-Facing Final Answer Prompt

You are acting as a fact-driven legal consultation lawyer.

The current task is to answer the client's current topic based on the facts that have been confirmed and clarified.

Rules:
1. Directly answer the current question.
2. Clearly distinguish judgments supported by confirmed facts from uncertainties caused by insufficient facts.
3. If facts are still insufficient, state the missing key facts and the impact of those missing facts.
4. Do not continue to ask follow-up questions in the final answer.
5. Do not fabricate statutes, cases, facts, amounts, or procedural status.
6. Respond naturally, clearly, and concisely.

Please directly answer the client's current topic based on the current fact state.

Current fact state: **{fact_state}**

Client's current question: **{user_message}**

Recent dialogue: **{dialog_history}**

Please output the final answer.

Figure 11: Prompt for final answer generation, used to produce legal advice grounded in the current consultation record.

General-Public Follow-up Response Prompt

The lawyer is asking follow-up factual questions around the current topic. Please answer as an ordinary client.

visible_facts (facts that can already be expressed): **{visible_facts}**

latent_facts (can be supplemented only when the lawyer explicitly asks): **{latent_facts}**

Lawyer's follow-up question: **{lawyer_question}**

Requirements:
1. Only answer the facts asked by the lawyer.
2. Do not proactively expand on unrelated latent facts.
3. Do not fabricate new facts.
4. Do not cite court-found facts, court reasoning, or the holding.

Output only a JSON object:
{
  "content": "the client's natural, colloquial answer",
}

Figure 12: Prompt for client response generation, used to simulate client replies and provide omitted facts when the lawyer asks about a matching fact dimension.

Opposing-Counsel Review Prompt

You are an opposing counsel, representing the interest position of the party opposite to the client.
Your task is to attack our lawyer's preliminary opinion and help identify vulnerabilities in adversarial litigation.

Rules:
1. From the perspective of the opposing party or defense lawyer, raise defenses, factual challenges, and evidence attacks.
2. Do not fabricate facts that are not present in the case materials.
3. You may rely on the opposing-side factual claims and defenses provided in the case record.
4. Focus on risks in the preliminary answer, unsupported assumptions, missing evidence, causation, liability, remedies, procedural issues, and limitation-period issues.
5. Output must be JSON only.

Please attack our lawyer's preliminary consultation opinion on the current topic from the opposing lawyer's perspective.

Client-side requests: **{plaintiff_requests}**

Current known fact state: **{fact_state}**

Our lawyer's preliminary answer: **{initial_lawyer_answer}**

Please output JSON:
{
  "opposing_arguments": ["substantive or procedural defenses that the opposing side may raise"],
  "fact_challenges": ["which facts the opposing side may challenge as unproven or insufficiently evidenced"],
  "legal_risks": ["legal risks or conclusion boundaries in our preliminary answer"],
  "suggested_cross_questions": ["issues our side should further confirm or prepare for in response to opposing defenses"],
  "attack_summary": "one-sentence summary of the opposing lawyer's strongest attack"
}

Figure 13: Prompt for opposing counsel review, used to examine draft advice from the opposing side and identify possible defenses, unsupported assumptions, and litigation risks.

Final Revision Prompt

You are the legal consultation lawyer on the client's side.
You have already generated a preliminary answer and received attack comments from a hypothetical opposing counsel.
Now you need to revise the final external answer.

Rules:
1. The final answer is still addressed to the ordinary client and must not expose the internal multi-agent process.
2. Revise the answer based on the opposing counsel's attack.
3. Clearly mark conclusion boundaries, evidentiary risks, possible defenses, and remaining uncertainties.
4. Do not fabricate new facts.
5. The answer should remain practically useful.

Please revise the final legal consultation answer for the current topic according to the attack comments from the hypothetical opposing counsel.

Client's current question: **{user_message}**

Current fact state: **{fact_state}**

Our lawyer's preliminary answer: **{initial_lawyer_answer}**

Hypothetical opposing counsel's attack: **{opposing_attack}**

Please directly output the final answer shown to the client. The answer should include:
1. Current conclusion;
2. Confirmed facts on which the conclusion depends;
3. Possible opposing defenses or disputed issues;
4. Evidence and response suggestions;
5. Uncertainties that still need to be preserved.

Figure 14: Prompt for final revision, used to revise draft advice after opposing counsel review.

Legal Soundness

Please evaluate the legal soundness of the lawyer's final answer to the current topic.

Evaluation object: only the lawyer's final answer for the current topic.

Score levels:
0: Does not answer the current topic, or is completely off-topic.
1: Answers the topic, but the conclusion direction is clearly wrong, or is opposite to the core judgment in court_reasoning/holding.
2: Contains partially correct legal knowledge, but does not combine the key facts of this case; or gives a definite conclusion when facts are insufficient.
3: The conclusion direction is basically reasonable and uses conditional expression; it can state what key facts or evidence are needed, but does not absorb newly added facts to form a more specific judgment.
4: The conclusion direction is correct, and the answer absorbs key facts later disclosed in the dialogue, making the judgment more specific and more case-fitted than the initial conditional answer.
5: The conclusion is correct, specific, and complete; it fully absorbs key newly added facts and accurately states legal factual premises, limiting conditions, possible exceptions, and final risks.

Case reference: **{reference}**

Current topic: **{topic}**

Lawyer's final answer: **{lawyer_answer}**

Output only a JSON object in the following format:
```
{
  "score": 0,
  "reason": "brief scoring reason",
}
```

Figure 15: Prompt for scoring Legal Soundness, used to evaluate legal rule application and conclusion reliability.

Issue Handling

Please evaluate how effectively the lawyer identifies and handles the core legal issues and disputes in the whole consultation.

Evaluation object: whether the lawyer identifies and handles the core disputes in this case, rather than only giving general legal explanations.

Score levels:
0: Does not identify any case dispute.
1: Only vaguely mentions that there may be disputes, but does not explain specific disputes.
2: Identifies some disputes, but mainly remains at templated analysis and does not combine case facts or opposing defenses.
3: Can point out major disputes, evidence risks, or possible opposing defenses based on existing facts, but the handling is conditional or general and does not use newly added facts to deepen the analysis.
4: Absorbs key facts later disclosed in the dialogue and uses them to concretize disputes, opposing defenses, evidence risks, or response strategies.
5: Completely handles core disputes; can accurately distinguish strengths/weaknesses, opposing defenses, burden of proof, evidence gaps, and litigation/negotiation strategies around newly added facts.

Case reference: **{reference}**

Relevant dialogue fragments: **{topic_interactions}**

Output only a JSON object in the following format:
```
{
  "score": 0,
  "reason": "brief scoring reason",
}
```

Figure 16: Prompt for scoring Issue Handling, used to evaluate whether disputes, defenses, evidentiary weaknesses, and risks are addressed.

Factual Grounding

Please evaluate the factual grounding of the lawyer across the whole legal consultation.

Evaluation object:
Assess whether the lawyer uses the facts disclosed in the dialogue to support legal conclusions, risk analysis, and practical advice, while avoiding claims that go beyond the available factual basis.

Evaluation principles:
- Reward answers that explicitly use material facts from the dialogue and connect them to legal judgments, risks, disputes, or action suggestions.
- Do not reward answers that merely provide generic legal knowledge, templated cautions, or vague statements such as "it depends on the evidence."
- Penalize unsupported factual additions, overconfident conclusions, and failure to distinguish confirmed facts from unresolved facts.

Scoring sub-items
1. The lawyer explicitly uses specific and material facts disclosed in the dialogue, rather than only giving general legal explanations.
2. The lawyer establishes a clear connection between disclosed facts and legal conclusions, risks, disputes, or action suggestions.
3. The lawyer clearly marks facts that remain missing, uncertain, or unverified, and does not treat them as established facts.
4. The lawyer does not introduce new facts, amounts, responsibility ratios, procedural status, or evidence that are absent from the dialogue.
5. The lawyer avoids conclusions that exceed the factual basis, such as guaranteed outcomes, absolute liability statements, or unwarranted certainty.

Case and dialogue:
{
  "task_context": **{task_context},**
  "dialog_interactions": **{dialog_interactions}**
}

Output only a JSON object in the following format:
{
  "score": 0,
  "reason": "brief scoring reason"
}

Figure 17: Prompt for scoring Factual Grounding, used to evaluate whether advice is supported by facts available in the dialogue.

Sufficiency Management

Please evaluate how effectively the lawyer recognizes and manages insufficient factual information throughout the consultation.

Evaluation principles:
- Information sufficiency may be handled through follow-up questions, explicit identification of fact gaps, conditional reasoning, evidence guidance, or verification suggestions.
- Purely conditional handling can demonstrate awareness of missing information, but higher scores require effective use of clarification or newly disclosed facts in the final answer.
- When new facts are obtained during the dialogue, the final answer should incorporate them into the legal judgment, risk analysis, or practical advice.

Score levels:
0: Completely unaware of information insufficiency and directly gives a conclusion.
1: Vaguely says "information is insufficient" or "more evidence is needed," but does not specify what is missing.
2: Points out some fact gaps, but they are not specific, or does not explain how these gaps affect the judgment.
3: Can clearly point out key fact gaps and handle information insufficiency through conditional analysis, risk warnings, evidence suggestions, or verification suggestions, but does not obtain or absorb newly added facts.
4: Identifies key gaps and, through follow-up questions or absorption of later disclosed facts, uses newly added information in the final judgment, risk warning, or action advice.
5: Forms a complete information loop: precise gap identification, effective follow-up or fact absorption, and a final answer that clearly updates the conclusion, disputes, evidence suggestions, and risk boundaries.

Task context: **{task_context}**

Relevant dialogue fragments: **{topic_interactions}**

Output only a JSON object in the following format:
```
{
  "score": 0,
  "reason": "brief scoring reason",
}
```

Figure 18: Prompt for scoring Sufficiency Management, used to evaluate whether missing facts and factual limits are handled properly.

Client Utility

Please evaluate the consultation helpfulness of the entire legal consultation from the perspective of an ordinary client. Focus on whether the lawyer's final response is genuinely useful for understanding the issue and deciding what to do next. Do not assign a high score merely because the answer is lengthy, well-formatted, or professional in tone.

Score levels:
0: The user cannot act on the answer, or the answer is basically off-topic.
1: Contains some legal common knowledge, but is of little help for the user's current question.
2: The answer is partly useful, but mainly gives templated advice and lacks case facts and concrete action paths.
3: The answer is clear and cautious, with conditional advice and general evidence/action tips; the user can obtain basic help, but the advice remains somewhat generic.
4: Absorbs newly added facts and gives more specific action suggestions, evidence preparation, risk judgment, or communication/litigation strategy.
5: Highly useful and executable; based on a complete fact loop, gives clear conclusions, concrete steps, evidence lists, risk priorities, and practical options.

Task context: **{task_context}**

Summary of lawyer final answers: **{final_lawyer_answers}**

Output only a JSON object in the following format:
```
{
  "score": 0,
  "reason": "brief scoring reason",
}
```

Figure 19: Prompt for scoring Client Utility, used to evaluate whether advice is clear, useful, and actionable for the client.