

Case-based Statute Retrieval for Multi-turn Legal Consultation with Fact-level and Rationale-level Mixed Evidence Graphs

Wanli Wu^{1,2}, Shizhu He^{1,3,*}, Yanchao Hao^{1,3}, Jun Zhao^{1,3}, Kang Liu^{1,3}

¹The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China

²School of Future Technology, University of Chinese Academy of Sciences, Beijing, China

³School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

*Correspondence: shizhu.he@nlpr.ia.ac.cn

Abstract

Statute retrieval in multi-turn legal consultation is challenging because users often describe evolving facts in colloquial language, while statutes formulate legal conditions in abstract terms. This mismatch makes direct query–statute matching insufficient to capture the intermediate evidence needed to determine statutory applicability. Legal cases can bridge this gap by documenting how concrete facts and legal rationales support applicable statutes. However, case retrieval is noisy: factually similar cases may concern different legal issues, whereas cases with dissimilar surface facts may share the same applicability rationale. We propose FR-MEG, a Fact-level and Rationale-level Mixed Evidence Graph for turn-level statute retrieval. FR-MEG organizes case-derived evidence into two complementary views that model case–statute relationships through factual descriptions and legal rationales. The two views are calibrated and fused with full-corpus LegalDense retrieval, enabling case evidence to improve statute ranking without limiting candidates to graph-covered statutes. Experiments on LexRAG show that FR-MEG improves Recall@10 from 38.83% to 43.81% and NDCG@10 from 25.64% to 28.23% over the LegalDense backbone.

1 Introduction

Multi-turn legal consultation statute retrieval aims to retrieve and rank statutory articles applicable at the current dialogue turn from the full statute corpus, given the user’s current query and the preceding dialogue history. This task is essential for real-world legal assistance because queries from non-professional users often exhibit a substantial semantic gap between colloquial factual descriptions and formal statutory provisions (Su et al., 2024). LexRAG further shows that statute retrieval remains a critical bottleneck in multi-turn legal retrieval-augmented generation (RAG), as existing

retrievers often fail to recall the statutory articles required for reliable downstream generation (Li et al., 2025). For example, a user may describe a situation as “*my boss suddenly fired me without compensation*,” whereas the corresponding legal issues may involve concepts such as unlawful termination, economic compensation, and liabilities arising from labor disputes.

Most existing statute retrieval methods nevertheless formulate the task as direct query–statute matching. Sparse retrievers such as BM25 rely on lexical overlap between queries and statute texts (Robertson and Zaragoza, 2009), while dense retrievers encode queries and statutes in a shared semantic space (Karpukhin et al., 2020; Shao et al., 2020). Reranking and hybrid retrieval methods further improve the resulting rankings (Nogueira and Cho, 2019; Cormack et al., 2009; Chen et al., 2024), and statute-structure methods exploit relations among statutory articles (Louis et al., 2023). However, these methods primarily draw evidence from the query and the statute corpus itself, leaving the semantic gap between lay factual narratives and formal statutory language unresolved. For example, a user may describe an unpaid private loan as “*fraud*,” although the legally relevant issue may instead be a civil lending dispute. Direct matching may therefore overemphasize criminal statutes while missing the applicable civil provisions. This limitation arises not only from imperfect matching models, but also from the absence of intermediate evidence connecting informal factual descriptions to statutory applicability.

Prior cases provide such intermediate evidence by documenting how concrete factual circumstances are associated with legal issues and applicable statutory provisions. IL-PCSR demonstrates the inherent relationship between prior-case retrieval and statute retrieval: legally similar cases often cite similar statutes, and exploiting this dependency can improve retrieval performance (Paul

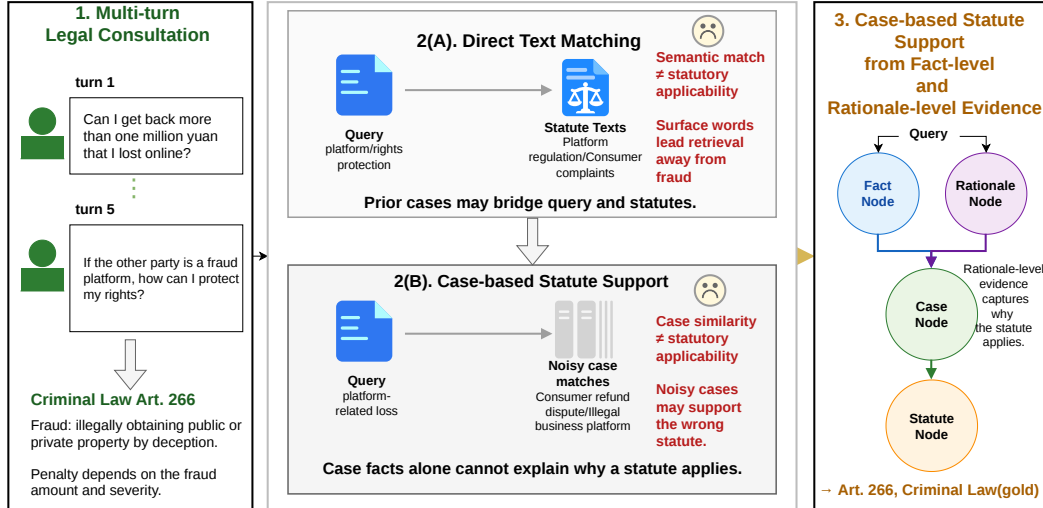


Figure 1: Motivating example of case-based statute retrieval. Direct query–statute matching may fail when colloquial descriptions differ from statutory language, whereas direct case retrieval may introduce factually similar but legally irrelevant cases.

et al., 2025).

Nevertheless, case retrieval is unreliable when treated as a standalone support signal. Unlike general textual similarity, legal case similarity depends not only on factual resemblance, but also on the legal rationale connecting facts to statutory consequences. Cases with similar factual patterns may involve different legal issues or applicable statutes, whereas cases with different surface facts may share similar legal rationales. Recent studies similarly show that effective legal retrieval benefits from modeling legal issues, legal elements, structured relations, and reasoning paths beyond surface factual overlap (V et al., 2025; Kim et al., 2025; Li et al., 2026; Tang et al., 2025; Kondo et al., 2025).

As illustrated in Figure 1, prior cases can bridge a consultation query to applicable statutes, but factually similar yet legally mismatched cases may also introduce misleading signals. To use case evidence more selectively, we organize it into fact-level and rationale-level views in **FR-MEG** (Fact-level and **R**ationale-level **M**ixed **E**vidence **G**raph).

Building on this graph, we develop **FR-MEG Retrieval**, a graph-supported framework for case-enhanced statute retrieval in multi-turn legal consultation. Rather than directly transferring statutes from retrieved cases, the framework treats case-derived graph rankings as calibrated auxiliary signals and fuses them with full-corpus LegalDense retrieval. This design allows case evidence to improve statute ranking without restricting the candi-

date space to graph-covered statutes. Experiments on LexRAG (Li et al., 2025) show that FR-MEG Retrieval consistently improves the LegalDense backbone. Diagnostic analyses further examine the robustness of these improvements and the effects of graph coverage. Our main contributions are as follows:

FR-MEG Construction. We construct a dual-view mixed-evidence graph that organizes prior-case knowledge into structured fact-level and rationale-level representations, allowing cases to provide targeted evidence for statute retrieval rather than undifferentiated textual matches.

Graph-Supported Full-Corpus Retrieval. We develop a calibrated fusion framework that integrates graph-derived evidence with full-corpus dense retrieval. It preserves statutes retrieved outside the graph while allowing supported statutes to receive additional evidence from relevant cases.

Comprehensive Evaluation and Analysis. We evaluate FR-MEG Retrieval on LexRAG against lexical, dense, hybrid, reranking, statute-graph, and case-expansion baselines. It improves Recall@10 from 38.83% to 43.81% and NDCG@10 from 25.64% to 28.23% over the LegalDense backbone. Grouped five-fold cross-validation, link-threshold sensitivity experiments, and coverage analyses further examine its robustness and the conditions under which case evidence is effective.

2 Related Work

2.1 Legal Statute Retrieval for Legal Consultation

Statute retrieval is a central component of retrieval-augmented legal consultation systems because downstream answers must be grounded in applicable legal provisions, following the broader RAG paradigm of grounding generation in retrieved external evidence (Lewis et al., 2021). Recent benchmarks have moved this task closer to real user scenarios. STARD introduces statute retrieval queries derived from non-professional legal consultations, showing that real-life queries differ substantially from formal legal questions (Su et al., 2024). LexRAG further extends this setting to multi-turn legal consultation and shows that retrieval remains a key bottleneck for legal RAG systems (Li et al., 2025). Beyond benchmarks, G-DSR incorporates statute topology into dense statute retrieval through graph neural networks (Louis et al., 2023), while IL-PCSR studies prior-case and statute retrieval jointly (Paul et al., 2025). These studies establish statute retrieval as a realistic and important task, but existing approaches mainly focus on direct query–statute matching, statute topology, or dependencies between case and statute retrieval, without jointly modeling fact-level and rationale-level case evidence for full-corpus statute ranking. In contrast, our work targets turn-level statute retrieval in multi-turn consultation and builds a unified fact- and rationale-level evidence graph for calibrated statute ranking.

2.2 Retrieval Methods for Statute Retrieval

Existing retrieval methods provide strong baselines for statute retrieval. Non-graph approaches include sparse lexical matching, dense bi-encoder retrieval, hybrid sparse–dense fusion, and cross-encoder reranking (Robertson and Zaragoza, 2009; Karpukhin et al., 2020; Cormack et al., 2009; Nogueira and Cho, 2019). These methods primarily estimate relevance through query–document matching, and therefore may miss legal applicability evidence that is not directly expressed in the query or statute text.

Graph-structured retrieval offers a complementary direction by organizing documents, entities, or legal units through relational structures (Peng et al., 2024; Zhang et al., 2025). Recent graph-based RAG methods, such as GraphRAG, LightRAG, HippoRAG, and PathRAG, use graph-enhanced

indexing or graph traversal to capture dependencies beyond isolated text chunks (Edge et al., 2025; Guo et al., 2025; Gutiérrez et al., 2025; Chen et al., 2025). In the legal domain, G-DSR incorporates the topology of statutory provisions into dense statute retrieval, showing that legislative structure can provide useful retrieval signals (Louis et al., 2023).

Our work is related to graph-enhanced retrieval in its use of structured evidence, but differs in both evidence source and retrieval objective. Rather than relying on statute topology or general document graphs, we construct FR-MEG, a case-derived fact–rationale evidence graph, and use it as calibrated auxiliary support for full-corpus statute retrieval in multi-turn legal consultation.

2.3 Legal Case Retrieval and Case-based Statute Evidence

Legal case retrieval aims to find prior cases that are legally relevant to a query case or legal question. Existing studies show that such relevance cannot be reduced to ordinary semantic similarity. LeCaRD and LeCaRDv2 provide Chinese legal case retrieval benchmarks with expert relevance annotations, while SAILER shows the importance of case structure and legal elements (Ma et al., 2021; Li et al., 2023b,a). More recent work further shows that factual overlap alone is insufficient: AQGr observes that legal issues are often under-modeled (V et al., 2025), while LegalSearchLM and GLIER show that lexical, embedding-based, or black-box dense matching may retrieve legally irrelevant cases without explicit modeling of legal elements and juridical logic (Kim et al., 2025; Li et al., 2026). ReaKase-8B similarly suggests that legal knowledge and rationale-oriented representations can help retrieve legally meaningful cases (Tang et al., 2025).

These findings are important for statute retrieval because prior cases can provide useful but noisy evidence. IL-PCSR shows that similar cases tend to cite similar statutes and that exploiting this dependence can improve retrieval (Paul et al., 2025). However, this relation should be used carefully: factually similar cases may support different statutes when their legal issues or applicability logic diverge. Facts-to-law knowledge graph studies further suggest that linking factual descriptions, legal norms, and reasoning paths can better capture such applicability patterns (Kondo et al., 2025). In our setting, LeCaRDv2 and ClaimGen-CN provide

the case foundation for constructing case-derived evidence: LeCaRDv2 provides criminal cases for legal case retrieval, while ClaimGen-CN expands the coverage to civil cases with factual narratives and legal claims (Li et al., 2023b; Zhou et al., 2025). Building on this line of work, we treat prior cases not as direct retrieval outputs to be transferred, but as structured fact- and rationale-level evidence for supporting statute ranking.

3 The Proposed Method

3.1 Task Formulation

We study turn-level statute retrieval in multi-turn legal consultation. Let the dialogue history before turn t be $H_t = \{(q_1, a_1), \dots, (q_{t-1}, a_{t-1})\}$, where q_i denotes a user query and a_i denotes the system response. Given the current user query q_t and its dialogue history H_t , the retriever ranks statutory articles from a statute corpus $\mathcal{S} = \{s_1, \dots, s_N\}$. Each turn is associated with a set of applicable gold statutes $\mathcal{G}_t \subseteq \mathcal{S}$, which may contain one or more statutes. The objective is to produce an ordered ranking of statutes in which the gold statutes appear as high as possible:

$$\pi_t = \text{rank}(q_t, H_t, \mathcal{S}).$$

3.2 FR-MEG Construction

Figure 2 presents the overall FR-MEG framework. FR-MEG is a case-derived evidence graph that organizes prior cases, fact-level evidence, rationale-level evidence, and statutory provisions. We construct it from two heterogeneous Chinese legal case datasets: LeCaRDv2 for criminal cases and ClaimGen-CN for civil cases.

Unified case representation. Each aligned case is converted into a CASERECD:

$$c = (\text{id}, d, f, \mathcal{R}, \mathcal{L}, \mathcal{M}),$$

where id is the case identifier, d is the legal domain, f is a fact summary, $\mathcal{R} = \{r_1, \dots, r_m\}$ is a set of atomic rationale elements, $\mathcal{L}$ is the set of linked statutes, and $\mathcal{M}$ stores metadata such as civil causes or criminal charges. Criminal cases provide structured charges, facts, and statute references, which are directly normalized as case–statute evidence links. Civil cases are often claim-centered and lack explicit statute citations, so we use LLM-assisted normalization to extract rationale elements and infer candidate statute links. Candidate links

are subsequently filtered by an LLM-based verifier using label- and confidence-based acceptance rules.

Graph nodes and edges. FR-MEG uses case nodes as hubs and connects them to fact nodes, rationale-level nodes, and statute nodes. Each case has a fact node represented by its fact summary, and its legal reasoning is decomposed into three to five atomic rationale elements. The main evidence paths are fact $\rightarrow$ case $\rightarrow$ statute and rationale node $\rightarrow$ atomic rationale $\rightarrow$ case $\rightarrow$ statute, allowing fact-level and rationale-level matches to support statutes through prior cases.

Rationale-level hierarchy. The rationale-level layer is built from atomic rationale elements extracted from cases. We pool these elements across cases, embed them, and cluster them into leaf rationale-pattern nodes. Each leaf node represents a fine-grained legal applicability pattern shared by a group of atomic rationale elements. To make these patterns retrievable, we summarize each leaf node with an LLM-generated label, keywords, query aliases, and a short applicability description. We then recursively merge related leaf nodes into higher-level hierarchy nodes, producing a multi-level rationale hierarchy. During graph construction, each atomic rationale element keeps its link to the original case. Thus, a matched hierarchy node can be expanded to its descendant leaf nodes, and matched leaf nodes can be traced through inverted links to atomic rationale elements, source cases, and finally linked statutes.

Indexes for retrieval. To support efficient retrieval, we build specialized indexes over FR-MEG: a fact embedding index for matching query facts to case fact nodes, a rationale-pattern index for matching query legal elements to pattern labels, query aliases, and centroids, an inverted index for tracing matched rationale nodes to atomic rationale elements and source cases, and a case–statute table for expanding cases to linked statutes. These indexes provide fact-level evidence from similar case facts and rationale-level evidence from shared legal applicability patterns.

3.3 FR-MEG Retrieval

Given a query turn (H_t, q_t) , FR-MEG Retrieval combines three parallel routes: full-corpus LegalDense retrieval, Fact-level Support, and Rationale-level Support.

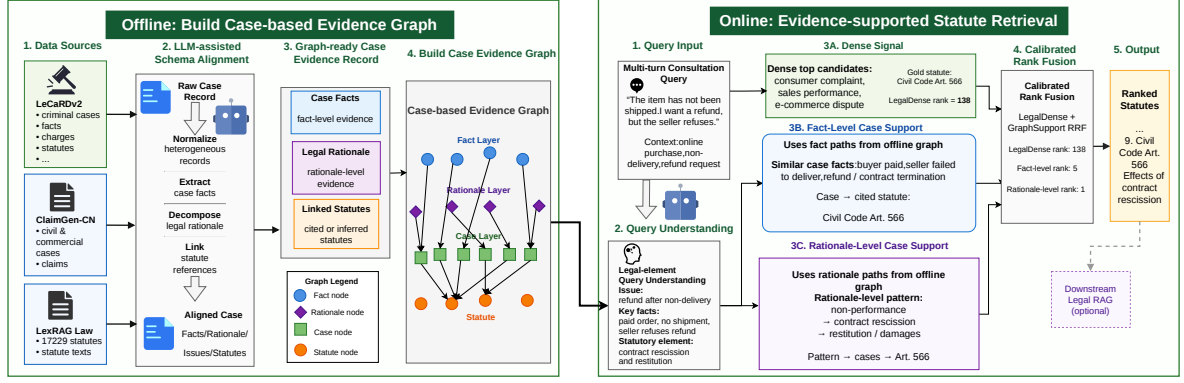


Figure 2: Overview of FR-MEG construction and retrieval. The offline stage normalizes heterogeneous legal cases into graph-ready evidence records and builds a case-based evidence graph with fact, rationale, case, and statute layers. The online stage combines full-corpus dense statute retrieval with fact-level and rationale-level case support through calibrated rank fusion.

LegalDense backbone. LegalDense converts the dialogue history and current query into a legal-element query view, embeds it, and retrieves statutes from the full corpus:

$$\text{score}_{\text{dense}}(s) = \cos(\mathbf{e}_{q_t}, \mathbf{e}_s),$$

where $\mathbf{e}_{q_t}$ is the embedding of the legal-element query view and $\mathbf{e}_s$ is the statute embedding. This route preserves full-corpus coverage when graph evidence is unavailable.

Fact-level Support. Fact-level Support matches the query against the fact embedding index, maps retrieved fact nodes to cases, and expands these cases to linked statutes. It captures factual similarity and mainly serves as a candidate-expansion signal.

Rationale-level Support. Rationale-level Support matches the legal-element query against rationale-pattern labels, query aliases, and centroid representations. Matched rationale nodes are expanded through inverted links to leaf rationale-pattern nodes, atomic rationale elements, cases, and linked statutes. This route captures legal applicability patterns that may be missed by surface factual similarity.

Calibrated graph support. The two graph routes are combined into a single graph-support ranking. Let $r_f(s)$, $r_{\text{rat}}(s)$, $r_d(s)$, and $r_g(s)$ denote the ranks of statute s in Fact-level Support, Rationale-level Support, LegalDense, and graph-support rankings, respectively. We compute an inner reciprocal rank fusion (RRF) score (Cormack

et al., 2009):

$$\text{score}_{\text{graph}}(s) = \frac{w_f}{k_g + r_f(s)} + \frac{w_{\text{rat}}}{k_g + r_{\text{rat}}(s)}, \quad (1)$$

where the corresponding term is set to zero if a statute is absent from a route. We use $k_g = 30$, $w_f = 0.3$, and $w_{\text{rat}} = 0.6$. The inner RRF score induces a graph-support ranking over statutes, whose rank is denoted by $r_g(s)$.

Final fusion. Finally, FR-MEG Retrieval fuses the full-corpus LegalDense ranking with graph support:

$$\text{score}_{\text{final}}(s) = \frac{w_l}{k + r_d(s)} + \frac{w_g}{k + r_g(s)}. \quad (2)$$

We use $k = 20$, $w_l = 1.0$, and $w_g = 0.65$. If a statute is absent from the graph-support ranking, the graph term is set to zero.

Because case-derived evidence is incomplete, graph evidence is used as calibrated support rather than as a hard retrieval constraint: LegalDense preserves full-corpus coverage, while Fact-level Support and Rationale-level Support promote statutes with supporting case evidence.

4 Experimental Setup

4.1 Dataset and Metrics

Task and evaluation data. We evaluate turn-level statute retrieval in multi-turn legal consultation on LexRAG (Li et al., 2025). The retrieval corpus contains 17,229 statutory articles, and the evaluation set contains 5,065 consultation turns: 4,781 turns have at least one gold statute and 284 turns have no gold annotation. We use the 4,781 gold-bearing

Scope	Queries	Gold	Meaning
RAW_ALL	4,781	5,499	Turns with at least one annotated gold statute; used for the main evaluation.
A-ALL-COVERED	1,809	1,954	All gold statutes are covered by FR-MEG.
B-PARTIALLY-COVERED	169	417	Some gold statutes are graph-covered.
C-NONE-COVERED	2,803	3,128	No gold statute is graph-covered.
B-COVERED-GOLD	169	214	Covered gold mentions within B; diagnostic only.
B-UNCOVERED-GOLD	169	203	Uncovered gold mentions within B; diagnostic only.

Table 1: Post-hoc graph coverage statistics. Only 39.43% of unique-per-query gold mentions are graph-covered, so FR-MEG is evaluated as calibrated support over full-corpus LegalDense retrieval rather than as hard graph routing.

turns as the primary RAW_ALL scope for Recall, while corrected NDCG values use all 5,065 turns when explicitly marked.

External evidence used to build FR-MEG. FR-MEG is built from external case evidence, not LexRAG evaluation queries. Graph construction uses 262,940 cases: 55,192 criminal cases from LeCaRDv2 (Li et al., 2023b) and 207,748 civil cases from ClaimGen-CN (Zhou et al., 2025). After schema alignment, the graph contains 262,940 case fact nodes, 1,249,624 atomic rationale nodes, 12,465 rationale leaf clusters, 26,894 rationale hierarchy nodes, and 17,229 statute nodes. Accepted case–statute links cover 1,244 statutes, making graph evidence incomplete relative to the full statute corpus.

Coverage strata. Because case-derived graph evidence does not cover every statute, we conduct a post-hoc graph-coverage analysis. Turns are split into A-ALL-COVERED, where all gold statutes are graph-covered; B-PARTIALLY-COVERED, where only some gold statutes are graph-covered; and C-NONE-COVERED, where no gold statute is graph-covered. These strata are diagnostic only and are never used for inference-time routing. Table 1 summarizes the coverage space.

Metrics. We report Recall@ K and NDCG@ K for $K \in \{1, 3, 5, 10, 30, 50, 100\}$. Recall is micro-averaged over gold statute mentions and reported on RAW_ALL unless otherwise stated. Corrected NDCG uses all 5,065 turns in the denominator, with no-gold turns contributing zero. Baseline NDCG values derived from gold-query averages are multiplied by 4781/5065. To examine possible over-retrieval, we additionally report Precision@10 and F1@10. We assess stability using grouped five-fold cross-validation over 1,013 dialogues (5,065 turns), with no dialogue shared across folds (seed=20260711). For each held-out fold, we select among 64 FR-MEG configurations on the other four folds using micro Recall@10 and

evaluate the selected configuration on the held-out fold. The reported OOF results pool predictions from all five held-out folds. Route ablations reuse the full-model-selected parameters without separate retuning.

4.2 Baselines

Compared systems. We compare FR-MEG with lexical, dense, hybrid, reranking, statute-graph, and case-expansion baselines. BM25 is implemented with bm25s using the Lucene variant with $k_1 = 1.3$ and $b = 0.75$ (Robertson and Zaragoza, 2009). Query likelihood baselines use Jelinek–Mercer smoothing with $\lambda = 0.9$ and Dirichlet smoothing with $\mu = 1000$ (Ponte and Croft, 1998; Zhai and Lafferty, 2004). Direct Dense embeds the original dialogue input without legal-element transformation, whereas LegalDense uses the legal-element query view described in Section 3.3. BM25+Dense RRF uses $k = 60$ with equal BM25 and dense weights (Karpukhin et al., 2020; Cormack et al., 2009). The no-graph reranking baseline applies bge-reranker-v2-m3 to the top-100 BM25+Dense candidates (Nogueira and Cho, 2019; Liu et al., 2025; Chen et al., 2024). The statute-structure graph baseline propagates scores over neighboring statutes with $\alpha \in \{0.05, 0.1, 0.2, 0.3\}$ and one or two hops (Louis et al., 2023). The case-expansion baseline retrieves top- K fact/case nodes for $K \in \{50, 100, 200, 500\}$ and keeps at most 20 linked statutes per case (Paul et al., 2025). Pure Case Expansion uses only the resulting case-derived statute ranking. BM25+Dense+Case adds this case-derived ranking as a third route to equal-weight RRF with $k = 60$; its reported configuration uses $K = 50$ and keeps at most 20 linked statutes per case. To test whether graph gains can be explained by statute popularity alone, the Graph-Frequency Prior baseline ranks statutes by their accepted case-link frequency and fuses this rank-

ing with LegalDense using RRF with $k = 20$ and weights 1.0/0.65.

4.3 Implementation Details

Representations, models, and configuration. All statute retrieval methods rank the same 17,229-article corpus. We use the history-aware standalone query provided by the official LexRAG rewriter and do not re-concatenate the raw dialogue history. Fact-level Support directly encodes this standalone query. The query rewriter produces three to six rationale tags; LegalDense uses at most three tags, while Rationale-level Support uses all generated tags. Dense statute retrieval reuses the cached Qwen2-1.5B-family law embeddings from the official LexRAG dense setting (Li et al., 2023c). For newly computed query and case embeddings, we use the local gte-Qwen2-1.5B encoder, producing 1,536-dimensional normalized embeddings. We use DeepSeek v4 Flash (DeepSeek-AI, 2026) for LLM-assisted schema alignment and civil statute-link verification. Candidate links with cosine similarity of at least 0.57 are sent to the verifier, with at most three candidates reviewed per case and a top-1 fallback when no candidate passes the gate. The verifier assigns each link a label and a confidence score on a 0–100 scale: STRONG links are retained, REJECT links are discarded, and PARTIAL links are retained only when their score is at least 60 (Ziems et al., 2024; Wang et al., 2024). We additionally vary the acceptance threshold for PARTIAL links over {40, 50, 60, 70, 80} in the sensitivity analysis. The dense baseline keeps at most three legal-element tags per query. Fact-level Support searches a fact embedding index built from 262,940 case fact summaries with the prefix “case fact:”. Rationale-level Support uses the cached rationale hierarchy built with the 1.5B Qwen-family encoder, whose leaf clusters are organized into L1/L2/L3/root hierarchy nodes with cached DeepSeek v4 Pro (DeepSeek-AI, 2026) summary labels and centroid representations. All graph lookups are served from JSONL/SQLite indexes rather than an in-memory graph. FR-MEG Retrieval first fuses Fact-level Support and Rationale-level Support with inner RRF using $k = 30$ and weights 0.3/0.6, then fuses graph support with full-corpus LegalDense using outer RRF with $k = 20$ and weights 1.0/0.65. The final ranked list is truncated to the top 100 statutes for evaluation, and no reranker is used.

Efficiency and deployment considerations. FR-

MEG separates costly graph construction from on-line retrieval. Case normalization, civil statute-link verification, rationale extraction, hierarchy construction, and representation encoding are performed offline, and their outputs are cached for reuse. At retrieval time, relative to the LegalDense backbone, FR-MEG additionally performs nearest-neighbor searches over the fact and rationale indexes, follows precomputed case–statute links, and applies two lightweight RRF operations. It therefore does not repeat the LLM-assisted graph-construction procedure for each consultation query. This design trades offline preprocessing and index storage for improved retrieval effectiveness, increasing Recall@10 from 38.83% to 43.81%. The graph can also be updated independently when new cases become available, without changing the full-corpus LegalDense index. Nevertheless, we do not report a wall-clock or monetary cost comparison because the graph was constructed using cached outputs and models under different hardware configurations.

5 Results and Analysis

5.1 Overall Retrieval Performance

Table 2 shows that FR-MEG Retrieval achieves the best overall performance on RAW_ALL, outperforming lexical, dense, hybrid, reranking, statute-graph, case-expansion, and triple-fusion baselines across all reported metrics. Compared with LEGALDENSE, FR-MEG improves Recall@10 from 38.83% to 43.81% and NDCG@10 from 25.64% to 28.23%, with R@5 increasing from 30.17% to 33.97%. These gains show that fact-level and rationale-level evidence provide useful support beyond legal-element dense retrieval.

5.2 Coverage-aware Analysis

Table 3 shows that the benefit of graph evidence is coverage-sensitive. On A-ALL-COVERED, FR-MEG improves Recall@10 over LEGALDENSE from 46.11% to 62.79%, showing strong headroom when relevant statutes are graph-reachable. On B-PARTIALLY-COVERED, FR-MEG also improves overall Recall@10, but the gains are concentrated on covered gold statutes, indicating that graph evidence mainly helps when the applicable statutes are reachable through case-derived links. On C-NONE-COVERED, however, LEGALDENSE remains stronger than FR-MEG, confirming that case-derived graph evidence

Method	R@1	R@3	R@5	R@10	R@30	R@50	R@100	NDCG@10
BM25	6.87	13.53	17.02	22.73	32.30	37.75	45.12	14.40
QLD	6.31	12.80	15.77	20.89	30.48	34.95	42.44	13.34
Direct Dense	11.18	20.60	25.93	33.39	46.32	53.17	61.92	20.49
LEGALDENSE	13.84	24.86	30.17	38.83	52.61	59.74	67.54	25.64
BM25+Dense RRF	10.64	19.66	24.55	32.02	47.14	53.72	62.16	20.70
BM25+Dense+BGE	14.20	25.02	30.84	38.30	51.74	57.77	62.16	25.94
Statute-Graph	8.91	18.24	23.91	32.52	47.90	54.65	62.96	19.89
Pure Case Expansion	5.96	11.82	14.77	19.51	23.06	23.40	23.42	12.05
BM25+Dense+Case	11.04	19.69	25.35	34.82	48.68	55.77	64.47	21.54
FR-MEG Retrieval	14.84	26.90	33.97	43.81	55.88	61.50	69.47	28.23

Table 2: Overall retrieval results, reported as percentages. Recall is evaluated on RAW_ALL, while NDCG@10 uses all 5,065 turns, with no-gold turns contributing zero.

should support rather than replace full-corpus retrieval.

5.3 Route Ablation

Table 4 isolates the contributions of the two FR-MEG routes. Adding Fact-level Support to LEGALDENSE improves raw Recall@10 from 38.83% to 42.94% and NDCG@10 from 25.64% to 27.71%. Adding Rationale-level Support gives a larger top-rank gain, reaching 43.59% Recall@10 and 28.19% NDCG@10. This suggests that factual case matching is useful, while rationale-level patterns are more directly aligned with statute applicability.

The two routes are complementary but partially overlapping. Their fusion yields the best Recall@10 and NDCG@10, although the improvement over Rationale-level Support alone is modest.

The A/C columns further confirm that graph support is most effective when relevant statutes are graph-covered, while LEGALDENSE remains stronger on C-NONE-COVERED. This supports

using graph evidence as calibrated support rather than as a hard graph-based retriever.

5.4 Robustness and Sensitivity Analysis

Using the dialogue-level grouped five-fold partition described in the experimental setup, FR-MEG outperforms LEGALDENSE in all five folds. The paired Recall@10 improvement is 5.00 percentage points, with a 95% confidence interval of [4.12, 5.91] and $p = 0.0002$. A tuned BM25+LegalDense baseline reaches 40.66% Recall@10, remaining below the 43.83% achieved by FR-MEG. These results indicate that the improvement is consistent across dialogue groups rather than being driven by a particular evaluation split.

The Precision@10 and F1@10 improvements in Table 5 show that the Recall gains are not obtained by indiscriminately increasing the number of retrieved statutes. The Graph-Frequency Prior improves Recall@10 over LEGALDENSE by 1.24 percentage points, with a 95% confidence interval of [0.87, 1.63] and $p = 0.0002$, but remains

Method	raw	A	B	C	B-cov	B-uncov	NDCG@10
Direct Dense	0.3339	0.3854	0.1966	0.3200	0.1916	0.2020	0.2049
LEGALDENSE	0.3883	0.4611	0.2470	0.3616	0.2523	0.2414	0.2564
Pure Case Expansion	0.1951	0.5097	0.1847	0.0000	0.3598	0.0000	0.1205
BM25+Dense+Case	0.3299	0.6305	0.2854	0.1480	0.4720	0.0887	0.2037
FR-MEG Retrieval	0.4381	0.6279	0.3070	0.3370	0.4065	0.2020	0.2823

Table 3: Coverage-aware key comparison. Columns raw/A/B/C report Recall@10 under RAW_ALL, A-ALL-COVERED, B-PARTIALLY-COVERED, and C-NONE-COVERED. B-cov and B-uncov split gold mentions inside B. NDCG@10 uses all 5,065 turns. A/B/C are diagnostics only, never inference-time routing.

Method	R@1	R@10	R@100	NDCG@10	A R@10	C R@10	A NDCG
LEGALDENSE	13.84	38.83	67.54	25.64	46.11	36.16	31.42
+ Fact-level Support	14.48	42.94	69.18	27.71	59.77	34.24	43.42
+ Rationale-level Support	15.04	43.59	69.38	28.19	62.18	33.76	45.29
+ Both / FR-MEG	14.84	43.81	69.47	28.23	62.79	33.70	45.92

Table 4: Route ablation with compact coverage diagnostics, reported as percentages. Both routes improve candidate coverage, while Rationale-level Support gives stronger top-rank performance. Their fusion yields the best Recall@10 and NDCG@10. A/C columns show that graph support is most effective when relevant statutes are graph-covered.

Method	R@10	R@100	N@10	P@10	F1@10
LEGALDENSE	38.83	67.54	25.64	4.22	7.60
Freq. Prior	40.06	65.76	26.90	—	—
FR-MEG	43.83	69.07	28.24	4.76	8.58

Table 5: Robustness diagnostics, reported as percentages. N@10 denotes NDCG@10, and Freq. Prior denotes the Graph-Frequency Prior. Precision@10 and F1@10 examine whether the Recall gains result from indiscriminate over-retrieval. The FR-MEG row reports pooled out-of-fold results from grouped five-fold cross-validation.

substantially below FR-MEG. This indicates that statute popularity in the case graph explains only part of the improvement.

We also vary the acceptance threshold for PARTIAL civil statute links over $\{40, 50, 60, 70, 80\}$. The corresponding numbers of retained civil links are 177,073, 142,423, 132,358, 95,135, and 91,792, respectively. Across these settings, Recall@10 remains between 43.41% and 43.99%, while NDCG@10 remains between 27.97% and 28.29%. The limited variation suggests that FR-MEG is not highly sensitive to the default PARTIAL threshold of 60.

5.5 Additional Diagnostics

Reranking diagnostic. Table 6 reranks the fixed FR-MEG top-100 candidate pool with BGE variants. Since the candidate pool is unchanged, R@100 remains fixed at 69.47%. However, reranking consistently lowers raw Recall@10 and substantially weakens performance on A-ALLCOVERED, where graph evidence is most useful. This suggests that the local reranker does not preserve the graph-derived ranking advantage, so we report FR-MEG without reranking as the main result.

Method	raw	A	C	R@100
FR-MEG	43.81	62.79	33.70	69.47
BGE text	38.57	48.00	34.65	69.47
BGE mark.	38.48	48.31	34.40	69.47
BGE real	36.72	44.06	33.92	69.47
BGE shuf.	36.22	42.02	34.53	69.47

Table 6: Fixed-pool reranking diagnostic, reported as percentages. Columns report R@10 except R@100.

Generation diagnostic. Table 7 reports an LLM-as-judge diagnostic on 500 LexRAG turns. FR-MEG obtains more wins than losses against LegalDense under both judges, although ties remain common. Adding rationale-level evidence directly to the prompt produces weaker and less stable results. Therefore, this diagnostic provides supporting evidence that retrieval gains can transfer to downstream generation, but it should not be interpreted as a conclusive generation evaluation.

Comparison	Q-W	Q-T	Q-L	D-W	D-T	D-L
FR-MEG vs. LegalDense	40.40	31.00	28.60	26.20	55.60	18.20
+Rationale text vs. FR-MEG	31.00	35.00	34.00	16.80	67.60	15.60

Table 7: LLM-as-judge generation diagnostic, reported as percentages. Q-W/T/L and D-W/T/L denote win/tie/loss rates judged by Qwen3.6-Flash and DeepSeek v4 Pro, respectively.

6 Conclusion

We presented FR-MEG, a Fact-level and Rationale-level Mixed Evidence Graph for turn-level statute retrieval in multi-turn legal consultation. FR-MEG integrates case-derived fact- and rationale-level evidence with full-corpus LegalDense retrieval, improving statute retrieval on LexRAG without restricting candidates to graph-covered statutes. Coverage-aware and robustness analyses further show that the gains are consistent and cannot be explained solely by statute popularity or the verification threshold.

7 Limitations

The main limitation of FR-MEG is coverage. Because the graph is constructed from available case datasets, only part of the LexRAG gold statute space is reachable through case-derived evidence. Graph support is therefore most effective for graph-covered turns and must be combined with full-corpus retrieval to remain robust. In addition, civil case normalization and statute linking rely on LLM-assisted preprocessing. Although confidence filtering is used to reduce noisy links and threshold sensitivity is evaluated, we do not conduct large-scale human validation of the accepted civil case–statute links, so residual systematic errors may remain. Our experiments are limited to Chinese legal resources and the LexRAG benchmark; whether the resulting evidence structures and retrieval gains generalize to other jurisdictions, languages, or legal domains remains to be studied. Finally, this work primarily focuses on statute retrieval. While we include a small LLM-as-judge generation diagnostic on 500 consultation turns, a full human evaluation of downstream legal answer quality is left for future work.

8 Ethical Considerations and Artifact Use

Intended use and risks. This work studies statute retrieval for legal consultation in a research setting. FR-MEG is intended to retrieve candidate statutes and should not be used as legal advice or as a substitute for professional legal judgment. Potential risks include retrieval errors, noisy case–statute links, and over-reliance on automatically retrieved legal provisions.

Artifacts, licenses, and data handling. We use LexRAG for evaluation, LeCaRDv2 and ClaimGenCN as external case evidence, and their associated statute/case resources under their stated licenses or terms of use. These artifacts are used consistently with their research-oriented purposes. The constructed FR-MEG graph and retrieval pipeline are intended for research on statute retrieval; any release of derived artifacts will follow the access conditions and redistribution terms of the original datasets and models. Because legal cases may contain sensitive factual descriptions, we do not attempt to identify individuals and do not release raw case text beyond what is permitted by the original datasets.

AI assistance disclosure. Generative AI tools were used for language polishing and editing support. They were not used to generate new research ideas, experimental results, or claims. All content, experiments, analyses, and final manuscript decisions were verified by the authors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376270) and the Independent Research Project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems.

References

- Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2025. [Pathrag: Pruning graph-based retrieval augmented generation with relational paths](#). *Preprint*, arXiv:2502.14902.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Preprint*, arXiv:2402.03216.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.
- DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitanaky, Robert Osazuwa Ness, and Jonathan Larson. 2025. [From local to global: A graph rag approach to query-focused summarization](#). *Preprint*, arXiv:2404.16130.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. [Lightrag: Simple and fast retrieval-augmented generation](#). *Preprint*, arXiv:2410.05779.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. [Hipporag: Neurobiologically inspired long-term memory for large language models](#). *Preprint*, arXiv:2405.14831.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Online. Association for Computational Linguistics.

- Chaeun Kim, Jinu Lee, and Wonseok Hwang. 2025. [Legalsearchlm: Rethinking legal case retrieval as legal elements generation](#). *Preprint*, arXiv:2505.23832.
- Ryoma Kondo, Riona Matsuoka, Takahiro Yoshida, Kazuyuki Yamasawa, and Ryohei Hisano. 2025. [Capturing legal reasoning paths from facts to law in court judgments using knowledge graphs](#). *Preprint*, arXiv:2508.17340.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. [Sailer: Structure-aware pre-trained language model for legal case retrieval](#). *Preprint*, arXiv:2304.11370.
- Haitao Li, Yifan Chen, Yiran Hu, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025. [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, pages 3606–3615, New York, NY, USA. Association for Computing Machinery.
- Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023b. [Lecardv2: A large-scale chinese legal case retrieval dataset](#). *Preprint*, arXiv:2310.17609.
- Minghan Li, Tianrui Lv, Chao Zhang, and Guodong Zhou. 2026. [Glier: Generative legal inference and evidence ranking for legal case retrieval](#). *Preprint*, arXiv:2604.23779.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023c. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Zheng Liu, Chaofan Li, Shitao Xiao, Yingxia Shao, and Defu Lian. 2025. [Llama2vec: Unsupervised adaptation of large language models for dense retrieval](#). *Preprint*, arXiv:2312.15503.
- Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. [Finding the law: Enhancing statutory article retrieval via graph neural networks](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2761–2776, Dubrovnik, Croatia. Association for Computational Linguistics.
- Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. [Lecard: A legal case retrieval dataset for chinese law system](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2342–2348, New York, NY, USA. Association for Computing Machinery.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with bert](#). *Preprint*, arXiv:1901.04085.
- Shounak Paul, Dhananjay Ghumare, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2025. [Il-pcsr: Legal corpus for prior case and statute retrieval](#). *Preprint*, arXiv:2511.00268.
- Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. [Graph retrieval-augmented generation: A survey](#). *Preprint*, arXiv:2408.08921.
- Jay M. Ponte and W. Bruce Croft. 1998. [A language modeling approach to information retrieval](#). In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Hsuan-Lei Shao, Yi-Chia Chen, and Sieh-Chuen Huang. 2020. [Bert-based ensemble model for statute law retrieval and legal information entailment](#). In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers*, page 226–239, Berlin, Heidelberg. Springer-Verlag.
- Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. [STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10658–10671, Miami, Florida, USA. Association for Computational Linguistics.
- Yanran Tang, Ruihong Qiu, Xue Li, and Zi Huang. 2025. [Reakase-8b: Legal case retrieval via knowledge and reasoning representations with llms](#). *Preprint*, arXiv:2510.26178.
- Vishnuprabha V, Daleesha M Viswanathan, Rajesh R, and Aneesh V Pillai. 2025. [Augmented question-guided retrieval \(aqgr\) of indian case law with llm, rag, and structured summaries](#). *Preprint*, arXiv:2508.04710.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Chengxiang Zhai and John Lafferty. 2004. [A study of smoothing methods for language models applied to ad hoc information retrieval](#). *ACM Transactions on Information Systems*, 22(2):179–214.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, Yi Chang, and Xiao Huang. 2025. [A survey of graph retrieval-augmented generation for customized large language models](#). *Preprint*, arXiv:2501.13958.

Siying Zhou, Yiquan Wu, Hui Chen, Xueyu Hu, Kun Kuang, Adam Jatowt, Chunyan Zheng, and Fei Wu. 2025. [ClaimGen-CN: A large-scale Chinese dataset for legal claim generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12296–12323, Suzhou, China. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, 50(1):237–291.

A LLM Use and Prompts

LLM use in experiments. We use LLMs at several controlled points in the experimental pipeline. During graph construction, LLMs are used to normalize heterogeneous criminal and civil case records into a unified case representation. For criminal cases, the model compresses existing structured fields such as facts, court reasoning, results, charges, and cited statutes into graph-ready records. For civil cases, where the source data is mainly claim-centered and often lacks explicit statute citations or court reasoning, the model is used as a conservative annotator to infer rationale elements and candidate statute references from the available claim information. Civil candidate statute links are further checked by an LLM-based verifier using the label- and confidence-based acceptance rules described in Section 4.3.

During retrieval, LLMs are not used to directly score or rank statutes. Instead, an LLM is used only to rewrite each consultation query into three to six legal-element or rationale-style tags for matching the rationale-level hierarchy. LLMs are also used offline to generate retrieval labels, keywords, and aliases for rationale-level leaf clusters, making the rationale-level graph searchable. The final statute ranking itself is produced by dense retrieval, graph expansion, and calibrated reciprocal-rank fusion, without an LLM reranker in the main result.

We also use LLMs in auxiliary diagnostics, including answer generation and LLM-as-judge eval-

uation. These diagnostics are used only to examine whether retrieval improvements can transfer to downstream answer quality; they are not used to construct FR-MEG, tune the main retrieval method, or compute the primary statute-retrieval metrics. Importantly, LexRAG evaluation queries and gold statute labels are not used to construct the graph or its case-derived evidence. FR-MEG is built only from external case datasets and the statute corpus, and A/B/C coverage labels are used solely for post-hoc analysis rather than inference-time routing.

Criminal case alignment prompt.

The formal preprocessing run used `alignment_prompt_name=prompt1`. For criminal cases, structured fields are compressed into graph-ready case facts, rationale elements, legal elements, and statute numbers.

```
System:
You are a rigorous Chinese legal information-structuring
model. Output only a JSON object that can be parsed
by json.loads. Do not output explanations or Markdown.

User:
You are processing criminal judgment samples. Output only
the minimal JSON needed for downstream post-
processing. Do not repeat fields that are already
obvious from the input, and do not output Markdown.

Output schema:
{
  "facts_summary": "string",
  "court_reasoning": "string",
  "judgment_result": "string",
  "statute_numbers": ["string or int"],
  "reasoning_steps": ["string"],
  "legal_elements": ["string"]
}

Rules:
1. Use only information that already appears in the input.
2. facts_summary must be based only on fact.
3. court_reasoning must be based only on reason.
4. judgment_result must be based only on result.
5. statute_numbers should keep only the article numbers
   already present in article; do not expand them into
   full statute names.
6. reasoning_steps should contain 3 to 5 items, and each
   item should express only one minimal judgment step.
7. legal_elements should contain 3 to 6 short phrases.
8. Output JSON only.

Input sample:
{case_json}
```

Civil case alignment prompt. For civil cases, the original records are claim-centered. The prompt asks the model to act as a cautious annotator and infer a conservative intermediate representation for graph construction.

```
System:
You are a cautious Chinese civil virtual judge. Output only
a JSON object that can be parsed by json.loads. Do
not output explanations or Markdown.
```

User:
You are a cautious civil virtual judge. The current input contains only the plaintiff's alleged facts, claims, and cause of action. It is not a judgment document and does not mean that the court has already found the facts.

Your task is not mechanical extraction. Instead, from the perspective of a hypothetical adjudicator, summarize the case, identify dispute issues, organize the review steps, and provide a conservative possible-support assessment based on the plaintiff's allegations.

Output only the minimal JSON needed for downstream post-processing. Do not output empty court-reasoning fields, empty judgment-result fields, or empty statute fields. Do not output Markdown.

Output schema:

```
{
  "facts_summary": "string",
  "claims": ["string"],
  "reasoning_steps": ["string"],
  "dispute_focus": ["string"],
  "legal_elements": ["string"],
  "tentative_outcome": "string"
}
```

Rules:

1. facts_summary must be based only on plaintiffFactSegment. Write it as a case summary, and do not pretend that these are facts found by the court.
2. claims must be based only on claims. Light denoising and deduplication are allowed.
3. reasoning_steps should contain 3 to 5 items and must reflect the review order of a virtual judge, rather than simply repeating the claims.
4. dispute_focus should contain at least 2 items, describing the core dispute issues most likely to affect the scope of support.
5. legal_elements should contain at least 3 items, preferably written as legal-element phrases that a judge would examine.
6. tentative_outcome must use conservative wording, such as "if ... is established, then it may generally ..." or "based on the current information, it tends to ...", but still requires evidentiary support."
7. Do not write tentative_outcome as an actual judgment.
8. Output JSON only.

Input sample:

```
{case_json}
```

Query-to-rationale prompt. This prompt rewrites a consultation query into several abstract legal rationale tags for matching the rationale-level hierarchy.

System:
You are a query rewriting model for the FR-MEG rationale hierarchy. Your goal is to rewrite the user's question into several concise texts suitable for matching reasoning labels, keywords, and common phrasings. Output only valid JSON.

User:
Rewrite the following query into 3 to 6 rationale-hierarchy retrieval tags.

Requirements:

1. Each tag_query should resemble a knowledge-graph node label, not a complete consultation question.
2. aliases should include 2 to 3 synonymous or commonly used phrasings to help match different surface labels.
3. keywords should include core legal keywords, and should not include irrelevant case details.
4. Do not output statute numbers, and do not directly answer the query.

Output JSON schema:

```
{
  "elements": [
    {
      "tag_query": "retrieval tag",
      "aliases": [
        "alternate phrasing"
      ],
      "keywords": [
        "keyword"
      ]
    }
  ]
}
```

query:

```
{query}
```

Rationale-level leaf label prompt. This prompt labels rationale leaf clusters so that each cluster can be used as a retrieval target in the rationale-level graph.

System:
You are a Chinese legal knowledge graph and legal NLP researcher. Your task is to generate a high-quality legal retrieval label for a cluster of reasoning texts. The label should be useful for legal retrieval and should not be merely a generic crime name or cause of action. Output only one valid JSON object, and do not output Markdown.

User:
Generate a positioning label and a retrieval label for the following reasoning cluster. The parent_label may be broad, while cluster_label and child_tags should precisely match the leaf node. The output should support both upward merging and downward retrieval from the leaf node.

Requirements:

1. parent_label is the upper-level legal issue. cluster_label is the precise reasoning label of the current leaf cluster.
2. tag_text should be directly usable for embedding. It should include parent_label, cluster_label, core keywords, and common query phrasings.
3. If the leaf cluster is not sufficiently atomic, set should_split=true, and write the next-level sub-labels in child_tags.
4. If should_split=false, child_tags may be empty or []. Only fill child_tags when should_split=true.
5. child_tags and split_candidates must be short labels that can be used for clustering. Do not write full sentences.
6. Output valid JSON, and keep fields compact.

Output JSON schema:

```
{
  "cluster_label": "8-18 Chinese-character concise label",
  "element_type": "legal_relationship|duty|breach|damage|causation|liability|evidence|sentencing|procedure|remedy|other",
  "parent_label": "upper-level legal issue",
  "tag_text": "retrieval text, including upper-level category, precise label, core keywords, and common query phrasings, within 80 Chinese characters",
  "summary": "one sentence explaining the role of this leaf cluster in adjudicative reasoning, within 45 Chinese characters",
  "keywords": ["3-8 Chinese keywords"],
  "query_aliases": ["1-4 user query phrasings"],
  "should_split": false,
  "child_tags": [],
  "split_candidates": [],
  "confidence": 0.0
}
```

Cluster data:

```
{cluster_json}
```

Auxiliary answer generation prompt. The following prompt was used only for auxiliary generation diagnostics, not for the main statute retrieval result.

You are a Chinese legal consultation assistant. Please answer the user's current question based on the given dialogue and candidate statutes.

Requirements:

1. Answer only based on the provided statutes and evidence. Do not fabricate statutes that are not provided.
2. If the evidence is insufficient, explicitly state the uncertain parts.
3. The answer should first give a conclusion, and then briefly explain the legal basis.
4. If you use a statute, cite it with labels such as [S1], [S2].
5. Do not output the retrieval process or model information.

Dialogue history:
{dialogue_context}

Current question:
{current_query}

Candidate statutes:
{candidate_statutes_block}

Optional reasoning-pattern evidence:
The following content comes from legal applicability reasoning patterns in similar cases. It is used to help understand the applicability conditions of statutes, but it cannot replace the statutory text.
{reasoning_evidence_block}

Please generate the answer:

- legal advice.
7. clarity: whether the answer is clear, readable, and suitable for an ordinary consultation user.
 8. overall: overall quality.

Dialogue history:
{dialogue_context}

Current question:
{current_query}

Reference answer:
{reference_answer}

Gold statutes:
- {gold_title}
 {gold_text}

Candidate answer A:
{answer_a}

Candidate answer B:
{answer_b}

Please output JSON:

```
{
  "winner": "A|B|tie",
  "confidence": 0.0,
  "scores": {
    "A": {"legal_correctness": 1, "factual_alignment": 1, "
      legal_reasoning": 1, "grounding": 1, "completeness":
      1, "safety_calibration": 1, "clarity": 1, "overall":
      1},
    "B": {"legal_correctness": 1, "factual_alignment": 1, "
      legal_reasoning": 1, "grounding": 1, "completeness":
      1, "safety_calibration": 1, "clarity": 1, "overall":
      1}
  },
  "error_flags": {"A_hallucinated_statute": false, "
    B_hallucinated_statute": false, "A_unsupported_claim":
    false, "B_unsupported_claim": false},
  "brief_reason": "One sentence explaining the reason for
    the decision."
}
```

Auxiliary LLM-as-judge prompt. The following prompt was used to compare generated answers anonymously with two judge models. It was used only for auxiliary diagnostics.

You are a strict evaluator of Chinese legal answers. Please evaluate the quality of candidate answers to a user's legal consultation question.

You will see:

1. A multi-turn consultation dialogue and the current question.
2. A dataset reference answer.
3. Gold relevant statutes and their contents.
4. Two anonymous candidate answers, A and B.

Please note:

- Do not prefer an answer simply because it is longer.
- Do not penalize an answer merely because its wording differs from the reference answer. Evaluate only whether the legal substance is correct.
- If a candidate answer cites statutes that were not provided, misinterprets statutes, or ignores key facts, it should be penalized.
- If the two answers are close in quality, you may judge them as tie.
- You do not know which system produced A or B. Do not guess.
- Output JSON only, and do not output Markdown.

Scoring dimensions, each from 1 to 5:

1. legal_correctness: whether the legal conclusion and applicable statutes are correct.
2. factual_alignment: whether the answer accurately responds to the facts in the dialogue and the current question.
3. legal_reasoning: whether the answer explains the key legal elements, applicability conditions, and reasoning process.
4. grounding: whether the answer is faithful to the provided statutes and evidence, without unsupported extension.
5. completeness: whether the answer covers the main dispute issues in the user's question.
6. safety_calibration: whether the answer avoids over-certainty, misleading promises, or inappropriate