"I Don't Know What to Say": A Fact-Filling Questionnaire Method to Help Non-Experts Talk to LegalAI Assistant

Yuting Huang[1,2], **Yiquan Wu**[1*], **Meiotng Guo**[1], **Ang Li**[1], **Xiaozhong Liu**[3]
Keting Yin[1], **Fei Wu**[1], **Kun Kuang**[1*]

[1]Zhejiang University, Hangzhou, China
[2]Fashton Technology, Hangzhou, China
[3]Worcester Polytechnic Institute, Worcester, USA
{yutinghuang, wuyiquan, guomeitong, leeyon, yinkt, wufei, kunkuang}@zju.edu.cn
xliu14@wpi.edu

Abstract

Artificial intelligence has become increasingly prevalent in the legal domain. However, LegalAI systems often struggle with vague user queries that lack essential legal details, leading to suboptimal performance in practical applications. To address this challenge, we propose FactFiller, a novel approach that dynamically generates questionnaires to help users refine their input queries. Our method leverages an iterative training process that collects valuable questionnaires, eliminating the need for human annotation. Additionally, we introduce a "case-law-quiz" cascading retrieval process, ensuring that the generated questions and answer options are directly linked to specific legal provisions. Through the user study and the downstream task experiments, we demonstrate that FactFiller, while remaining easy for non-experts to understand, not only improves the completeness of queries but also ensures the performance of various domain-specific models in downstream legal tasks.

1 Introduction

Recent advancements in large language models (LLMs), have greatly expanded the scope of legal artificial intelligence (LegalAI), enabling applications such as legal consultation Q&A (Shi et al., 2024; Büttner and Habernal, 2024), legal judgment prediction (LJP) (Wu et al., 2023; Tong et al., 2024), and legal case retrieval (LCR) (Li et al., 2023b; Wiratunga et al., 2024). Despite this progress, LegalAI systems often fail to provide reliable assistance in real-world settings.

The root cause lies in a fundamental issue: **non-expert users frequently formulate vague or incomplete queries that lack essential legal-specific details**. Especially in the context of legal consultation, as shown in Figure 1, a workplace injury consultation might omit critical information,

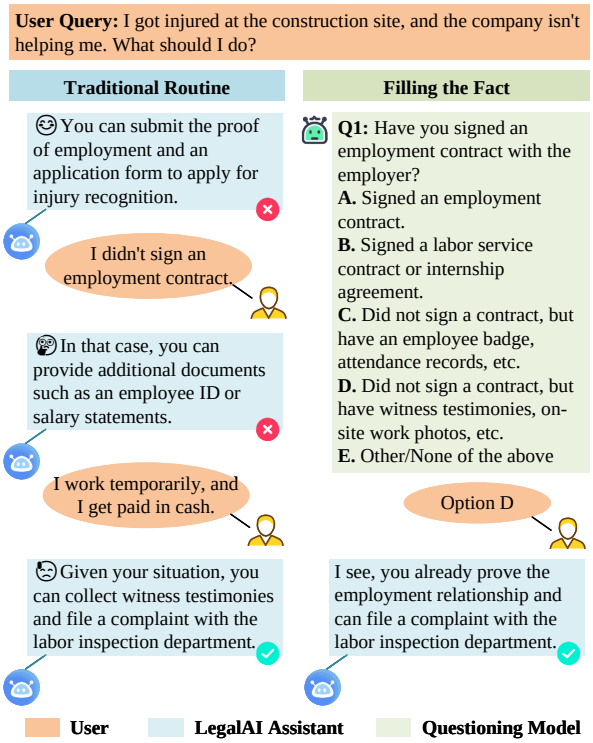

Figure 1: In traditional workflows, vague or incomplete user queries often lead LegalAI assistant to produce unhelpful responses. To address this, we use questionnaires generated by FactFiller to gather case-related facts, enabling LegalAI assistant to make more accurate judgments.

such as unsigned labor contracts or missing payroll records. When users repeatedly receive unhelpful or incorrect responses from LegalAI, their trust in the system is undermined.

Recent studies suggest that LLMs can actively ask clarifying questions to elicit missing information (Li et al., 2023a; Andukuri et al., 2024). Yet in legal settings, open-ended clarification is often ineffective: non-expert users may struggle to interpret such questions, feel overwhelmed, or still fail to provide precise legal details. Human–computer interaction research suggests that reducing cognitive load improves both trust and

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 12193–12210

task performance (Huang, 2018). Compared to open-ended questions, structured multiple-choice questionnaires provide explicit and constrained options, simplifying decision-making and easing information elicitation (Cau and Spano, 2025). This motivates the use of questionnaire-style counter-questions for legal fact completion.

The implementation faces two major technical challenges: **(1) How to generate questions and options with substantive legal significance that directly influence subsequent reasoning?** The questions must not only be legally meaningful, so that different choices correspond to distinct legal consequences, but also remain understandable to users. Superficial yes/no questionnaires contribute little to downstream tasks. **(2) How to train the questioning model in the absence of labeled "query–quiz" datasets and the prohibitive cost of manual annotation?** Such datasets are virtually nonexistent, and constructing them would require extensive annotation by domain experts with deep legal knowledge. Therefore, it is crucial to design an unlabeled training paradigm that can leverage raw legal documents.

To tackle these challenges, we present **FactFiller**, a fact-filling questionnaire method that bridges the gap between non-expert queries and expert LLM models. FactFiller combines two key components. First, we propose an unlabeled iterative training paradigm that progressively improves the questioning model by selecting pedagogically valuable training instances. Second, we introduce a "case–law–quiz" cascading retrieval pipeline: relevant cases are retrieved to identify frequently cited legal provisions, and these provisions are then transformed into quizzes, ensuring that generated questions and options are grounded in authoritative law.

We evaluate FactFiller through complementary human evaluations and downstream task experiments. Experts assess questionnaire quality in terms of relevance, comprehensiveness, and helpfulness, while non-experts evaluate ease of understanding and expressive support. We further demonstrate that queries completed with FactFiller consistently improve LegalAI performance on court view generation, legal case retrieval, and confusing charge prediction, providing evidence of both human usability and downstream effectiveness.

In summary, our main contributions are:

1. We identify vague queries as a key problem in LegalAI and propose questionnaire-based fact-filling as a design solution.
2. We propose FactFiller, a fact-filling questionnaire method that introduces an unlabeled iterative training paradigm and a cascading retrieval pipeline to generate legally grounded, user-friendly questions.
3. We conduct two experiments: a Human evaluation (with both experts and non-experts) and a downstream task evaluation, demonstrating that FactFiller is easy to use while enhancing factual completeness and guaranteeing the performance of LegalAI models.

2 Related Work

2.1 Legal Artificial Intelligence

In recent years, advances in artificial intelligence, particularly deep learning and large language models, have enabled a growing number of intelligent applications in the legal domain. Representative tasks include court view generation (Li et al., 2024b; Liu et al., 2024), which automates the drafting of judicial opinions from case facts and claims; legal case retrieval (Li et al., 2023b; Wiratunga et al., 2024), which identifies relevant precedents through case or query matching; legal judgment prediction (Wu et al., 2023; Tong et al., 2024), which forecasts trial outcomes such as charges and penalties; and legal consultation chatbots (Shi et al., 2024; Büttner and Habernal, 2024), which provide accessible legal assistance by processing natural-language user queries.

Although these methods achieve strong performance on their respective benchmarks, their effectiveness is often limited in real-world settings, where user queries are frequently ambiguous or incomplete, preventing LegalAI systems from fully realizing their practical potential.

2.2 Proactive Questioning

Proactive questioning with LLMs has been explored as a way to elicit missing information from users. GATE (Li et al., 2023a) employs open-ended clarification questions to better capture human preferences, while STaR-GATE (Andukuri et al., 2024) extends this idea through large-scale role-playing data. Ask-before-Plan (Zhang et al., 2024) focuses on predicting clarification needs during task planning and invoking external tools accordingly.

In the legal domain, D3LM (Wu et al., 2024)

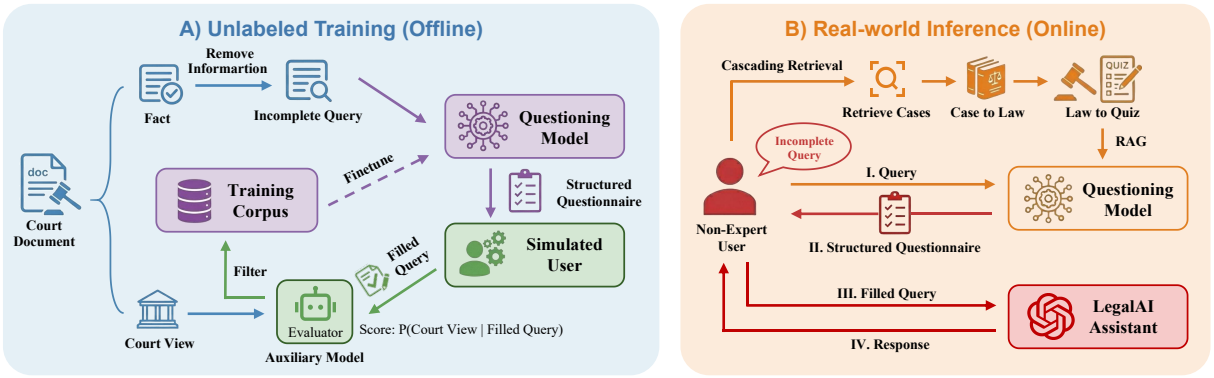

Figure 2: The proposed FactFiller framework. (a) An unlabeled iterative training paradigm simulates missing factinformation, generates questionnaires, and uses an auxiliary model to evaluate fact-completion ability for building the training corpus. (b) At inference time, cascading retrieval grounds questionnaires in law, and completed graduates improve downstream LegalAI responses.

simulates lawyers by generating open-ended professional questions to improve legal consultation. However, these approaches primarily rely on open-ended interaction, which can be difficult for non-expert users to respond to in a legally precise manner. In contrast, our work focuses on structured, questionnaire-based questioning to guide non-experts in providing legally meaningful facts.

3 The Proposed Fact-Filling Method: FactFiller

In this section, we introduce the FactFiller Framework as shown in Figure 2, which consists of two main components: (1) a training paradigm without labeled data, (2) a novel "case–law–quiz" cascading retrieval architecture.

3.1 Unlabeled Iterative Training Paradigm

Labeled "query–quiz" corpora do not exist, making it necessary to design an unlabeled training paradigm. As shown in Figure 2a, we start from court documents containing both factual descriptions and court views. Since the court view closely depends on complete factual information, we introduce an auxiliary legal expert model f_{AM} to define an indirect evaluation metric: the probability of generating the court view from a given query. A higher probability indicates that the query captures more complete facts. To construct such queries, we first perform a preprocessing step, where LLMs discard part of the legal facts and rephrase the factual paragraphs into incomplete user queries.

Formally, we aim to learn a questioning model f_{QM} by maximizing the following objective:

$$\max_{f_{\text{QM}}} \log P\left(C \mid Q + f_{\text{QM}}(Q) + A; f_{\text{AM}}\right), \quad (1)$$

where C denotes the court view, Q is an incomplete user query, $f_{\text{QM}}(Q)$ represents the generated questionnaire, and A denotes the simulated user answers. For an autoregressive auxiliary model f_{AM}, the conditional log-probability is computed as:

$$\log P(C \mid x; f_{\text{AM}}) = \frac{1}{T} \sum_{t=1}^{T} \log P(c_t \mid c_{<t}, x; f_{\text{AM}}). \quad (2)$$

where $C = (c_1, c_2, \ldots, c_T)$ is the token sequence of the court view.

The iterative procedure is then formalized in Algorithm 1. In each round, the system executes five steps: **(1)** the questioning model f_{QM} generates multiple candidate questionnaires for each query; **(2)** user selections are simulated based on the complete facts; **(3)** the simulated answers are concatenated with the incomplete query, and the auxiliary model f_{AM} estimates the probability of generating the correct court view; **(4)** the best-performing questionnaire for each query is retained in the supervised fine-tuning (SFT) corpus, replacing lower-quality ones; and **(5)** the questioning model is fine-tuned on this evolving corpus, enabling it to generate increasingly effective questionnaires in subsequent rounds without requiring manual annotation.

Since we adopt an iterative process to search for and identify effective questionnaires, the diversity of generated questionnaires is crucial for the training efficiency and performance of the model. To ensure that the questioning model can generate diverse questionnaires for the same query Q during the iterative process, we adopted two generation strategies shown in Figure 3.

First, at the between-cases level, we train the

Algorithm 1: Training the questioning model

Input : Number of iterations n; Number of questions in the questionnaires m; Number of generated questions p; Dataset D; Base model $f_{\text{QM}}^{(0)}$

Output : Questioning model $f_{\text{QM}}^{(n)}$

Initialize the training corpus P = [];
for $k \leftarrow 1, n$ **do**
  Initialize the SFT Dataset T = [];
  foreach *instance* $X_i \in D$ **do**
    $F_i, C_i, Q_i \leftarrow X_i$; // Fact F_i, Court View C_i and Query Q_i
    $R \leftarrow \texttt{Retrieval}(Q_i)$; // Cascading Retrieval R
    $O \leftarrow f_{\text{QM}}^{(k-1)}(Q_i, R)$; // Generate questions and options O
    $A \leftarrow \texttt{Simulator}(O, F_i)$; // Simulate user choices A
    foreach *indicates* $ind \in \texttt{Combination}(p, m)$ **do**
      $S_{ind} \leftarrow \texttt{Evaluator}(O_{ind}, A_{ind}, C)$; // Calculate the generation probability
      $P_i \leftarrow P_i \cup \{(O_{ind}, A_{ind}, S_{ind})\}$; // Expand the training corpus
    $O, A, C \leftarrow \texttt{Sort}(P_i)_0$; // Find optimal questionnaire based on the generation probability
    $T \leftarrow T \cup \{(Q_i, O)\}$; // Add this pair to SFT Dataset
  $f_{\text{QM}}^{(k)} \leftarrow \texttt{SFT}(f_{QM}^{(0)}, T)$;

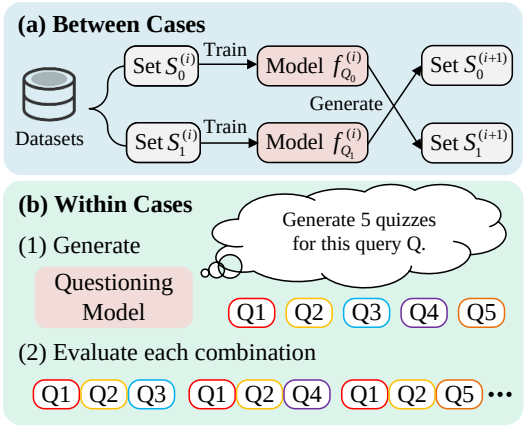

Figure 3: Strategies for ensuring questionnaire diversity. (a) Between cases: the model is trained on one subset and generates questionnaires for another. (b) Within cases: multiple quizzes are generated and evaluated to select the best combination as the final questionnaire.

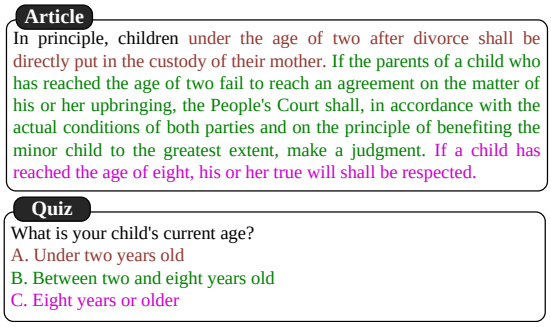

Figure 4: Convert legal provisions into quizzes. Different options correspond to different judgments as specified in the provisions.

model on one subset and then generate questionnaires for another subset. This approach prevents the model from simply replicating the questionnaire generation it learned from the training data for the same Q, ensuring it creates questionnaires that differ from those in the training set.

Secondly, at the within-case level, the questioning model generates multiple quizzes for the query Q at once, ensuring diversity within this batch. We then use combinatorial enumeration to find the optimal questionnaire combination, constructing an SFT dataset for training the questioning model.

3.2 Cascading Retrieval

Compared to open-ended questions, questionnaires allow the system to guide users through predefined options. To be effective, questionnaires must be legally meaningful, such that different user choices lead to different judgments under the law. Therefore, questionnaire generation requires grounding in relevant legal provisions.

To align retrieved provisions with questionnaire outputs, we construct a quiz database offline. This design is motivated by the structure of the law itself. Most legal provisions are expressed as conditional logic (Rahman and Dango, 2017), which can be naturally transformed into multiple-choice questions. As illustrated in Figure 4, the law specifies different approaches for children in different age groups, enabling the expert model to provide more tailored responses once the relevant condition is identified. We first filter out non-normative provisions (Hermann, 2025) that do not support legal reasoning, and then convert the remaining, condition-based provisions into questions and options.

However, directly retrieving legal provisions from user queries is challenging. Legal texts are highly formal and structured, whereas user queries

are typically colloquial and fact-oriented, resulting in low semantic similarity and reduced retrieval accuracy (Gao et al., 2024). To address this gap, we propose a **case–law–quiz** cascading retrieval framework (Figure 2b), which first retrieves similar cases based on factual similarity and then maps them to relevant legal provisions and quizzes.

First, we perform vector retrieval based on the user's query Q to identify the top-n_c similar cases $\mathbf{C}$ from the cases database.

Next, we conduct a "case–law" retrieval, calculating the frequency $\mathbf{Y}$ of each legal provision appearing in these similar cases, expressed by the following formula:

$$Y_i = \sum_{c \in \mathbf{C}} \mathbf{1}_{\{law_i \in c\}}, \tag{3}$$

where Y_i represents the frequency of the i-th legal provision, law_i represents the i-th legal provision, c is a similar case, and $\mathbf{1}_{\{\}}$ denotes the indicator function. Based on the occurrence frequency Y, we rank the legal provisions to obtain the top-n_p most cited provisions.

Thirdly, a "law–quiz" mapping is performed to retrieve the corresponding quizzes for these laws from the quiz database.

The higher the citation frequency of a provision, the more important it becomes in related cases. This suggests that we should check if the user's query has omitted information related to this provision and generate a questionnaire for supplementation. Here n_c and n_p are hyperparameters.

4 Experiments

We evaluate FactFiller through both human evaluations and downstream LegalAI performance.

4.1 Implementation Details

We report the main implementation details of our experiments. Full settings and hyperparameters are provided in Appendix A.

4.1.1 Datasets

We collected 17,865 court documents from China Judgments Online, covering 90 commonly encountered legal causes over the past two years, including 75 civil and 15 criminal categories. Among these, 90% were used for training and 10% for testing.

4.1.2 Retrieval Database

We additionally constructed a Case Database comprising 82,257 court cases that are disjoint from the training and test sets. Together with relevant legal provisions obtained from the China Legal Quick Reference Handbook[1], they jointly form the retrieval corpus for the case–law–quiz pipeline.

4.1.3 Models

The Questioning Model and Auxiliary Model were trained on the Qwen2.5-7B-Instruct model (Yang et al., 2024). We chose GPT-4o-mini (Hurst et al., 2024) to simulate users completing the questionnaires under conditions where the complete facts are available, as its moderate performance realistically reflects that non-experts may also make errors when facing complex questions. The computational costs are reported in Appendix A.6.

4.1.4 Questionnaire

Each questionnaire presents three multiple-choice questions simultaneously, without any logical dependency between them. To support flexibility, each question includes an "Other" option that allows users to skip or indicate that none of the listed choices apply.

4.1.5 Baselines

We compare **GATE** (Li et al., 2023a), which generates open-ended clarification questions and our method **FactFiller** under an ablation setting that includes four progressively enhanced configurations of questionnaire-based prompting: (1) **Multiple-Choice (Baseline)**, where the Qwen2.5-7B-Instruct model directly generates multiple-choice questions; (2) **+ Direct Retrieval**, where legal provisions are retrieved based on vector similarity and provided as additional input when generating the questionnaire; (3) **+ Cascaded Retrieval**, which replaces direct retrieval with a case–law–quiz cascade retrieval pipeline; (4) **+ Training (FactFiller)**, which further improves the questioning model through the unlabeled iterative training paradigm.

4.2 Human Evaluation

We conducted two complementary human evaluations: one with domain experts and the other with lay participants. Expert users were invited to assess the quality of the questionnaires along professional dimensions, while non-expert users were asked to evaluate their comprehensibility and usefulness.

4.2.1 Expert Evaluation

This evaluation evaluates the quality of questionnaires generated by different methods from a pro-

[1] https://github.com/LawRefBook/Laws

Method	(a) Expert						(b) Non-Expert			
	Relevance		Comprehensiveness		Helpfulness		Ease of Understanding		Expressive Support	
	Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value
Multiple-Choice (Baseline)	3.89 ± 0.16	—	3.51 ± 0.15	—	3.69 ± 0.16	—	$\mathbf{3.97 \pm 0.10}$	—	3.93 ± 0.11	—
+ Direct Retrieval	3.85 ± 0.15	$0.4882^{n.s.}$	3.58 ± 0.12	$0.2306^{n.s.}$	3.72 ± 0.13	$0.6817^{n.s.}$	3.95 ± 0.12	$0.6267^{n.s.}$	3.95 ± 0.10	$0.6308^{n.s.}$
+ Cascaded Retrieval	3.95 ± 0.14	$0.2306^{n.s.}$	3.79 ± 0.12	0.0001^{**}	4.01 ± 0.13	$<0.0001^{**}$	3.93 ± 0.11	$0.4254^{n.s.}$	3.94 ± 0.10	$0.8417^{n.s.}$
+ Training (FactFiller)	$\mathbf{4.19 \pm 0.12}$	$<0.0001^{**}$	$\mathbf{3.85 \pm 0.11}$	$<0.0001^{**}$	$\mathbf{4.04 \pm 0.12}$	$<0.0001^{**}$	3.95 ± 0.11	$0.6452^{n.s.}$	$\mathbf{4.03 \pm 0.10}$	0.0492^{*}

Table 1: Human evaluation results for (a) Expert and (b) Non-Expert evaluations. Scores are reported as mean ± 95% CI, with the best results highlighted in **bold**. Significance levels are marked as * for p < 0.05, ** for p < 0.01, and n.s. for not significant and p-values indicate statistical significance compared against baseline.

fessional legal perspective, focusing on relevance, comprehensiveness, and practical helpfulness. We recruited 10 participants with formal legal education. Each legal expert was randomly assigned 25 consultation scenarios, each consisting of an event description and a short user query. For each scenario, experts reviewed four questionnaires produced by different methods in randomized order and rated them on a 5-point Likert scale along the three dimensions. Further details on the experimental design, participants, and analysis can be found in Appendix B.1.

As shown in Table 1a, FactFiller consistently outperformed the baseline across all dimensions, improving relevance from 3.89 to 4.19, comprehensiveness from 3.51 to 3.85, and helpfulness from 3.69 to 4.04. All improvements were statistically significant ($p < 0.0001$). Cascaded retrieval further enhanced questionnaire quality, particularly in comprehensiveness and helpfulness, whereas direct retrieval alone yielded limited or non-significant gains. Overall, these results demonstrate that FactFiller produces questionnaires that are more relevant, comprehensive, and practically useful for legal experts.

4.2.2 Non-Expert Evaluation

We further evaluate the questionnaires from the perspective of lay users, focusing on ease of understanding and expressive support. We recruited 50 participants without formal legal training. Each participant was randomly assigned 10 consultation scenarios. Further details can be found in Appendix B.2.

As shown in Table 1b, FactFiller maintained comparable ease of understanding to the baseline, with no statistically significant difference (3.97 to 3.95, $p = 0.6452$), indicating that incorporating legal grounding did not increase cognitive burden. In contrast, expressive support improved significantly, with the mean score increasing from 3.93 to 4.03 ($p = 0.0492$). Although this improvement does not reach a more stringent significance threshold (e.g., $p < 0.01$), it aligns with the observation that lay users may have limited awareness of which facts are most informative for legal professionals. Overall, these results suggest that FactFiller helps lay users better articulate case facts without compromising questionnaire comprehensibility.

4.3 Downstream LegalAI Performance

We evaluated whether queries completed with questionnaires improve the performance of domain-specific LegalAI models on real tasks.

4.3.1 Downstream Tasks

We evaluated the impact of questionnaire-based query completion on three downstream tasks. For details on the implementation of downstream tasks, please refer to Appendix C.

Court View Generation Court View Generation aims to generate court views based on plaintiffs' claims and factual descriptions (Li et al., 2024b). We conducted experiments on the test split of our dataset and adopted Qwen2.5-72B-Instruct as the expert model to accommodate the long document context. Performance was evaluated using ROUGE-{1,2,L}, BLEU-{1,2,3}, and an LLM-as-judger based on Qwen-plus.

Legal Case Retrieval Legal Case Retrieval aims to retrieve relevant historical cases given a query (Leburu-Dingalo, 2024). We used the Chinese Civil Case Retrieval Dataset (Ye and Li, 2024), which contains 1,146 queries and 114,600 candidate cases. Case representations were obtained using the bge-large-zh-v1.5 embedding model (Xiao et al., 2024), and retrieval performance was evaluated using Recall@50, 100, 200, 500, 1000.

Confusing Charge Prediction Confusing Charge Prediction (Li et al., 2024a) focuses on distinguishing highly similar and easily confused

Group	Method	Court View Generation						
		ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	LLM-Judger
Baselines	Original	37.41 ± 0.38	16.89 ± 0.31	23.86 ± 0.34	18.18 ± 0.61	11.83 ± 0.42	8.42 ± 0.33	2.33 ± 0.04
	Open-ended GATE	37.86 ± 0.40	17.55 ± 0.34	24.53 ± 0.34	20.26 ± 0.64	13.35 ± 0.46	9.55 ± 0.37	2.40 ± 0.04
Ablation	Multiple-Choice	37.03 ± 0.40	16.89 ± 0.32	23.86 ± 0.34	18.86 ± 0.62	12.14 ± 0.43	8.60 ± 0.34	2.36 ± 0.04
	+ Direct Retrieval	37.64 ± 0.40	17.18 ± 0.33	24.14 ± 0.35	19.40 ± 0.64	12.68 ± 0.45	9.09 ± 0.36	2.38 ± 0.04
	+ Cascaded Retrieval	38.80 ± 0.38	17.87 ± 0.31	24.70 ± 0.34	20.46 ± 0.65	13.41 ± 0.45	9.62 ± 0.35	2.40 ± 0.04
	+ Training (**FactFiller**)	**40.51 ± 0.38**	**19.01 ± 0.32**	**25.80 ± 0.34**	**24.32 ± 0.65**	**16.12 ± 0.45**	**11.81 ± 0.36**	**2.43 ± 0.04**

Table 2: Performance on the downstream task of Court View Generation, including both baseline methods and an ablation study of FactFiller. Values are reported as mean ± 95% CI, with the best results highlighted in **bold**.

Group	Method	Legal Case Retrieval				
		Recall@50	Recall@100	Recall@200	Recall@500	Recall@1000
Baselines	Original	15.80 ± 2.11	22.35 ± 2.41	31.40 ± 2.69	48.41 ± 2.89	62.37 ± 2.80
	Open-ended GATE	18.38 ± 2.24	26.52 ± 2.56	36.84 ± 2.79	53.92 ± 2.89	68.40 ± 2.69
Ablation	Multiple-Choice	18.93 ± 2.27	27.58 ± 2.59	37.54 ± 2.81	54.98 ± 2.89	69.14 ± 2.68
	+ Direct Retrieval	18.94 ± 2.27	27.26 ± 2.58	37.33 ± 2.80	54.62 ± 2.88	68.37 ± 2.69
	+ Cascaded Retrieval	19.14 ± 2.28	27.30 ± 2.58	37.89 ± 2.81	55.55 ± 2.88	70.14 ± 2.65
	+ Training (**FactFiller**)	**19.30 ± 2.29**	**27.70 ± 2.59**	**38.10 ± 2.81**	**56.03 ± 2.88**	**70.67 ± 2.64**

Table 3: Performance on the downstream task of Legal Case Retrieval.

Group	Method	Confusing Charge Prediction					
		P-I	V-A	F & E	E & MPF	AP & DD	Average
Baselines	Original	73.75 ± 3.05	61.57 ± 5.65	84.72 ± 3.24	81.90 ± 4.11	72.77 ± 4.38	74.94
	Open-ended GATE	73.58 ± 3.05	60.28 ± 5.71	85.09 ± 3.27	78.76 ± 4.36	73.23 ± 4.40	74.19
Ablation	Multiple-Choice	75.09 ± 3.03	61.54 ± 5.77	84.53 ± 3.30	81.45 ± 4.16	73.61 ± 4.46	75.24
	+ Direct Retrieval	75.28 ± 2.97	61.89 ± 5.73	84.72 ± 3.30	79.88 ± 4.56	72.77 ± 4.49	74.91
	+ Cascaded Retrieval	75.06 ± 2.98	**61.97 ± 5.71**	84.23 ± 3.40	**83.48 ± 4.38**	75.61 ± 4.40	76.07
	+ Training (**FactFiller**)	**75.31 ± 3.06**	61.89 ± 5.70	**85.62 ± 3.30**	81.96 ± 4.22	**79.10 ± 4.38**	**76.78**

Table 4: Performance on the downstream task of Confusing Charge Prediction.

criminal charges. Following GCI (Liu et al., 2021), we evaluated five confusing charge sets (P-I, V-A, F&E, E&MPF, and AP&DD) using 2,316 samples from CAIL2018 (Xiao et al., 2018). We employed Qwen-2.5-72B-Instruct to predict charges based on ambiguous case facts and completed questionnaires, and report accuracy for each charge set as well as their average.

4.3.2 Experimental results

We first conducted experiments on three LegalAI downstream tasks to evaluate the ability to complete facts using questionnaires. The experimental results indicate that **FactFiller can effectively supplement the missing information in the user query through the three questions in the questionnaire**, aiding expert models in making more accurate judgments.

In the Court View Generation task as shown in Table 2, the ROUGE-series metrics improved by an average of 2.15, and the BLEU-series metrics improved by an average of 4.61. The GATE model for generating open-ended questions provides poor guidance to users, allowing FactFiller without training to outperform GATE.

In the Legal Case Retrieval task as shown in Table 3, Recall@50, Recall@500, and Recall@1000 increased by 3.50%, 7.62%, and 8.30% respectively. The questionnaire does not always enhance the performance of downstream tasks, **a low-quality questionnaire can disrupt expert models.** For instance, "+Direct Retrieval" may retrieve irrelevant provisions due to the difficulty in directly retrieving legal articles, leading to unrelated questions.

In the five subsets of Confusing Charge Prediction as shown in Table 4, FactFiller achieved an average improvement of 1.84%. GATE performed poorly in this task, as users' potentially inaccurate responses led to a decline in the performance of the expert model.

5 Discussion

To further validate the effectiveness and practicality of our approach, we organize the discussion around a representative case study and the training

User Query	I worked at a company, but later terminated my employment relationship with them. The company refused to acknowledge the labor relationship during my time there and claimed that they were not obligated to provide me with severance pay or compensation for unused annual leave.		
FactFiller (w training)	1. The length of time you worked at the company is: A. Less than six months B. More than six months but less than one year C. More than one year but less than six years D. More than six years but less than twelve years E. More than twelve years	2. What is the status of your labor contract with the company? A. A written labor contract has been signed. B. No written labor contract has been signed, but a labor relationship has been established. C. A labor contract was signed before employment began. D. No written labor contract has been signed, and no labor relationship has been established. E. Other.	3. Does the employer exhibit any of the following behaviors? A. Forcing the employee to work through violence, threats, or illegal restrictions on personal freedom B. Violating regulations by commanding risky operations that endanger the employee's personal safety C. None of the above

Figure 5: Questionnaires generated by various methods for a labor contract dispute consultation.

framework.

5.1 Case Study

We use a labor contract dispute consultation case to analyze FactFiller, as illustrated in Figure 5. Our FactFiller method generates three questions closely related to severance pay, each backed by relevant legal provisions. The first question inquires about the length of employment, as Article 47 of the Labor Contract Law of the People's Republic of China stipulates the standards for economic compensation based on years of service. The second question concerns the establishment of the labor contract, which relates to Article 10 regarding contract signing. The third question asks whether the employer engaged in any misconduct, which pertains to Article 38 that outlines conditions under which an employee may terminate the contract. We provide a more detailed analysis of various methods in Appendix E.

5.2 Analysis on Training Framework

Since no labeled data is available for training, we iteratively search for the best training corpus. As shown in Figure 6, over 10 iterations, the cumulative generation probabilities on both the training and test sets gradually increase, indicating that the quality of the training corpus improved with each iteration. Additionally, the stepwise probabilities of the questioning model also improve, demonstrating that **the proposed training paradigm can indeed produce a better questioning model.**

We further validated the performance of the model on the test set after being trained with the best corpus, random corpus, and worst corpus. As shown in Table 5, the results indicate that the best corpus outperforms the random corpus, and the random corpus outperforms the worst corpus. This further confirms that the questioning model can effectively learn well-designed questionnaires.

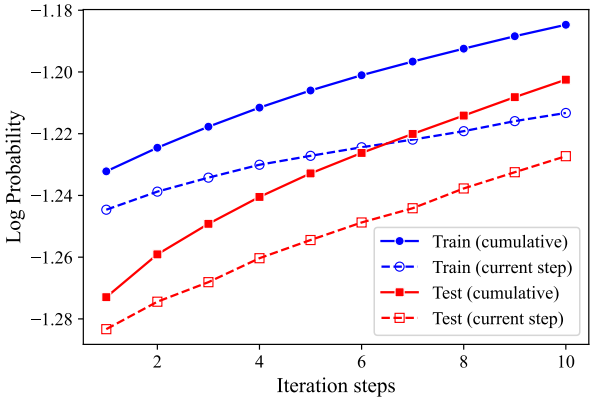

Figure 6: Log-probabilities reflecting fact completeness on the training and test sets versus iteration steps.

Strategy	Best	Random	Worst
Log Probability	**-1.23**	-1.28	-1.31

Table 5: Log-probabilities reflecting fact completeness on test-set after training with different corpora.

6 Conclusion

In this paper, we introduce FactFiller, a interaction method that guides users to provide critical case information through automatically generated multiple-choice questionnaires. FactFiller introduces a training paradigm that does not require labeled data, using the probability of court view generated by an auxiliary model based on the filled fact query as its optimization goal. Iteratively discovers and identifies the optimal corpus. Through "case–law–quiz" cascade retrieval, it ensures that the questions and options in the questionnaire are legally significant. Experimental results show that the generated questionnaires achieve high quality in terms of relevance, comprehensiveness, helpfulness, ease of understanding, and expressive support, while also ensuring the performance of expert models across three downstream LegalAI tasks.

Limitations

This paper has two potential limitations. First, the current approach generates a questionnaire with three quizzes for each user query, but not all scenarios require a questionnaire to fill in facts. Some queries may be sufficiently complete, requiring fewer quizzes or none at all. In real-world applications, effectively triaging user queries could enhance the user experience. However, this paper focuses on questionnaire generation, so this issue is not discussed.

Second, this paper primarily focuses on Chinese laws and judicial documents and does not include experimental data for other regions. Theoretically, our framework can be applied to any language and country, as FactFiller is based on LLMs and is not constrained by language. However, the focus of this paper is not on multilingual capabilities, so no experiments were conducted to discuss related issues.

Ethics Statement

LegalAI systems have the potential to enhance fairness and accessibility in the legal domain, facilitating legal consultations for individuals without legal expertise. By leveraging automated systems, these solutions can democratize legal assistance and provide more equitable access to legal services. Our proposed method, FactFiller, contributes to the development of LegalAI by enhancing the completeness of user queries, ensuring that even vague queries are transformed into useful and actionable information for downstream tasks. Additionally, all datasets and models used in this work are publicly available. We strictly adhere to their usage policies and provide appropriate citations.

Acknowledgments

This work was supported in part by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037, 2024C01259), National Key Research and Development Program of China (2024YFE0203700), and National Natural Science Foundation of China (62376243), Key R&D Program of Hangzhou (2025SZDA0254), Ant Group, Chongqing Ant Consumer Finance Co., Ant Group through CCF-Ant Research Fund. All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in German civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027, St. Julian's, Malta. Association for Computational Linguistics.

Federico Maria Cau and Lucio Davide Spano. 2025. Exploring the impact of explainable ai and cognitive capabilities on users' decisions. *arXiv preprint arXiv:2505.01192*.

Cheng Gao, Chaojun Xiao, Zhenghao Liu, Huimin Chen, Zhiyuan Liu, and Maosong Sun. 2024. Enhancing legal case retrieval via scaling high-quality synthetic query-candidate pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Miami, Florida, USA. Association for Computational Linguistics.

Mikołaj Hermann. 2025. Application and applicability of the law in relation to its validity. In *Languages of the Law*, pages 189–202, Cham. Springer Nature Switzerland.

Yi-Hung Huang. 2018. Influence of instructional design to manage intrinsic cognitive load on learning effectiveness. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(6):2653–2668.

Yuting Huang, Meitong Guo, Yiquan Wu, Ang Li, Xiaozhong Liu, Keting Yin, Changlong Sun, Fei Wu, and Kun Kuang. 2025. Appealcase: A dataset and benchmark for civil case appeal scenarios. *arXiv preprint arXiv:2505.16514*.

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

Tebo Leburu-Dingalo. 2024. Towards a framework for legal case retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3078–3078.

Ang Li, Qiangchao Chen, Yiquan Wu, Xiang Zhou, Kun Kuang, Fei Wu, and Ming Cai. 2024a. From graph to word bag: Introducing domain knowledge to confusing charge prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7469–7479, Torino, Italia. ELRA and ICCL.

Ang Li, Yiquan Wu, Yifei Liu, Kun Kuang, Fei Wu, and Ming Cai. 2024b. Enhancing court view generation with knowledge injection and guidance. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5896–5906, Torino, Italia. ELRA and ICCL.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023a. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023b. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1035–1044.

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

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1928–1941, Online. Association for Computational Linguistics.

Yifei Liu, Yiquan Wu, Ang Li, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2024. Unleashing the power of LLMs in court view generation by stimulating internal knowledge and incorporating external knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2782–2792, Mexico City, Mexico. Association for Computational Linguistics.

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

Shahid Rahman and Bernadette Dango. 2017. Conditionals and legal reasoning. elements of a logic of law.

Juanming Shi, Qinglang Guo, Yong Liao, Yuxing Wang, Shijia Chen, and Shenglin Liang. 2024. Legal-lm: Knowledge graph enhanced large language models for law consulting. In *International Conference on Intelligent Computing*, pages 175–186. Springer.

Suxin Tong, Jingling Yuan, Peiliang Zhang, and Lin Li. 2024. Legal judgment prediction via graph boosting with constraints. *Information Processing & Management*, 61(3):103663.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

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

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Fuda Ye and Shuangyin Li. 2024. Milecut: A multi-view truncation framework for legal case retrieval. In *Proceedings of the ACM on Web Conference 2024*, pages 1341–1349.

Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, and Tat-Seng Chua. 2024. Ask-before-plan: Proactive language agents for real-world planning. *arXiv preprint arXiv:2406.12639*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Experiment Settings

We report implementation details, including model settings and prompts.

A.1 Settings of LLMs

All large language models (LLMs) use the default system prompt `"You are a helpful assistant."` We set top-p to 0.8 and temperature to 0.7 for all models, except GPT-4o, which is used for automated evaluation in terms of practicality, comprehensiveness, and guidance, and runs with temperature set to 0 to ensure deterministic scoring.

A.2 Hyper-parameters of FactFiller

The training and test datasets used by FactFiller are summarized in Table 6. FactFiller iterates for $n = 10$ rounds. For the training data, it uses within-cases level generation to generate $p = 6$ questions at once, selecting $m = 3$ questions to form the questionnaire. For the test data, it generates $p = 3$ questions at once, selecting $m = 3$ questions to form the questionnaire.

A.3 Settings of Supervised Fine-tuning

The auxiliary model was trained on 111,612 court judgment documents, which do not overlap with the FactFiller training/test sets or the retrieval database. The model takes the fact description as input and learns to generate the reasoning and judgment sections of the court decision.

We fine-tuned the questioning model and the auxiliary model from Qwen2.5-7B-Instruct using the llama-factory framework (Zheng et al., 2024) with 4 NVIDIA H800 GPUs and LoRA, setting the LoRA rank to 32 and LoRA alpha to 64.

The optimization was done using DeepSpeed ZeRO-3 and the AdamW optimizer, with a learning rate of 5×10^{-5}. The total train batch size is set to 16 (4 per GPU), and training was conducted for 1 epoch.

Type	Dataset
# Types of Case Causes	90
# Training Cases	16,078
# Test Cases	1,787
Avg. Length in Query	175
Avg. Length in Fact	1,472
Avg. Length in Court View	993

Table 6: Dataset Statistics.

A.4 Prompts for questioning model

A.4.1 Questioning without retrieval

The "Multiple Choice" method without retrieval uses the following prompts.

> You are a professional lawyer, and a user seeks legal advice from you. You need to ask the user 3 multiple-choice questions to gain a more comprehensive understanding of the situation.
> # User Query
> {query}
> # Multiple-Choice Questions
> In this section, please provide 3 multiple-choice questions based on the user's inquiry, with 3-5 options for each question.

A.4.2 Questioning with retrieval

The retrieval-based methods, including "+ Direct Retrieval", "+ Cascaded Retrieval", and "+ Training (FactFiller)", share the following prompt.

> You are a professional lawyer, and a user seeks legal advice from you. You need to ask the user 3 multiple-choice questions to gain a more comprehensive understanding of the situation.
> # Reference
> {reference}
> # User Query
> {query}
> # Multiple-Choice Questions
> In this section, please provide 3 multiple-choice questions based on the user's inquiry, with 3-5 options for each question.

A.5 Prompts for Simulated User Answers

The following are the prompts used by the model to simulate answering the questionnaire based on the complete facts.

> You are a user who is seeking legal consultation, and you know the information provided in the content below. You need to answer several multiple-choice questions.
> # Content
> {content}
> # Questions and Options
> {questionnaire}

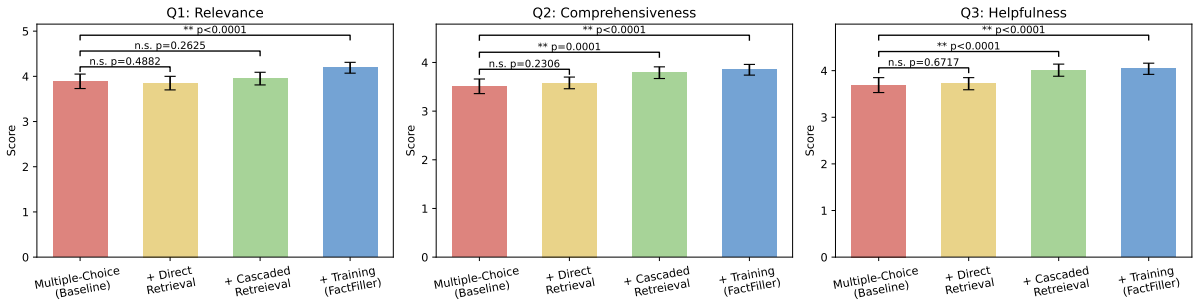

Figure 7: Expert evaluation of Relevance, Comprehensiveness, and Helpfulness across different methods. Error bars indicate 95% confidence intervals. Significance levels are marked as * for $p < 0.05$, ** for $p < 0.01$, and n.s. for not significant.

> # Reasoning
> Analyze each question and explain your reasoning for the chosen answer to each question.
> # Answers
> Provide your final answers here, separated by spaces.

A.6 Efficiency and Cost Analysis

A.6.1 User-side latency

From the user perspective, FactFiller introduces only minimal latency. On average, questionnaire generation takes 0.191 seconds, including 0.176 seconds for case–law–quiz retrieval and 0.015 seconds for the questioning model's first-token generation. The average time to produce a complete questionnaire of three questions with options is 1.248 seconds. For comparison, in the downstream task of Court View Generation (CVG), the overall response time is 16.775 seconds, meaning that FactFiller adds only a 7.4% increase in total processing time.

A.6.2 Provider-side resource usage

On the provider side, training can be completed within a single day using 4×NVIDIA H800 GPUs. The auxiliary model required approximately 4,956,909 GFLOPs and took 612 minutes (about 10.2 hours). Each self-training round of the questioning model required 506,846 GFLOPs, and a total of 10 rounds took about 10 hours. Training times may vary depending on the specific hardware configuration.

A.6.3 Estimated monetary cost

The estimated monetary cost of FactFiller is also relatively low. Renting 4×NVIDIA H100 GPUs for 24 hours on Lambda Labs is quoted at around $239.04 USD, and the cost for H800 GPUs would be lower. In addition, simulated user responses were generated using GPT-4o-mini across 17,865 instances over 10 rounds. Based on API usage logs, the average cost was approximately $0.0005 per instance, resulting in a total cost of about $90 USD.

B Human Evaluation

B.1 Human Evaluation 1: Expert Evaluation

This evaluation was designed to test questionnaire Quality. In particular, we examined whether the questionnaires generated by different methods achieve high quality in terms of relevance, comprehensiveness, and practical helpfulness from a professional legal perspective.

B.1.1 Experiment Design

Each expert participant was first presented with an introduction to the system and the purpose of the evaluation, followed by an informed consent form. Participants were then randomly assigned to consultation scenarios, each of which included a description of an event and the corresponding short user query.

For each scenario, we presented four questionnaires generated by four different methods, with the order of presentation randomized. After reviewing each questionnaire, experts were asked to evaluate it across three dimensions, each rated on a 5-point Likert scale (higher scores indicate better quality).

Q1. **Relevance:** Is the content of the questionnaire directly relevant to the user's consultation?

Q2. **Comprehensiveness:** Does the questionnaire sufficiently cover the essential factual elements of the case, taking into account different possible situations?

ID	Gender	Age	Education	Occupation	ID	Gender	Age	Education	Occupation
NE1	M	32	High School	Engineer	NE26	M	59	Junior High and Below	Laborer
NE2	M	23	Secondary Vocational	Product Manager	NE27	M	27	Secondary Vocational	Product Manager
NE3	F	54	High School	Service Worker	NE28	M	48	High School	Service Worker
NE4	F	31	Junior College	Salesperson	NE29	F	38	High School	Salesperson
NE5	M	37	Secondary Vocational	Service Worker	NE30	F	52	High School	Freelancer
NE6	F	57	High School	Accountant	NE31	F	55	High School	Manager
NE7	F	47	High School	Laborer	NE32	M	26	High School	Salesperson
NE8	F	35	Secondary Vocational	Freelancer	NE33	F	25	High School	Freelancer
NE9	F	23	High School	Student	NE34	F	38	Secondary Vocational	Freelancer
NE10	M	43	High School	Laborer	NE35	M	39	Secondary Vocational	Freelancer
NE11	F	34	Secondary Vocational	Service Worker	NE36	F	33	Junior College	Self-Employed
NE12	F	56	High School	Retiree	NE37	M	53	High School	Laborer
NE13	M	49	High School	Laborer	NE38	M	34	Secondary Vocational	Service Worker
NE14	F	32	Junior College	Administrator	NE39	F	27	High School	Homemaker
NE15	F	44	Secondary Vocational	Salesperson	NE40	F	49	High School	Salesperson
NE16	M	27	High School	Laborer	NE41	M	39	Secondary Vocational	Product Manager
NE17	F	30	Junior College	Service Worker	NE42	M	16	Secondary Vocational	Student
NE18	F	45	Junior College	Salesperson	NE43	M	31	High School	Government Official
NE19	F	57	Junior High and Below	Retiree	NE44	F	35	Secondary Vocational	Accountant
NE20	M	39	High School	Manager	NE45	M	39	High School	Laborer
NE21	F	54	High School	Retiree	NE46	M	56	High School	Engineer
NE22	F	31	Secondary Vocational	Freelancer	NE47	F	38	Junior High and Below	Laborer
NE23	M	27	Secondary Vocational	Product Manager	NE48	M	39	Secondary Vocational	Self-Employed
NE24	M	33	Secondary Vocational	Government Official	NE49	F	34	High School	Freelancer
NE25	M	61	Secondary Vocational	Laborer	NE50	M	27	Secondary Vocational	Service Worker

Table 7: Backgrounds of non-expert participants, including gender, age, education, and occupation.

Q3. **Helpfulness:** Do the additional facts elicited by the questionnaire help in addressing the user's query?

The scenario pool contained 50 distinct consultation cases. Each expert was randomly assigned 25 scenarios, with 4 questionnaires per scenario, leading to 100 evaluations per participant.

B.1.2 Participants

We recruited 10 participants with formal legal education backgrounds, including 2 PhD holders, 4 Master's degree holders, and 4 Bachelor's degree holders. The group consisted of 8 females and 2 males. On average, participants spent 2,562 seconds completing the evaluation. Each participant received approximately $14 in compensation, corresponding to an hourly rate of about $20.

B.1.3 Results

We computed the mean scores of relevance, comprehensiveness, and helpfulness across scenarios, along with 95% confidence intervals, and conducted paired t-tests between methods. The results are summarized in Table 1a and Figure 7.

Relevance. Compared to the baseline, FactFiller with cascading retrieval and the unlabeled training paradigm achieved a significant improvement in relevance, increasing the mean score from 3.89 to 4.19 ($p < 0.0001$). This indicates that experts judged the questionnaires to be more directly aligned with the consultation queries. Notably, when legal provisions were retrieved directly, the mean score slightly decreased from 3.89 to 3.85 (not statistically significant), suggesting that direct retrieval may introduce irrelevant provisions and lower the perceived quality of the questionnaire. By contrast, cascading retrieval improved the score to 3.95, demonstrating that more accurate provision retrieval contributes positively to perceived relevance.

Comprehensiveness. FactFiller significantly improved comprehensiveness, with mean scores rising from 3.51 to 3.85 ($p < 0.0001$). Experts noted that FactFiller questionnaires more fully captured the essential factual elements of the case and better addressed diverse possible circumstances. Direct retrieval, cascading retrieval, and the training paradigm all contributed positively.

Helpfulness. Finally, FactFiller significantly enhanced the perceived helpfulness of the questionnaires, increasing the mean score from 3.69 to 4.04 ($p < 0.0001$). Experts reported that questionnaires generated by FactFiller were more effective in eliciting legally meaningful information and in supporting responses to the user's consultation. Direct retrieval, cascading retrieval, and the training paradigm all contributed positively.

Method	Overall		Below High School		High School		Junior College	
	Ease	Support	Ease	Support	Ease	Support	Ease	Support
Multiple-Choice (Baseline)	**3.97 ± 0.10**	3.93 ± 0.11	**4.04 ± 0.15**	4.08 ± 0.12	**3.82 ± 0.17**	3.67 ± 0.21	**4.08 ± 0.33**	4.02 ± 0.25
+ Direct Retrieval	3.95 ± 0.12	3.95 ± 0.10	4.04 ± 0.16	4.12 ± 0.12	3.79 ± 0.17	3.67 ± 0.21	3.91 ± 0.38	4.03 ± 0.21
+ Cascaded Retrieval	3.93 ± 0.11	3.94 ± 0.10	4.03 ± 0.14	4.12 ± 0.13	3.77 ± 0.20	3.69 ± 0.19	3.95 ± 0.30	4.02 ± 0.30
+ Training (FactFiller)	3.95 ± 0.11	**4.03 ± 0.10**	4.01 ± 0.14	**4.13 ± 0.12**	3.78 ± 0.21	**3.82 ± 0.19**	3.94 ± 0.34	**4.29 ± 0.26**

Table 8: Non-expert evaluation of Ease of Understanding (Ease) and Expressive Support (Support) across education levels. Values are reported as mean ± 95% CI. The best result for each metric is highlighted with **bold** font.

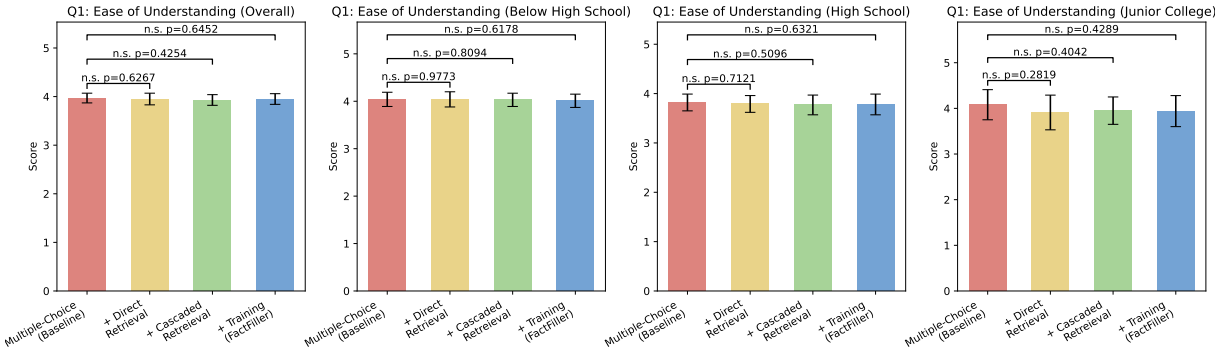

Figure 8: Non-expert evaluation of Ease of Understanding across different methods, grouped by education level.

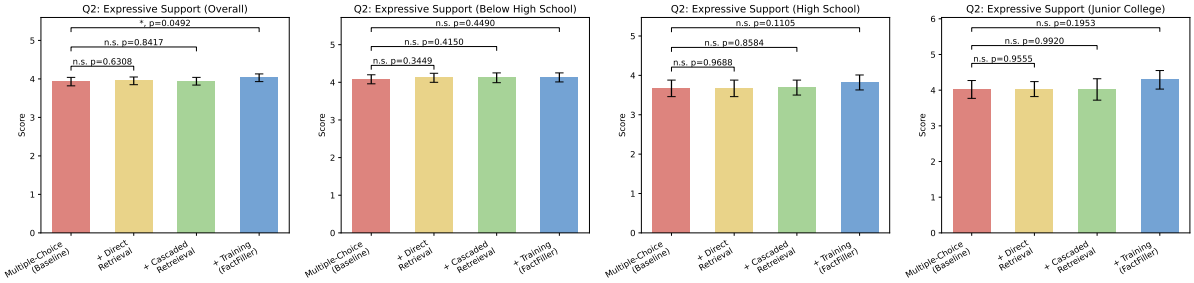

Figure 9: Non-expert evaluation of Expressive Support across different methods, grouped by education level.

B.2 Human Evaluation 2: Non-Expert Evaluation

This evaluation was designed to test comprehensibility for non-experts. In particular, we examined whether lay participants without formal legal training found the questionnaires easy to understand and whether the questionnaires supported them in expressing the facts of their cases.

B.2.1 Experiment Design

Each non-expert participant was first presented with an introduction to the system and the purpose of the evaluation, followed by an informed consent form. Participants were then randomly assigned to consultation scenarios, each of which included a description of an event and the corresponding short user query. They were asked to imagine themselves in a similar situation seeking legal advice, with the system returning a questionnaire in response to the user query.

For each scenario, we presented four questionnaires generated by four different methods, with the order of presentation randomized. After reviewing each questionnaire, participants were asked to evaluate it across two dimensions, each rated on a 5-point Likert scale (higher scores indicate better quality):

Q1. **Ease of Understanding:** Is the questionnaire easy to understand, that is, can you readily read the questions and select an appropriate option?

Q2. **Expressive Support:** Does the questionnaire help you better describe the event, that is, does it enable a lawyer to understand your situation more clearly?

The scenario pool contained 50 distinct consulta-

tion cases, consistent with the human evaluation for experts. Each non-expert was randomly assigned 10 scenarios, with 4 questionnaires per scenario, leading to 40 evaluations per participant.

B.2.2 Participants

As summarized in Table 7, we recruited 50 non-expert participants from diverse backgrounds. The sample included 26 females and 24 males, with a mean age of 39 years (range 16–61). None of the participants had received formal legal training. Educational attainment was at most junior college: 3 had completed middle school or below, 18 vocational school, 24 high school, and 5 junior college. On average, participants spent 1,180 seconds completing the evaluation. Each participant received approximately $5 in compensation, corresponding to an estimated hourly rate of $15.

B.2.3 Results

We computed overall mean scores and further analyzed ease of understanding and expressive support across education levels. For all measures, we report 95% confidence intervals and conducted paired t-tests between methods. The results are summarized in Table 8, with Ease of Understanding shown in Figure 8 and Expressive Support shown in Figure 9.

Ease of Understanding. Compared to the baseline, the ease of understanding score of FactFiller decreased slightly from 3.97 to 3.95. This minor decline was observed across all education-level subgroups, and paired t-tests confirmed that the differences were not statistically significant overall ($p = 0.6452$) or within any subgroup. Although legal provisions are generally difficult for lay users to interpret, these results suggest that cascading retrieval and the unlabeled training paradigm did not increase the difficulty of the questionnaires and did not impose additional cognitive burden on non-expert participants.

Expressive Support. In contrast, FactFiller significantly improved expressive support, with mean scores increasing from 3.93 to 4.03 ($p = 0.0492$). Subgroup analyses by education level also showed consistent upward trends, but these differences were not statistically significant. This suggests that although non-expert users generally perceived FactFiller questionnaires as more helpful for describing their situation, they often lacked awareness of which specific questions or legal elements are most informative for professionals or LegalAI systems, leading to greater uncertainty and variation in their responses.

C Implementation of Down-Stream Tasks

We evaluate FactFiller across three representative LegalAI downstream tasks.

C.1 Court View Generation

The Court View Generation (CVG) aims to generate court views based on the plaintiff's claims and the fact descriptions (Li et al., 2024b).

C.1.1 Dataset

We used the test set from our self-constructed dataset, which includes 90 causes and a total of 1,787 judicial documents.

C.1.2 Expert Model

We used the Qwen2.5-72B-Instruct model as the expert model. By concatenating the user's query with the answered questionnaire using prompts, the model outputs the court's reasoning and judgment. While it is not specifically trained on legal-domain data, existing benchmark (Huang et al., 2025) have shown that Qwen2.5-72B-Instruct performs competitively in civil law scenarios, especially in the CVG task. It has been reported to outperform both closed-source and domain-specific LLMs, demonstrating strong legal reasoning capabilities when provided with detailed factual context.

C.1.3 Prompt

The Confusing Charge Prediction task uses the following prompts:

> Based on the given document and supplementary information, please predict the court's View and output the sections titled "Reasoning" and "Judgment."
> # Document
> {query}
> # Supplementary Information
> {questionnaire}
> # Court View

C.1.4 Metrics

We evaluated performance using ROUGE similarity metrics (Lin, 2004) (including ROUGE-1, ROUGE-2, and ROUGE-L), BLEU similarity metrics (Papineni et al., 2002) (including BLEU-1, BLEU-2, and BLEU-3), as well as scoring from the Qwen-plus.

C.2 Legal Case Retrieval

Legal Case Retrieval (LCR) is a specialized information retrieval task in the legal domain. It aims to retrieve relevant legal cases from a large database of historical cases based on a given query (Leburu-Dingalo, 2024).

C.2.1 Dataset

We utilized the Chinese Civil Case Retrieval Dataset (C3RD) (Ye and Li, 2024), which contains 1,146 queries, each corresponding to 100 candidates. We processed these into a candidate pool of 114,600 entries.

C.2.2 Expert Model

We performed retrieval using vector search, employing the bge-large-zh-v1.5 model (Xiao et al., 2024) fine-tuned on the C3RD-extended[2] training dataset as the embedding model. Candidate cases were retrieved based on cosine similarity.

C.2.3 Metrics

We compared the metrics Recall@50, Recall@100, Recall@200, Recall@500, and Recall@1000 to evaluate the performance of the LCR task.

C.3 Confusing Charge Prediction

Confusing Charge Prediction involves predicting the correct charges for legal cases when the charges are highly similar and easily confused (Li et al., 2024a).

C.3.1 Dataset

We followed the settings of GCI (Liu et al., 2021), which selected five similar charge sets that are difficult to distinguish in practice. These include Personal Injury (Intentional Injury, Murder, and Involuntary Manslaughter), Violent Acquisition (Robbery, Seizure, and Kidnapping), F&E (Fraud and Extortion), E&MPF (Embezzlement and Misappropriation of Public Funds), and AP&DD (Abuse of Power and Dereliction of Duty). The data is sourced from 15% of the Chinese AI and Law Challenge (CAIL2018) (Xiao et al., 2018), comprising a total of 2,316 samples.

C.3.2 Expert Model

We selected the Qwen2.5-72B-Instruct model as the expert model, which selects a charge from the list of charges based on ambiguous facts and the answered questionnaire.

[2] https://aistudio.baidu.com/datasetdetail/205651

C.3.3 Prompts

The Confusing Charge Prediction task uses the following prompts:

> Please determine which charge from {charges} applies based on the given document and additional information.
> # Document
> {query}
> # Additional Information
> {questionnaire}
> # Charge Prediction
> Select one from {charges} and do not output any other content.

C.3.4 Metrics

We evaluated task performance by measuring the accuracy of each of the five similar charge sets individually, as well as their average accuracy.

D Proportion of Users Choosing "Other"

The questionnaire generated by FactFiller consistently includes three multiple-choice questions, each containing an "Other" option, which users can select when they are unsure how to answer.

Therefore, the proportion of users selecting "Other" also serves as an indirect indicator of questionnaire quality, reflecting both the relevance to the user's query and the ease of answering the questions. A higher proportion of "Other" selections suggests that the questions may be confusing or not well aligned with users' needs.

Figure 10 presents a line chart showing the proportion of "Other" responses across different training iteration steps. The results demonstrate that as the number of training iterations increases, the proportion of "Other" responses gradually decreases,

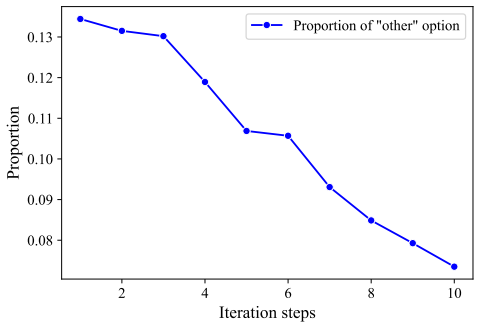

Figure 10: The proportion of simulated user selections of "Other" on the test set verus training iteration steps.

User Query	I worked at a company, but later terminated my employment relationship with them. The company refused to acknowledge the labor relationship during my time there and claimed that they were not obligated to provide me with severance pay or		
Multiple Choice	1. Why do you request the company to pay compensation and wage differences? A. The arbitration committee confirmed the labor relationship B. The company did not pay the wage difference for unused annual leave C. Request intervention from the labor inspection team D. The company did not pay wages during employment	2. What action do you plan to take next? A. Negotiate with the company B. File a lawsuit in court C. Request intervention from the labor inspection team D. Seek arbitration again from another labor arbitration institution	3. Do you need assistance from a lawyer? A. Yes B. No C. Indifferent
+ Direct Retrieval	1. Are you satisfied with the arbitration committee's decision? A. Satisfied B. Dissatisfied and plan to file a lawsuit C. Dissatisfied and plan to reapply for arbitration D. Dissatisfied and plan to negotiate with the company	2. What measures do you plan to take regarding the company's failure to pay your due compensation and wage differences? A. Negotiate with the company to reach a settlement B. File a complaint directly with the labor inspection team C. File a lawsuit with the People's Court D. Seek help and support from the labor union	3. Do you need the company to issue proof of termination or cessation of the labor contract? A. Yes B. No C. Indifferent D. Only if the company pays the compensation and wage differences
+ Cascaded Retrieval	1. What is the status of your labor contract with the company? A. A written labor contract has been signed. B. No written labor contract has been signed. C. A labor contract was signed before the termination of the employment relationship. D. No labor relationship was established. E. Other.	2. During your employment, did you take any annual leave? A. Took annual leave B. Did not take annual leave C. Uncertain D. Other	3. Has the company provided you with a labor contract? A. Provided a labor contract B. Did not provide a labor contract C. Uncertain if a labor contract was provided D. Other
+ Training (FactFiller)	1. The length of time you worked at the company is: A. Less than six months B. More than six months but less than one year C. More than one year but less than six years D. More than six years but less than twelve years E. More than twelve years	2. What is the status of your labor contract with the company? A. A written labor contract has been signed. B. No written labor contract has been signed, but a labor relationship has been established. C. A labor contract was signed before employment began. D. No written labor contract has been signed, and no labor relationship has been established. E. Other.	3. Does the employer exhibit any of the following behaviors? A. Forcing the employee to work through violence, threats, or illegal restrictions on personal freedom B. Violating regulations by commanding risky operations that endanger the employee's personal safety C. None of the above

Figure 11: Questionnaires generated by various methods for a labor contract dispute consultation.

indicating improved alignment between the generated questionnaires and users' queries.

Although the questionnaire in FactFiller without training, "+ Cascaded Retrieval", is also related to legal provisions, questions 1 and 3 are repetitive. "+ Direct Retrieval" incorrectly recalls legal provisions related to labor arbitration and fails to inquire about relevant factual information. Finally, the "Multiple Choice" method, which involves neither training nor legal retrieval, contains procedural and attitudinal questions rather than fact-related ones, resulting in relatively lower quality.

E Extended Case Study

We use a labor contract dispute consultation case to compare different methods, as illustrated in Figure 11. Training effectively enhances the quality of questionnaire generation. Although the questionnaire in FactFiller without training, "+ Cascaded Retrieval", is also related to legal provisions, questions 1 and 3 are repetitive. "+ Direct Law Retrieval" incorrectly recalls legal provisions related to labor arbitration and fails to inquire about relevant factual information. Finally, the "Multiple Choice" method, which involves neither training nor legal retrieval, contains procedural and attitudinal questions rather than fact-related ones, resulting in relatively lower quality.

F End-to-End Example of FactFiller

This section provides an illustrative end-to-end example of how FactFiller bridges the information

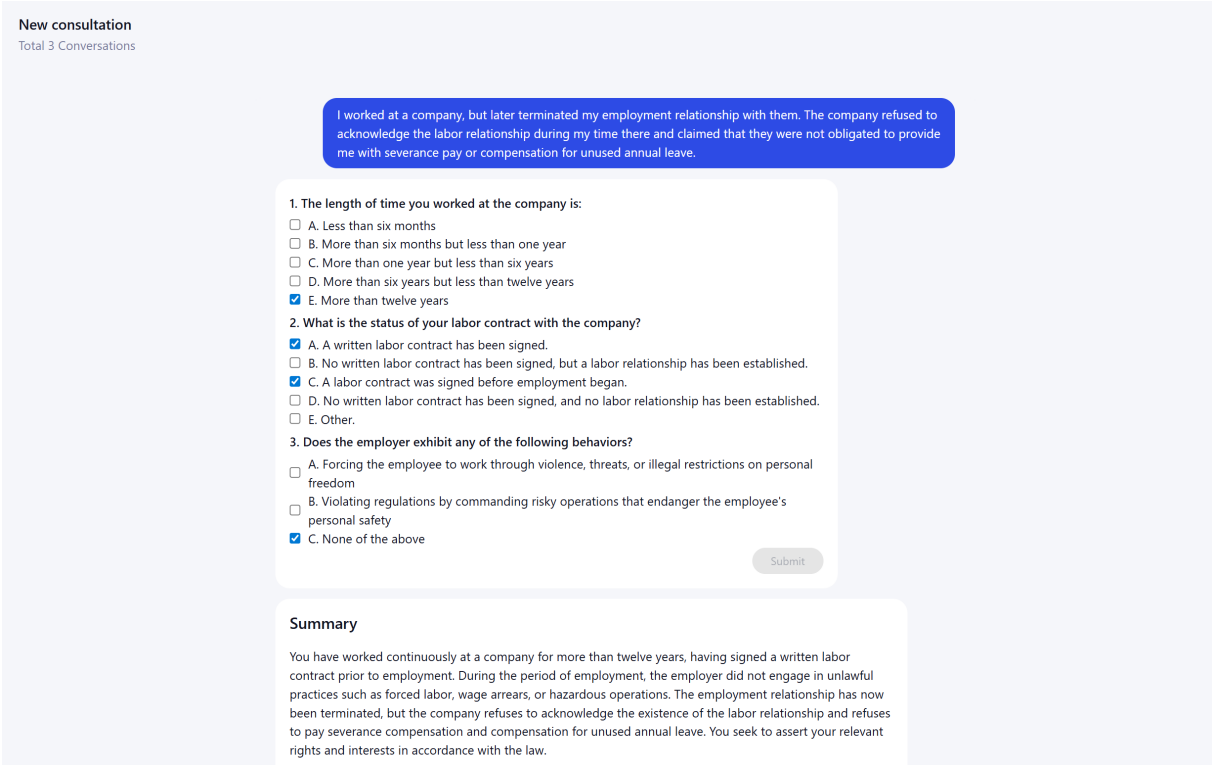

Figure 12: When non-expert users provide queries with missing facts, the system proactively asks clarifying questions through a structured questionnaire before providing an answer.

gap between non-expert users and expert LegalAI systems through questionnaire-based interaction.

As shown in Figure 12, when a user submits a vague or incomplete query, FactFiller first generates a structured multiple-choice questionnaire to proactively elicit missing legally relevant facts. The user answers the questions by selecting predefined options, which significantly reduces cognitive burden compared to open-ended clarification.

The selected answers can then be used by downstream expert models to produce more accurate and legally grounded responses.