JustifIT-QA: Evidence-Grounded Legal Question Answering in Italian

Anonymous ACL submission

Abstract

Grounded legal question answering requires retrieving the right statutory provisions, generating legally sound answers, and supporting them with explicit evidence. We introduce *JustifIT-QA*, an Italian statutory QA dataset with expert-written answers and fine-grained evidence spans extracted from legislative text. We also present *DARv2*, an improved dense retriever previous statutory article retrievers. On top of this retrieval layer, we define a judge-based feedback loop in which a large language model critiques and refines generated answers, and we evaluate the feasibility of using an LLM as a judge for legal answers. Results show that feedback improves expert-rated answer quality when relevant context is retrieved, while evidence extraction quality improves with model scale. Overall, the paper introduces a new resource and an evidence-grounded pipeline for Italian legal QA.

1 Introduction

Legal question answering requires replies that are not only fluent but also normatively grounded, since even minor inaccuracies may lead to incorrect interpretations of obligations, rights, or sanctions. This is particularly challenging for LLM-based systems: recent work shows that they may produce legally unsupported statements and fabricated authorities, and that long-form generation can remain only partially grounded even when conditioned on retrieved sources (Dahl et al., 2024; Stolfo, 2024). At the same time, grounded legal QA remains under-resourced outside English, and correctness is difficult to assess at scale in open-ended settings (Guha et al., 2023). Recent work on LLM-based judging further suggests both the promise of automatic evaluation and the need to carefully analyze its reliability and biases (Gu et al., 2025; Chen et al., 2024).

This paper addresses these challenges in the Italian statutory domain. First, we introduce a dataset with expert-written answers and fine-grained evidence extracted from legislative text, enabling evaluation beyond article-level retrieval. Second, we present a dense retriever designed for the legal context, with consistent gains on recall and rank-aware metrics. Third, we study an iterative feedback loop in which generated answers are refined through evidence-grounded critiques, improving expert-rated correctness when relevant context is retrieved. Finally, we provide a feasibility study of LLM-based judging for legal answers, analyzing agreement with experts and safety-critical error patterns. To support future research, we will be made publicly available the dataset, retrieval model[1], and the code[2].

2 Related Work

Legal NLP has recently seen the release of several benchmarks for retrieval and question answering, most of which focus on identifying relevant provisions and supporting downstream legal decision making. Statutory retrieval datasets such as BSARD (Louis and Spanakis, 2022) provide expert-annotated mappings between natural-language questions and applicable legal articles, while broader benchmark suites such as LexGLUE (Chalkidis et al., 2022) cover a range of legal tasks but are less focused on grounded QA. JuriFindIT introduces statutory article retrieval for the Italian domain and explicitly encodes cross-article references, enabling graph-aware modeling of inter-provision dependencies in retrieval (Dalla Noce et al., 2026). More recent resources move toward answer generation grounded in legal sources. LLeQA targets long-form legal question answering with retrieval over legislative material and answer generation conditioned on retrieved passages (Louis et al., 2024). Complementary work in other legal genres emphasizes pinpointing legally salient

[1] justifit/DARv2 - Huggingface [2] JustifIT-QA repository (code & dataset)

spans, as in contract understanding datasets such as CUAD (Hendrycks et al., 2021). Beyond English, datasets such as JEC-QA for judicial exam questions in Chinese highlight retrieval and reasoning over legal materials under different legal systems and question styles (Zhong et al., 2020). Despite this progress, non-English statutory QA remains comparatively under-resourced, and span-level evidence annotations paired with expert-written answers are still uncommon across jurisdictions and languages.

A second relevant line of work concerns retrieval-augmented generation, where generation is explicitly conditioned on retrieved documents to reduce reliance on parametric memory (Lewis et al., 2020). In parallel, a rapidly expanding body of research studies the use of large language models as evaluators of open-ended outputs. LLM-as-a-judge methods investigate how rubric-based prompting can approximate human preference judgments and support scalable evaluation (Zheng et al., 2023; Liu et al., 2023). Iterative self-refinement methods further explore how natural-language critique can be used to revise and improve model outputs across multiple passes (Madaan et al., 2023; Shinn et al., 2023). Finally, evidence identification and rationale extraction have been widely studied to make predictions inspectable and to ground outputs in supporting text, notably in fact verification and rationale benchmarks such as FEVER and ERASER (Thorne et al., 2018; DeYoung et al., 2020). These directions collectively motivate increasing attention to grounded generation, structured evaluation, and evidence-linked explanations in high-stakes domains such as law.

3 JustifIT-QA: Extending the Dataset with Answers and Fine-Grained Evidence

We introduce *JustifIT-QA*, an Italian statutory QA dataset built on top of JuriFindIT[3]. The dataset is designed to address two main limitations of that benchmark: the absence of expert-written answers and the lack of fine-grained evidence annotations. For each question, JustifIT-QA provides (i) an expert drafted answer and (ii) explicit annotations of the relevant text spans within the labelled articles drawn from the JuriFindIT corpus. This yields a stand-alone resource for grounded legal question answering, and evidence-based interpretability.

[3] jurifindit/JuriFindIT

The extension is built on the expert-authored portion of the original dataset and does not introduce new questions. Instead, it increases the annotation granularity of the existing 895 queries by adding answers and span-level evidence over previously labelled relevant articles, preserving continuity with the original benchmark while enabling supervision beyond retrieval.

3.1 Expert-Written Answers

For each expert-authored question, we collected a reference answer written by a legal annotation team covering the four domains represented in the dataset: civil law, criminal law, anti-money laundering and counter-terrorism regulations, and privacy. Using a team with matching domain expertise helps maintain consistency across question interpretation, article selection, and answer drafting. The answers were written from scratch by human experts, without generation or post-editing from model outputs. Each answer is designed to resolve the question directly, identify the applicable rule, and mention the relevant conditions or consequences when required. We report some examples in Appendix A.3.

3.2 Fine-Grained Evidence Annotation

In addition to writing answers, annotators marked the specific portions of the relevant articles that support them. Unlike the original dataset, which provides article-level supervision, JustifIT-QA introduces span-level evidence capturing the normative statements underlying the expert answer.

Evidence was annotated at paragraph, subparagraph, clause, or shorter span level when necessary. Annotators were instructed to mark only the text directly needed to justify the answer, rather than entire articles. To reduce anchoring effects, annotation followed a two-step workflow: experts first drafted the answer, and only afterwards identified the minimal supporting excerpts in the relevant articles. This backward procedure encourages evidence selection based on justification rather than topical overlap. The resulting annotations define an explicit many-to-many mapping between questions, answers, and statutory text: a question may require multiple articles, and each article may contribute multiple disjoint excerpts. We report some samples from the dataset in Table 3.

4 Dataset Statistics

We report descriptive statistics for the two new annotation layers in JustifIT-QA, i.e. expert-written

Area	mean	std	median	max
AML/CTF	24.85	52.83	5.5	360
Civil Code	1.07	0.31	1.0	4
Criminal Law	1.69	1.04	1.0	7
Privacy	32.88	33.37	22.0	141

Table 1: Number of highlighted evidence sections per question by normative area.

Area	mean	std	median	max
AML/CTF	3462.64	7248.08	760.0	41891
Civil Code	49.79	49.11	39.0	489
Criminal Law	156.01	277.23	76.5	2582
Privacy	4341.94	4461.66	2540.0	20748

Table 2: Total words in highlighted evidence sections per question by normative area.

answers and relevant evidence sections, stratified by normative area. For answers, we compute word-count statistics and extract linguistic profiling features with Profiling-UD (Brunato et al., 2020) (Appendix A.1). For evidence, we report both the number of annotated sections per question and their total length in words.

Evidence sections. Table 1 shows that evidence fragmentation differs substantially across domains. *Civil Code* and *Criminal Law* typically rely on a compact grounding regime, where a small number of sections is sufficient to justify the answer and the dispersion is limited. In contrast, *AML/CTF* and *Privacy* exhibit markedly more fragmented evidence, with a pronounced long tail: many instances require aggregating a large number of highlighted sections, suggesting multi-provision reasoning and the need to connect dispersed normative constraints.

Evidence length. Table 2 reports the total amount of highlighted text (in words) per question. The pattern mirrors the fragmentation results. *Civil Code* evidence is generally short and concentrated, consistent with the fact that answers can often be supported by localized statutory portions. *Criminal Law* evidence is typically longer than Civil Code but remains relatively bounded for most cases, reflecting a mix of concise legal grounding and occasional longer justifications. Conversely, *AML/CTF* and *Privacy* are characterized by substantially longer evidence, again with strong right-skew: for these domains, supporting an answer frequently requires collecting extended statutory passages and combining multiple sections.

Overall, the evidence statistics stratify the benchmark into two regimes. *Civil Code* and *Criminal Law* largely fall into a compact-evidence setting, where grounding can be achieved with limited and localized excerpts. *AML/CTF* and *Privacy* represent the most demanding conditions, characterized by fragmented and lengthy evidence with heavy-tailed distributions. This motivates reporting results by normative area, since model behavior may differ substantially between compact and long-context evidence settings.

5 Answer Refinement and Evidence Generation

Legal question answering requires outputs that are not only fluent but *normatively grounded*: even minor inaccuracies may trigger incorrect interpretations of obligations, rights, or sanctions. This makes hallucinations and unsupported details particularly harmful in professional legal settings, where LLM-generated content has been shown to produce legally inconsistent statements and fabricated references (Dahl et al., 2024; Huang et al., 2025). To address this risk, we adopt a pipeline (Figure 1) that ties generation to statutory sources and produces reviewable support. Given a question, we (i) retrieve a small set of relevant provisions, (ii) generate an answer conditioned on the retrieved context via retrieval-augmented generation (Lewis et al., 2020), (iii) iteratively refine the answer through evidence-based verification, and (iv) extract verbatim supporting excerpts from the same provisions, in line with long-form legal QA settings that combine retrieval with grounded answer generation (Louis et al., 2024). Together, these components aim to reduce unsupported statements while improving interpretability by linking the final answer to explicit textual evidence.

5.1 Article Retrieval

Given a legal question in natural language, the goal of statutory article retrieval is to return a small set of legislative provisions that are useful for answering the query. We formalize retrieval as a function

$$R : (q, C; k) \rightarrow F \tag{1}$$

that takes a query q and a corpus $C = \{a_1, \ldots, a_N\}$ of N articles, and outputs a ranked subset $F \subset C$ of size $k \ll N$. Let $s_\theta(q, a) \in \mathbb{R}$ be a parametric scoring function, and let $\mathrm{rank}_q(a)$ denote the position of article a when sorting the

question	relevant_document_spans	answer
A bank employee receives feedback from the FIU regarding a suspicious transaction report. Can they inform the customer or other parties involved?	{1739: [71–1236, 1700–3904, 6717–6954, . . .]}	No. The employee must not disclose the report to the customer or any involved parties.
What are the main duties of the Data Protection Officer (DPO) under the GDPR?	{10662: [58–1223, 1301–2702, 3404–4676, . . .]}	The DPO must advise the controller/processor, monitor GDPR compliance...
A company makes a wire transfer linked to a specific operation. Must it notify a public authority within a certain period after payment?	{13725: [2482–3079], 13726: [12–875], 14071: [97–1114, . . .]}	If the payment concerns an operation subject to reporting, the company must notify the competent authority...

Table 3: Examples from JustifIT-QA (translated into English). For each question, *relevant_document_spans* is a dictionary whose key is the identifier of a legal text within JuriFindIT corpus and whose value is a list of evidence spans, each represented by the start and end character offsets of the corresponding excerpt.

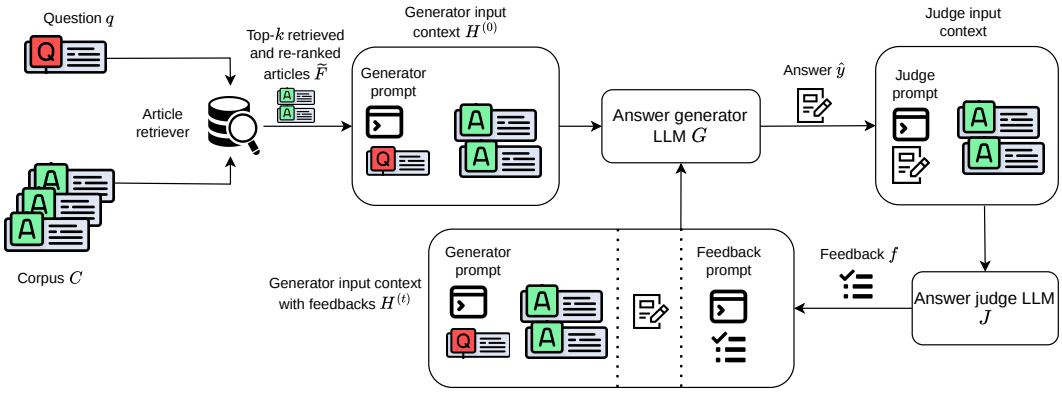

Figure 1: Overview of our pipeline. Given a question, an article retriever returns the top-k relevant provisions. A generator produces an answer conditioned on the retrieved articles. A judge model verifies the answer against the same articles and, when needed, produces feedback used to refine the next generation step. Prompt templates are shown in Appendix E.

.

corpus by decreasing score. The retrieved list is thus defined as:

$$F = R(q, C; k) = \{\, a \in C \mid \mathrm{rank}_q(a) \le k \,\} = \underset{a \in C}{\mathrm{TopK}}\ s_\theta(q, a). \tag{2}$$

We adopt a two-stage retrieval setup (Guo et al., 2022; Tao et al., 2023). In the first stage, we use a dense bi-encoder architecture (Reimers and Gurevych, 2019; Karpukhin et al., 2020) built on Transformer encoders (Vaswani et al., 2017), which maps queries and articles into a shared d-dimensional embedding space. A single encoder $\mathrm{Enc}_\theta(\cdot)$ is applied to both inputs, producing representations $\mathbf{h}_q = \mathrm{Enc}_\theta(q)$ and $\mathbf{h}_a = \mathrm{Enc}_\theta(a)$ with $\mathbf{h}_q, \mathbf{h}_a \in \mathbb{R}^d$. We compute each representation using the *last token* embedding. Relevance is estimated via cosine similarity, $s_\theta(q, a) = \cos(\mathbf{h}_q, \mathbf{h}_a) = \frac{\mathbf{h}_q \cdot \mathbf{h}_a}{\|\mathbf{h}_q\|\,\|\mathbf{h}_a\|}$, and all $a \in C$ are ranked according to $s_\theta(q, a)$.

In the second stage, a re-ranker refines the initial ranking by scoring each query–article pair jointly rather than comparing independently encoded representations. Given the candidate set returned by the bi-encoder, the re-ranker takes (q, a) as paired input and outputs a refined relevance score $\tilde{s}_\phi(q, a) \in \mathbb{R}$, which is used to reorder the candidates: $\widetilde{F} = \mathrm{RankK}_{a \in F}\ \tilde{s}_\phi(q, a)$, where $\tilde{s}_\phi(q, a)$ is the probability assigned by the re-ranker to the token *yes* for the paired input (q, a), i.e.,

$$\tilde{s}_\phi(q, a) = \frac{\exp(z_{\text{yes}}(q, a))}{\exp(z_{\text{yes}}(q, a)) + \exp(z_{\text{no}}(q, a))}, \tag{3}$$

with $z_{\text{yes}}(q, a)$ and $z_{\text{no}}(q, a)$ denoting the final logits associated with the tokens *yes* and *no*, respectively. For convenience, in the rest of this work we use the function R to indicate the retrieval and re-ranking pipeline.

Learning. We train the bi-encoder with a contrastive objective over a dataset $\mathcal{D} = \{(q_i, a_i^p)\}_{i=1}^N$ of question–positive-article pairs. Negatives are sampled via in-batch negatives (Karpukhin et al.,

2020): for a batch of size B, the negative set for q_k is $\mathcal{N}(q_k) = \{a_j \mid j \neq k,\ j \in [1, B]\}$. The loss is:

$$\mathcal{L}(q_i, a_i^p, \mathcal{N}(q_i)) = -\log \frac{\exp(s_\theta(q_i, a_i^p)/\tau)}{\sum_{a \in \{a_i^p\} \cup \mathcal{N}(q_i)} \exp(s_\theta(q_i, a)/\tau)}, \tag{4}$$

where τ is a temperature parameter.

5.2 Judge-Based Feedback Loop

To reduce hallucinations and enforce consistency with the retrieved statutory context, we introduce a feedback loop driven by a *judge* language model. LLM-based judges have recently emerged as a practical alternative to costly human evaluation, with evidence of useful correlation with expert preferences in controlled settings (Zheng et al., 2023; Liu et al., 2023). We formalize the judge as

$$J : (q, \widetilde{F}, \hat{y}) \to (\ell, f), \tag{5}$$

where ℓ is a compliance label and f is a feedback message. The judge evaluates $\hat{y}$ *only* against the retrieved articles $\widetilde{F}$, treating them as the authoritative source and thus directly targeting unsupported content, a central failure mode of neural generation (Maynez et al., 2020; Ji et al., 2023). Its decision follows expert-defined criteria covering support in $\widetilde{F}$, absence of contradictions, preservation of normative force and quantitative details, preservation of scope (e.g., subjects, conditions, exceptions, jurisdiction), and faithful integration without arbitrary additions; the prompt with the criteria is reported in Appendix E.2.

The label ℓ follows a three-level scale designed with a legal team: NON-COMPLIANT, PARTIALLY-COMPLIANT, and FULLY-COMPLIANT. This granularity is important in legal QA, where answers often include both core legal claims and ancillary details (e.g., minor conditions, procedural details, contextual remarks) enriching content of the answers. A binary label would conflate errors affecting the main conclusion with inaccuracies confined to secondary information. We therefore use NON-COMPLIANT for errors on core legal content, PARTIALLY-COMPLIANT for answers whose main claim is correct but include unsupported peripheral details, and FULLY-COMPLIANT for answers fully consistent with $\widetilde{F}$.

When $\ell \neq$ FULLY-COMPLIANT, the judge returns a feedback f identifying unsupported or contradicted statements in $\hat{y}$. Following iterative self-refinement approaches (Madaan et al., 2023), we implement revision in conversational form: the generator receives f as a new user turn while preserving the dialogue history. Let $\mathcal{H}^{(t)}$ denote the conversation state at iteration t; refinement is:

$$\hat{y}^{(t+1)} = G\big(\mathcal{H}^{(t)} \oplus f^{(t)}\big), \tag{6}$$

where $\oplus$ appends the new feedback to the history. The process stops when the judge returns FULLY-COMPLIANT or when a maximum number of iterations is reached. Algorithm 1 summarizes the procedure. The proposed loop implements evidence-grounded self-correction: the judge flags claims unsupported by $\widetilde{F}$, and the generator revises its answer while the statutory context remains fixed.

Algorithm 1 Conversational judge-based feedback loop for answer refinement.

Require: question q, corpus C, retrieval pipeline R, generator G, judge J, top-k, max steps T

$\widetilde{F} \leftarrow R(q, C; k)$
$\mathcal{H}^{(0)} \leftarrow \langle (q, \widetilde{F}) \rangle$
$\hat{y}^{(0)} \leftarrow G(\mathcal{H}^{(0)})$
$\mathcal{H}^{(0)} \leftarrow \mathcal{H}^{(0)} \oplus \hat{y}^{(0)}$
for $t = 0$ **to** $T - 1$ **do**
  $(\ell^{(t)}, f^{(t)}) \leftarrow J(q, \widetilde{F}, \hat{y}^{(t)})$
  if $\ell^{(t)} =$ FULLY-COMPLIANT **then**
    return $\hat{y}^{(t)}$
  end if
  $\mathcal{H}^{(t+1)} \leftarrow \mathcal{H}^{(t)} \oplus f^{(t)}$
  $\hat{y}^{(t+1)} \leftarrow G(\mathcal{H}^{(t+1)})$
  $\mathcal{H}^{(t+1)} \leftarrow \mathcal{H}^{(t+1)} \oplus \hat{y}^{(t+1)}$
end for
return $\hat{y}^{(T)}$

5.3 LLM-based Evidence Extraction

Beyond answer generation and verification, we explicitly extract textual *evidence* from the retrieved statutory context. This follows evidence-based QA paradigms (DeYoung et al., 2020; Thorne et al., 2018) and is particularly relevant in law, where verbatim, source-linked excerpts improve transparency and support post-hoc auditing.

Let q be a question, $C = \{a_1, \ldots, a_N\}$ the article corpus, and $\widetilde{F} = R(q, C; k)$ the retrieved set. Given the selected answer $y^\star$, we define an evidence extractor as

$$E : (q, \widetilde{F}, y^\star) \to \widehat{\mathcal{S}}, \tag{7}$$

where $\widehat{\mathcal{S}}$ is a set of predicted evidence spans. Each span is represented as $s = (id(a), \tau)$, where $a \in \widetilde{F}$ and τ is a verbatim excerpt from a; thus $\widehat{\mathcal{S}} = \{\hat{s}_1, \ldots, \hat{s}_m\}$. Gold evidence is defined analogously as $\mathcal{S} = \{s_1, \ldots, s_n\}$ with $s_i = (id(a_i), \tau_i)$, matching the dataset annotation protocol and enabling direct set-level comparison. Operationally,

E returns a structured list of $(id(a), \tau)$ pairs covering all excerpts needed to justify $y^\star$. This set-based formulation captures the many-to-many structure of legal justification, while restricting extraction to $\widetilde{F}$ aligns evidence with the same sources used for answering and limits unsupported citations. Combined with judge-based refinement, this yields answers that are both context-consistent and grounded, a requirement emphasized in long-form legal QA (Louis et al., 2024).

6 Experiments

We evaluate our framework by measuring its ability to generate legally compliant answers and accurately identify supporting evidence. The experiments aim to quantify the impact of iterative judge feedback on answer quality and to assess the precision of the extraction and retrieval components.

6.1 Answer Evaluation

Expert evaluation. To validate whether judge feedback improves answer quality in practice, we conduct an expert evaluation on 133 randomly sampled instances. Legal annotators assess generated answers using the same three-level compliance scale and criteria adopted by the judge (§ 5.2). For each instance, annotators are given the question, the generated answer, and the gold relevant articles, and are instructed to base their assessment solely on those articles. The sample includes a paired subset that enables direct comparison between pre- and post-feedback generations: for 48 questions, we collect both the initial answer generated without feedback and the judge-selected best answer after refinement. We use these labels both to characterize the overall compliance of judge-selected answers and to quantify the effect of feedback relative to the no-feedback baseline.

Automatic metrics. We measure reference-based similarity between generated and expert answers with METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2020). Pure n-gram overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are less suitable for long-form, open-ended answers (Wang et al., 2022; Xu et al., 2023). METEOR mitigates surface-form mismatch via stemming and synonym matching and has shown stronger correlation with human judgments than overlap-based measures (Sharma et al., 2017; Chen et al., 2022), while BERTScore captures semantic similarity through contextual embeddings and is therefore more robust to paraphrasing and variation in detail level.

6.2 Evidence Extraction Evaluation

Exact string matching between predicted and annotated evidence is brittle in legal texts, where harmless boundary differences or adjacent context may lead to mismatches. We therefore adopt a containment-based criterion computed *per sample* and then macro-averaged over the dataset. For the i-th question, let $\hat{T}_i = \{\hat{\tau}_{i,1}, \ldots, \hat{\tau}_{i,m_i}\}$ be the set of predicted excerpts and $T_i = \{\tau_{i,1}, \ldots, \tau_{i,n_i}\}$ the corresponding gold excerpts, pooled across retrieved articles. For a predicted excerpt $\hat{\tau} \in \hat{T}_i$ and a gold excerpt $\tau \in T_i$, we define:

$$\mathrm{GoldCont}(\hat{\tau}, \tau) = \frac{|Tok(\hat{\tau}) \cap Tok(\tau)|}{|Tok(\tau)|}, \qquad (8)$$

where $Tok(\cdot)$ denotes the set of normalized tokens (lowercasing and punctuation removal). A match is counted when $\mathrm{GoldCont}(\hat{\tau}, \tau) \geq 0.95$, with one-to-one matching to prevent multiple predictions from claiming the same gold excerpt. This yields per-sample TP_i, FP_i, and FN_i, from which we compute precision, recall, and F1, and finally macro-average over all samples.

6.3 Experimental Setup

Retriever and re-ranking. We fine-tune *jina-embeddings-v5-text*[4] (Akram et al., 2026), an encoder-only embedding model based on EuroBERT (Boizard et al., 2025), as the dense retriever used throughout the pipeline. The model has 239M parameters and supports inputs up to 8,192 tokens, making it suitable for long legislative texts. We train it on a mixture of real and synthetic questions: following prior work on statutory retrieval (Louis and Spanakis, 2022; Dalla Noce et al., 2026), we generate one synthetic question per article and combine these queries with the expert-authored questions as synthetic query generation has been shown to improve retrieval quality (Lawrie et al., 2025). Training uses a learning rate of 3.5×10^{-5} with linear decay, weight decay of 10^{-2}, batch size 768, and 35 epochs. At inference time, we first retrieve an initial candidate set and then re-rank the top-20 articles with *Qwen3-Reranker-4B* (Zhang et al., 2025); a more detailed description is provided in Appendix B.

Metrics. We evaluate retrieval with standard rank-aware metrics over the top-k list. Recall@k

[4] jinaai/jina-embeddings-v5-text-nano

Retriever	*Params*	*R@5*	*nDCG@5*	*MRR@5*	*MAP@5*
DAR	305M	76.02	74.19	77.20	69.85
DARv2	239M	**76.68**	**75.40**	**78.05**	**70.94**

Table 4: Retriever performance on the JuriFindIT validation set at k=5.

Setting	*Precision*	*Recall*	*F1*
Main correctness	87.16	84.31	85.71
Side correctness	79.84	72.79	76.15

Table 5: Judge (Qwen3-32B) performance against expert labels, treating correctness as the positive class.

measures the fraction of gold relevant articles retrieved in the top-k, MRR@k emphasizes how early the first relevant article appears, and nDCG@k and MAP@k reward systems that rank multiple relevant provisions higher in the list.

Generation, judging, and evidence extraction. We use *Qwen3-32B* (Team, 2025) [5] in *no-reasoning* mode for answer generation, and both *Qwen3-32B* and *Qwen3.5-27B* (Qwen Team, 2026) for compliance judging. The generator produces an answer conditioned on the question and retrieved articles, while the judge assigns a three-level compliance label and, when needed, provides feedback grounded in the same statutory context. Due to memory constraints, we cap the context to the top-5 retrieved articles and set the maximum number of feedback iterations to 2. For evidence extraction, we treat the task as structured prediction of verbatim supporting excerpts paired with article identifiers. To study scaling effects, we compare *Qwen3* 8B, 14B, and 32B under the same prompting setup, again using the top-5 retrieved articles as input context.

7 Results

7.1 Retriever performance

Table 4 reports retriever performance on the JuriFindIT validation set at k=5, to allow the comparison with the original work. Our trained retriever DARv2 improves over DAR (Dalla Noce et al., 2026) across all metrics. The gains are more pronounced on rank-aware measures than on Recall, indicating that DARv2 retrieves relevant provisions slightly more often within the top-5, also ranking them better. Summing the absolute improvements across the four reported metrics yields a cumulative gain of 3.81 points.

7.2 Judge-Based Evaluation and Feedback

Agreement with expert judgments. We first assess the judge as a proxy for expert evaluation by treating correctness as the positive class and reporting precision, recall, and F1 in Table 5; raw counts are provided in Appendix C. The judge is substantially more reliable on *main* content than on *side* information, with a better precision–recall trade-off on the core legal substance of the answer. In legal QA, precision is especially important because false positives correspond to answers that the judge approves despite being considered incorrect by experts, thereby increasing the risk of endorsing legally unreliable content. An error analysis suggests that retrieval recall is a key driver of expert-rated quality: answers judged incorrect by experts are associated with markedly lower average article recall than correct ones, both for main and side content, whereas false negatives exhibit recall levels comparable to true positives. This indicates that missing relevant context primarily harms answer quality, while many false negatives reflect conservative judge behaviour. Further details such as qualitative analysis are reported in Appendix D.

Judging with access to expert answers. We then evaluate the judge in a more informed setting, where it is additionally provided with the expert-written reference answer and the labelled articles together with the generated answer. Results are reported in Table 6; the corresponding counts are given in Appendix C. Access to the reference answer improves alignment with experts, but the models exhibit different operating profiles. Qwen3-32B provides the best overall balance, combining strong precision with very high recall, whereas Qwen3.5-27B in both thinking and non-thinking mode is more conservative: it attains higher precision, but at the cost of substantially lower recall, rejecting a larger fraction of answers that experts consider correct. In legal settings, this corresponds to a safety–coverage trade-off: a higher-precision judge is preferable when minimizing false acceptance is the primary concern, while a higher-recall judge is preferable when preserving valid answers is more important.

Effect of judge feedback. Finally, we examine whether judge feedback translates into better answers under expert assessment. Table 7 compares the proportion of expert-rated correct answers be-

[5] Qwen3/3.5 models.

Judge model	*Precision*	*Recall*	*F1*
Qwen3-32B	89.57	95.42	92.40
Qwen3.5-27B	93.89	80.39	86.62
Qwen3.5-27B (thinking)	94.57	79.74	86.52

Table 6: Judge performance on main correctness when provided with the expert-written reference answer and labelled articles.

Correct answers	*No feedback*	*With feedback*
Main correctness	76.74	83.72
Side correctness	69.77	74.42

Table 7: Percentage of expert-rated correct answers, comparing the initial answer (*No feedback*) and the judge-refined answer (*With feedback*).

Score	*Expert=0*	*Expert=1*
METEOR	25.34 ± 6.70	32.12 ± 11.34
BERTScore	66.80 ± 2.84	70.27 ± 4.11

Table 8: Similarity between generated answers and expert references, stratified by expert main correctness, reported as mean $\pm$ standard deviation.

Model	*Precision*	*Recall*	*F1*
Qwen3-8B	25.75	36.37	27.39
Qwen3-14B	34.70	46.64	36.26
Qwen3-32B	**41.07**	**53.08**	**42.55**

Table 9: Evidence extraction results on the JustifIT-QA validation set. Scores are macro-averaged across samples.

fore and after refinement. On the subset of questions with non-zero retrieval recall, feedback improves correctness on both main content and side details, indicating that judge-guided revision is effective when relevant statutory support is available. At the same time, the benefit is contingent on retrieval quality: when recall is zero, we found feedbacks counterproductive, as the model has no reliable evidence to ground the revision and producing unsupported changes. This confirms that the usefulness of iterative refinement depends on the availability of relevant context, and therefore on the quality of the retriever upstream.

7.3 Automatic Evaluations

We also compare generated answers against expert-written references using METEOR and BERTScore. Since expert references mainly capture the core legal content, we stratify the analysis by expert judgments of *main* correctness (Table 8). Both metrics are higher for answers judged correct by experts, but the separation is limited, especially for BERTScore. This suggests that reference-based similarity captures only partial alignment with expert judgments. These metrics should therefore be viewed as descriptive rather than correctness-based: incorrect answers may still score high, while correct ones may differ in wording or detail.

7.4 Evidence Extraction

Table 9 reports evidence extraction performance on the JustifIT-QA validation set under the setting introduced in § 6.2. We compare three models from the same family, varying only the number of parameters (Qwen3-8B/14B/32B). Performance improves monotonically with scale: Qwen3-32B achieves the best F1 and also the highest recall and precision, while Qwen3-14B attains intermediate results, and Qwen3-8B substantially lags behind. Overall, the gains are more pronounced in recall than in precision, suggesting that larger models are primarily better at covering expert-highlighted excerpts, while still producing a non-negligible amount of unmatched evidence. This trend indicates that model capacity is beneficial for near-verbatim legal evidence extraction, but the gap between recall and precision highlights the remaining difficulty of extracting only the strictly relevant portions under a high-containment constraint.

8 Conclusion

We introduced *JustifIT-QA*, an Italian statutory QA dataset with expert-written answers and fine-grained evidence spans, enabling evaluation beyond article-level retrieval. We also presented *DARv2*, which improves retrieval performance over the original JuriFindIT retriever, especially on rank-aware metrics. On top of this retrieval layer, we studied a judge-based feedback loop for answer refinement and showed that, when relevant context is available, it improves expert-rated correctness. We further found that LLM-based judging is a feasible proxy for evaluating core legal content, although performance drops on ancillary details, and that evidence extraction quality improves with model scale. In the end, our results highlight the importance of combining retrieval, verification, and source-linked evidence for grounded legal QA. We hope that this contribution will support future research on grounded and explainable legal QA.

Limitations

Our study has several limitations. First, due to memory constraints, all LLM-based components operate on at most the top-5 retrieved articles. While this budget keeps the pipeline tractable, it inevitably excludes relevant provisions for some questions, especially in domains such as AML/CTF and Privacy, where evidence is often distributed across multiple articles. A stronger retrieval stage could therefore further improve downstream generation and judging.

Second, the expert-written answers in JustifIT-QA are designed to be concise, precise, and directly responsive to the question. This makes them well suited for benchmarking grounded QA, but less representative of the longer and more nuanced answers often found in real legal practice, where lawyers may discuss alternative interpretations, procedural contingencies, or broader contextual implications. As a result, the dataset captures a focused notion of legal answer quality rather than the full spectrum of professional legal writing.

Third, the evaluation of answer quality remains partly constrained by the use of expert judgments on a sampled subset rather than on the full dataset. Although this is common in high-cost legal annotation settings, it limits the granularity with which we can analyse domain-specific failure modes. Relatedly, our analysis shows that reference-based automatic metrics such as METEOR and BERTScore do not reliably track legal correctness, which further increases dependence on expert assessment.

Fourth, the judge-based feedback loop is only as strong as the retrieved evidence it receives. When retrieval misses the relevant provisions, iterative refinement may amplify unsupported edits rather than correct them, as observed in the recall-zero cases. This makes the overall system highly dependent on the quality of the retrieval stage and highlights that feedback cannot compensate for missing evidence.

Finally, evidence extraction is evaluated with a containment-based matching criterion that is more robust than exact string matching, but still remains surface-oriented. It does not fully capture deeper equivalence relations such as paraphrastic justifications, alternative but legally valid excerpt boundaries, or implicit support distributed across multiple passages. More semantically informed evaluation of legal evidence extraction remains an open problem.

Ethical considerations

All questions, answers, and evidence annotations in JustifIT-QA were produced by legal experts, so the dataset does not contain personal or sensitive user data. Still, as with any resource built from legislative documents and annotation workflows, errors may arise from preprocessing, article segmentation, evidence extraction, or annotation inconsistencies. We therefore recommend caution when using the dataset for downstream training or evaluation, especially in settings where mistakes may propagate to end users.

More broadly, the pipeline studied in this work is intended to support research on grounded legal QA, not to replace professional legal advice. Although retrieval, judging, and evidence extraction improve transparency, generated answers may still omit relevant provisions or contain unsupported statements, particularly when retrieval fails. Responsible use is therefore essential, especially in high-stakes scenarios where errors could affect legal interpretation or decision making.

References

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. 2026. jina-embeddings-v5-text: Task-targeted embedding distillation. *Preprint*, arXiv:2602.15547.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André F. T. Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El Haddad, Manuel Faysse, Maxime Peyrard, Nuno Miguel Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *CoRR*, abs/2503.05500.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 7145–7151. European Language Resources Association.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4310–4330. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States. Association for Computational Linguistics.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *CoRR*, abs/2401.01301.

Niko Dalla Noce, Davide Colla, Sina Farhang Doust, Lorenzo De Mattei, and Davide Bacciu. 2026. Jurifindit: an italian legal retrieval dataset. In *Findings of the Association for Computational Linguistics: EACL 2026*. Association for Computational Linguistics. Accepted to appear.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

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

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans. Inf. Syst.*, 40(4):66:1–66:42.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: an expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

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 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Dawn J. Lawrie, Efsun Kayi, Eugene Yang, James Mayfield, Douglas W. Oard, and Scott Miller. 2025. Generate-distill: Training cross-language IR models with synthetically-generated data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 2926–2930. ACM.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in french. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6789–6803. Association for Computational Linguistics.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22266–22275. AAAI Press.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *Preprint*, arXiv:1706.09799.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alessandro Stolfo. 2024. Groundedness in retrieval-augmented long-form generation: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1537–1552, Mexico City, Mexico. Association for Computational Linguistics.

Chongyang Tao, Jiazhan Feng, Tao Shen, Chang Liu, Juntao Li, Xiubo Geng, and Daxin Jiang. 2023. CORE: cooperative training of retriever-reranker for effective dialogue response selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3102–3114. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling exemplification in long-form question answering via retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2079–2092, Seattle, United States. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9701–9708. AAAI Press.

A Dataset

A.1 Linguistic profiling of answers

To characterize the linguistic properties of the answers, we compute Profiling-UD features separately for the four normative areas (AML/CTF, Civil Code, Criminal Law, Privacy). Table 10 reports a concise set of indicators covering lexical complexity (chars/token, lexical density), POS distribution (%NOUN, %VERB), and syntactic complexity (dependency depth, dependency length, %subordinate clauses). The trends are consistent across measures: *Privacy* answers are syntactically more articulated, with longer sentences and stronger embedding, whereas *Criminal Law* displays the highest nominal share, reflecting a more nominalized style. *AML/CTF* and *Civil Code* show broadly similar POS profiles and overall syntactic patterns.

A.2 Linguistic profiling of real vs. synthetic questions

To characterize the linguistic gap between human-authored queries and our synthetic supervision, we report Profiling-UD statistics (Brunato et al., 2020) for (i) real, expert-annotated questions, (ii) JuriFindIT synthetic questions distributed with the dataset, and (iii) our new synthetic questions used in the new training runs. Table 11 summarizes the features (mean $\pm$ std).

Across features, the new synthetic questions are closer to real queries in length than the JuriFindIT synthetic set (higher tokens_per_sent), while remaining less subordinated (lower subordinate_proposition_dist and subordinate chain length). We hypothesize that this shift yields more retrieval-effective supervision for dense encoders: (i) slightly longer queries provide additional disambiguating context against near-duplicate statutory provisions, and (ii) a stronger nominal footprint (higher NOUN distribution) aligns better with the noun-heavy terminology that distinguishes legal articles (entities, institutions, and procedural objects). At the same time, reduced subordination may remove syntactic "noise" (incidental clauses)

Area	tok/sent	chars/tok	lex. dens.	%NOUN	%VERB	depth	dep. len	%subord.
AML/CTF	36.69	5.12	0.480	20.68	9.40	6.17	2.80	58.30
Civil Code	35.81	4.76	0.470	21.17	9.40	5.71	2.95	59.65
Criminal Law	44.06	4.91	0.480	24.04	8.14	6.43	2.95	62.60
Privacy	72.96	5.18	0.484	22.29	8.46	8.03	3.31	75.07

Table 10: Selected Profiling-UD statistics (means) computed on answer texts. *tok/sent*: tokens per sentence; *chars/tok*: characters per token; *lex. dens.*: content-word ratio; *depth*: average maximum dependency-tree depth; *dep. len*: average dependency link length (tokens); *%subord.*: subordinate proposition ratio.

Feature	Real (annot.)	JuriFindIT synth.	New synth.
tokens_per_sent	34.27 ± 12.00	24.82 ± 6.41	27.14 ± 7.05
char_per_tok	5.00 ± 0.51	5.03 ± 0.62	5.35 ± 0.73
lexical_density	0.50 ± 0.06	0.45 ± 0.06	0.47 ± 0.12
upos_dist_VERB	10.85 ± 3.99	6.66 ± 3.81	5.00 ± 4.07
upos_dist_NOUN	20.34 ± 4.60	22.94 ± 4.43	26.28 ± 9.49
avg_max_depth	5.45 ± 1.63	5.68 ± 1.75	5.44 ± 1.83
avg_links_len	2.92 ± 0.66	2.21 ± 0.41	3.02 ± 1.49
avg_subordinate_chain_len	1.12 ± 0.63	0.82 ± 0.61	0.60 ± 0.63
subordinate_proposition_dist	57.78 ± 25.45	40.81 ± 27.79	31.95 ± 32.70

Table 11: Profiling-UD features (mean $\pm$ std) for real and synthetic questions. "JuriFindIT synth." refers to the synthetic questions originally distributed with JuriFindIT; "New synth." refers to our newly generated synthetic questions.

that is weakly predictive of relevance, stabilizing query representations during contrastive training and improving separation from hard negatives.

A.3 Annotated Examples

This appendix provides two examples from the dataset, showing the question, the annotated relevant evidence span(s), and the corresponding expert-written answer.

Example 1. `Question:` If Anna gives Marco a watch to thank him for significant help, is this considered a donation?

`Relevant evidence:` *Article_id: 19778*, *"A gift made out of gratitude, in consideration of the recipient's merits, or as a special reward is also a donation."*

`Expert-written answer:` Yes. A gift made out of gratitude or in consideration of the recipient's merits is also considered a donation.

Example 2. `Question:` When the lease expires, Laura continues to live in the property and the landlord does not object. Can the contract be considered renewed?

`Relevant evidence:` *Article_id: 20758*, *"The lease is deemed renewed if, after its term has expired, the tenant remains and is allowed to retain possession of the leased property, or, in the case of a lease of indefinite duration, if notice has not been given in accordance with the previous article. The new lease is governed by the same conditions as the previous one, but its duration is that established for leases of indefinite duration."*

`Expert-written answer:` Yes. If, after the lease expires, the tenant remains in possession of the property and the landlord does not object, the lease is tacitly renewed under the same conditions as the previous one, but with indefinite duration, unless notice to vacate or termination had already been given.

B Re-ranking

To better isolate the contribution of reranking, we evaluate a two-stage pipeline where DARv2 first retrieves the Top-20 candidate articles for each query, and a reranker then reorders *only* these candidates. The DARv2@20 row in Table 12 reports the effectiveness of the original DARv2 ordering restricted to this Top-20 pool, while the remaining rows report the same metrics after reranking with different Qwen3 reranker variants. As this protocol keeps the candidate set fixed, improvements should be interpreted as gains in ranking quality conditional on candidate generation (i.e., a reranker cannot recover relevant articles that are not present in the Top-20 list). Overall, larger Qwen3 rerankers substantially improve early ranking quality over the DARv2@20 baseline: Qwen3-4B increases MAP from 0.766 to 0.834 and yields a strong boost in MRR@5 (0.785→0.892), indicating that a rel-

Model	MAP	MRR@5	nDCG@5
DARv2@20	76.64	78.54	78.68
Qwen3-Reranker-0.6B	70.55	74.59	73.95
Qwen3-Reranker-4B	83.40	89.17	84.81
Qwen3-Reranker-8B	84.57	89.26	85.61

Table 12: Pool-based reranking on the Top-20 candidates returned by DARv2. The DARv2@20 row evaluates the original DARv2 ranking restricted to the Top-20 pool; Qwen3 reranker rows report effectiveness after reordering the same candidate lists.

Support	*Expert=1*	*Expert=0*
Main eval.		
Judge=1	129	19
Judge=0	24	9
Side eval.		
Judge=1	99	25
Judge=0	37	20

Table 13: Confusion-matrix counts comparing the judge against expert labels for main and side correctness.

Support	*Expert=1*	*Expert=0*
Main eval. (Qwen3-32B)		
Judge=1	146	17
Judge=0	7	11
Main eval. (Qwen3.5-27B)		
Judge=1	123	8
Judge=0	30	20
Main eval. (Qwen3.5-27B, thinking)		
Judge=1	122	7
Judge=0	31	21

Table 14: Confusion-matrix counts for main correctness when the judge is provided with the expert-written reference answer and labelled articles.

evant article is more frequently promoted into the very top positions. Qwen3-8B provides the best overall reranking performance, further improving MAP to 0.846 and nDCG@5 to 0.856, suggesting a more reliable ordering among the highest-ranked candidates. In contrast, Qwen3-0.6B degrades performance across all metrics (e.g., MAP 0.705), highlighting that small rerankers may be insufficiently expressive for this legal-domain relevance assessment and can disrupt an already strong dense ordering.

C Confusion matrix evaluation

This appendix reports the confusion-matrix counts underlying the aggregate metrics discussed in § 7.2. While the main text focuses on precision, recall, and F1, the raw counts are useful to inspect the balance between false positives and false negatives, which have different implications in legal QA. In particular, false positives correspond to answers that the judge accepts although experts deem them incorrect, and are therefore the most safety-critical errors. False negatives, by contrast, reflect conservative rejections and mainly affect coverage.

Table 13 provides the raw counts for the comparison between judge and expert labels on main and side correctness. Consistent with the aggregate metrics in the main text, the judge is more reliable on main content than on side information: the number of true positives is higher and the false-positive rate is lower relative to the more heterogeneous side-content setting. This supports the interpretation that the judge can more reliably assess the core legal claim than ancillary details.

Table 14 reports the corresponding counts when the judge is additionally provided with the expert-written reference answer. These counts clarify the different operating profiles of the evaluated models. *Qwen3-32B* achieves stronger recall by recovering more expert-correct answers, whereas *Qwen3.5-27B* in standard mode is more conservative, with fewer false positives but substantially more false negatives. The thinking-mode variant of *Qwen3.5-27B* pushes this trade-off further: it reduces false positives even more, but at the cost of additional false negatives.

D Error Analysis on Model-Experts Agreement

We conducted a compact error analysis to better understand the disagreement patterns between the judge and expert annotators. The clearest signal is the role of retrieval recall. For *main* content, answers rated as incorrect by experts are associated with substantially lower average recall (approximately 0.5) than answers rated as correct (approximately 0.87). This suggests that, because the generator is explicitly constrained to rely on the retrieved statutory context, insufficient coverage of relevant articles directly harms answer quality. We observe the same trend for *side* content, where answers judged incorrect by experts have an average recall of about 0.55, compared with about 0.9 for correct answers.

A different pattern emerges for false negatives. Their average recall is comparable to that of true positives, both on main and side content. This indicates that these cases are less often due to miss-

ing evidence in the retrieved context, and more often to the judge being stricter than human annotators. In other words, low retrieval recall is primarily associated with genuinely poor answers, whereas false-negative judgments more typically reflect conservative model behavior rather than lack of supporting statutory material.

Overall, the analysis points to retrieval coverage as the main upstream factor affecting answer correctness, while judge-specific errors are more often attributable to thresholding or calibration effects.

Qualitative error patterns. A qualitative inspection of annotator comments reveals a small number of recurring failure modes. The most frequent one is *incomplete legal grounding*: many incorrect answers omit decisive provisions, paragraphs, or legally relevant conditions, leading to partial or unsupported conclusions. A second common pattern is *misinterpretation of the applicable norm*, where the model cites a relevant provision but applies it incorrectly or reverses its meaning. Annotators also frequently report *hallucinated or unnecessary extensions*, i.e., answers that go beyond the provided legal basis by introducing speculative roles, conditions, or consequences not supported by the retrieved articles. Relatedly, several errors stem from *wrong legal qualification*, such as focusing on a non-pertinent institution, sanction, or offense while missing the actually relevant legal basis. Finally, some answers are judged problematic because they are *too generic or evasive*: they restate parts of the retrieved text, but fail to resolve the question concretely. Overall, the comments suggest that answer quality is most compromised not by surface-level phrasing issues, but by missing or misapplied legal support and by overgeneration beyond the retrieved normative context.

E Prompt Templates

This appendix reports shortened English versions of the main prompts used in our pipeline. We omit repeated formatting instructions and minor implementation details, while preserving the core constraints that define each task.

E.1 Answer Generation Prompt

The generator is instructed to answer a legal question using only the retrieved statutory articles, with an emphasis on faithfulness, preservation of legal meaning, and explicit grounding in the provided documents.

You are an expert in Italian law. Your task is to answer, in Italian, a legal question Q using **only** the relevant articles D provided in the context. Your answer must be legally precise, grounded in D, and free of unsupported additions or distortions.

Follow these principles:

(1) Coverage and grounding. Include every material element in D that directly answers Q. Derive the answer only from what is explicitly stated or clearly implied by D. If the articles are insufficient, state this explicitly.

(2) No contradiction. Do not contradict the meaning, conditions, or consequences established in D. Do not mix unrelated provisions.

(3) Preserve legal force and precision. Maintain the same deontic force (e.g., "must" vs. "may") and preserve quantitative or temporal details exactly.

(4) Preserve scope. Do not alter who is bound, where the rule applies, when it applies, or the relevant exceptions and definitions.

(5) Faithful synthesis. You may summarize or rephrase for clarity, but do not change the legal meaning. Do not add external facts, precedents, or interpretations not present in D.

(6) Transparency. Explicitly mention the relevant legal facts and article identifiers when useful.

If information is missing or ambiguous, do not speculate. Instead, state that the answer cannot be fully determined from the provided articles.

Input: Question Q: `{question}`
Relevant articles D: `{documents}`

Output: Write a clear and legally precise answer in Italian that directly answers Q, remains strictly grounded in D, preserves legal meaning and scope, and explicitly refers to the relevant articles.

E.2 Judge Prompt for Feedback Generation

The judge evaluates whether a generated answer is faithful to the retrieved articles and, when necessary, provides targeted feedback for revision.

You are an expert legal verifier. Your task is to determine whether an answer R faithfully reflects the content and legal meaning of the relevant articles D.

You must identify which parts of R are unsupported, contradictory, or legally distorted with respect to D, and provide clear feedback so that the answer can be corrected.

Evaluate according to the following criteria:

(1) Coverage. Every relevant claim in R must be supported or clearly implied by D.

(2) No contradiction. R must not contradict any provision, condition, or rule in D.

(3) Normative force and precision. R must preserve deontic force and semantic polarity (e.g., "must" $\neq$ "may") and keep quantitative and temporal details unchanged.

(4) Scope. R must not alter subjects, jurisdiction, validity period, exceptions, definitions, or conditions stated in D.

(5) Faithful integration. Paraphrases or extensions are allowed only if they remain fully consistent with D and introduce no arbitrary information.

Assign one of the following labels:

- **COMPLIANT**: all criteria are satisfied;
- **PARTIALLY_COMPLIANT**: criteria (1)–(4) hold, but the answer includes non-essential unsupported additions;
- **NON_COMPLIANT**: one or more of criteria (1)–(4) are violated.

If the answer is not fully compliant, provide specific feedback indicating what must be corrected or removed. Avoid generic comments; each observation should be grounded in D.

Output (JSON only):

```
{ "evaluation": "COMPLIANT" | "PARTIALLY_COMPLIANT" | "NON_COMPLIANT",
"feedback": "<targeted correction instructions>"
}
```

E.3 Judge Prompt with Access to Expert Answers

In the reference-based setting, the judge compares the generated answer against the expert-written answer while consulting the legal documents.

You are a meticulous legal verifier. Your task is to determine whether a generated answer G is **COMPLIANT** with a reference answer L, while consulting the relevant documents D.

G is:

- **COMPLIANT** if it is consistent with L and may summarize or extend it without changing its legal meaning, normative force, or scope;
- **NON_COMPLIANT** if it contradicts, distorts, or substantially alters L, or adds incompatible legal content.

Use the following criteria: **(1) No contradiction.** G must not contradict any conclusion in L.
(2) No variation in normative force. G must preserve deontic force, polarity, and all numerical/temporal details.
(3) No variation in scope. G must not alter subjects, jurisdictions, validity conditions, exceptions, or definitions expressed in L.

General rules:

- Consider both L and G in light of the documents D.
- A claim in G that is absent from L may still be compliant if it is supported by D and compatible with L.
- Focus on substantive legal meaning, not stylistic or citation differences.

Input: Reference answer L: `{reference}`
Generated answer G: `{generated}`
Relevant documents D: `{documents}`

Output (JSON only):

```
{ "evaluation": "COMPLIANT" | "NON_COMPLIANT",
"reason": "<brief justification based on criteria (1)-(3)>",
"evidence": {
"contradictions": [...],
"incoherences": [...],
"supporting_facts": [...]
}
}
```