

Paper Type: Original Article

A Source-Grounded RAG Framework for Egyptian Criminal Law: Domain-Adapted Semantic Retrieval and Evidence-Grounded Generation

Aliaa Jamal ¹ , Aya Atef ¹ , Fawzya Said ¹ , Nouran Emad ¹ , Rania Ali ¹ , Mohammed Tarek ¹ 
and Shorok Osama ^{1,*} 

¹ Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt;
Emails: alliaagamall@gmail.com, ayabadrin667@gmail.com, sfawzya40@gmail.com, Nouraneldash@gmail.com,
raniaelnemer48@gmail.com, mohamedtarek24244@gmail.com, shorokosama@fci.zu.edu.eg.

Received: 10 May 2026

Revised: 20 Jul 2026

Accepted: 02 Oct 2026

Published: 07 Oct 2026

Abstract

Large language models can accelerate legal research, but legal deployment requires two properties that generic assistants do not guarantee: jurisdiction-specific retrieval and explicit grounding in authoritative text. This paper presents a retrieval-augmented generation (RAG) assistant specialized in Egyptian criminal law. The system constructs an Arabic legal knowledge base from scanned legal books and supplementary online legal materials, applies OCR cleaning and metadata-preserving chunking, and indexes dense representations for semantic retrieval. Its retrieval encoder is based on the Sentence-Transformers paraphrase-multilingual-mpnet-base-v2 model and is fine-tuned with corpus-derived anchor-positive-negative triplets using an InfoNCE objective. The generation pipeline is retrieval-gated: out-of-domain queries are rejected, retrieved passages are carried into the answer context with source metadata, and the system abstains when relevant legal evidence is not available. On the reported held-out in-domain evaluation, fine-tuning reduced validation loss from 0.51 to 0.04, increased MRR from 0.86 to 0.99, and increased ranking accuracy from 0.76 to 0.98. More importantly, the mean negative similarity decreased from 0.52 to 0.17, widening the positive-negative separation margin from 0.26 to 0.58. These results demonstrate improved legal-domain retrieval discrimination; they are not treated as a direct measure of end-to-end legal-answer correctness. The resulting architecture provides a reproducible foundation for source-cited Arabic legal research assistance while preserving a human-review boundary for high-stakes legal use.

Keywords: Egyptian Criminal Law, Retrieval-augmented Generation, Legal AI, Arabic NLP, Dense Retrieval, Contrastive Learning, Source Grounding, Large Language Models.

1 | Introduction

Legal research is a high-stakes information task: a useful answer must identify the controlling or relevant authority, preserve jurisdictional scope, and distinguish retrieved legal text from model-generated explanation. These requirements become especially important for Arabic legal systems, where general-purpose language models may have uneven coverage of local legislation and where legal materials are often distributed across



Corresponding Author: shorokosama@fci.zu.edu.eg



<https://doi.org/10.61356/j.scin.2026.3676>



Licensee SciNexuses. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

scanned books, government publications, and heterogeneous online sources. Retrieval-augmented generation (RAG) addresses part of this problem by conditioning generation on externally retrieved passages rather than relying exclusively on parametric model memory [1]. Dense retrieval further allows semantically related passages to be identified even when the wording of a user query differs from statutory terminology [2].

RAG, however, is not equivalent to legal reliability. Empirical studies of commercial legal research systems show that retrieval can reduce, but does not eliminate, unsupported or incorrect legal output [10]. Recent legal-RAG research therefore emphasizes retrieval quality, precise source selection, and post-retrieval verification rather than treating grounding as a binary property [7, 11]. For Arabic legal NLP, benchmark work such as ArabLegalEval and ALARB confirms both the growing importance of Arabic legal reasoning and the scarcity of jurisdiction-specific resources compared with English-language legal AI [8, 9].

The system studied in this paper addresses a narrower and measurable problem: building a source-grounded assistant for Egyptian criminal law around a domain-adapted Arabic retriever. The contribution is not the use of RAG by itself. Instead, the work integrates corpus construction from scanned Egyptian legal materials, supervised contrastive adaptation of a multilingual sentence encoder, retrieval gating with an abstention path, source metadata propagation, and a deployable multi-interface prototype.

1.1 | Research Questions and Contributions

The study is organized around three research questions:

- RQ1: Does domain-specific contrastive fine-tuning improve the retriever's ability to separate legally relevant passages from in-domain negatives?
- RQ2: How does fine-tuning change the geometry of the embedding space, particularly the separation between positive and negative legal passages?
- RQ3: How can a domain-adapted retriever be integrated with a retrieval-gated generation pipeline that exposes legal sources and abstains when supporting evidence is insufficient?

The resulting scientific and engineering contributions are:

- A corpus-processing pipeline for Arabic Egyptian criminal-law materials that combines OCR extraction, text normalization, chunk-level metadata, and semantic indexing.
- A domain-adapted dense retriever obtained by fine-tuning a multilingual sentence encoder on legal anchor-positive-negative triplets using a contrastive objective.
- An explicit evaluation of retrieval discrimination before and after fine-tuning, including loss, MRR, ranking accuracy, and positive-negative similarity separation.
- A retrieval-gated RAG architecture in which source retrieval, relevance checking, citation-carrying context, and abstention are first-class system components rather than post-hoc interface features.
- A deployment-oriented prototype linking the AI service to web, mobile, and messaging interfaces while retaining human review for legal interpretation and drafting.

2 | Related Work and Research Positioning

2.1 | Dense Retrieval and Retrieval-Augmented Generation

Dense Passage Retrieval demonstrated that dual-encoder representations can support scalable semantic retrieval by mapping queries and passages into a shared vector space [2]. Sentence-BERT established an efficient sentence-embedding formulation for semantic similarity and retrieval tasks [3]. RAG subsequently combined retrieval with sequence generation so that external documents could condition model outputs [1]. These ideas form the technical basis of the present system, but the legal domain adds requirements that generic open-domain question answering does not address: jurisdiction, article-level provenance, legal terminology, and the need to expose sources to a professional reviewer.

2.2 | Legal LLM Evaluation, RAG, and Hallucination Risk

LegalBench showed that legal reasoning is not a single capability and introduced a broad collection of expert-designed tasks for evaluating large language models [6]. LegalBench-RAG later focused specifically on the retrieval stage, emphasizing that precise retrieval of minimally sufficient legal text is a distinct problem from generation quality [7]. Independent evaluations of commercial legal research systems have also shown that RAG-based products can still produce erroneous or unsupported statements, reinforcing the need to evaluate retrieval and generation separately [10]. More recent production-oriented work adds reference-audit stages that check cited legislation and jurisprudence against authoritative sources before delivery [11].

2.3 | Arabic Legal NLP

Arabic legal NLP remains comparatively under-resourced. ArabLegalEval introduced a multitask benchmark for Arabic legal knowledge and reasoning, while ALARB extended evaluation toward multistep Arabic legal argumentation using court cases [8, 9]. These resources are important but target jurisdictions and tasks different from the Egyptian criminal-law retrieval problem considered here. The present work therefore concentrates on domain adaptation of the retrieval layer rather than claiming a general Arabic legal reasoning benchmark.

2.4 | Structural Comparison with Representative Work

Table 1 positions the proposed system relative to representative retrieval, legal-RAG, and Arabic legal-NLP work. The table is structural rather than a score-based ranking because the datasets, jurisdictions, and evaluation protocols are not directly comparable.

Table 1. Structural positioning of the proposed system against representative RAG and legal-NLP research.

Work	Primary focus	Language / jurisdiction	Retrieval adaptation	Source-grounded generation	Egyptian criminal-law scope
Lewis et al. [1]	General RAG	General / open-domain	Generic retriever	Yes, retrieved context	No
LegalBench [6]	Legal reasoning benchmark	Primarily English / multiple legal tasks	Not the main focus	No	No
LegalBench-RAG [7]	Legal retrieval benchmark	English legal corpora	Evaluates retrieval	Benchmark-oriented	No
ArabLegalEval [8]	Arabic legal knowledge benchmark	Arabic / Saudi legal sources	Not the main contribution	Evaluation-oriented	No
ALARB [9]	Arabic legal reasoning benchmark	Arabic / Saudi commercial cases	Regulation identification task	Evaluation-oriented	No
This work	Domain-adapted legal RAG	Arabic / Egyptian criminal law	Yes; contrastive fine-tuning	Yes; retrieval gate + source metadata	Yes

As Table 1 shows, the distinguishing feature of this work is the combination of an Egyptian criminal-law corpus, a domain-adapted Arabic dense retriever, and an operational source-gated generation pipeline. The novelty claim is therefore deliberately scoped to this combination rather than to RAG, legal question answering, or Arabic legal NLP individually.

3 | System Architecture and Trust Boundary

3.1 | Retrieval-Gated RAG Workflow

The end-to-end decision path is shown in Figure 1. A user query first passes through a domain classifier. Queries outside the supported legal scope do not enter the retrieval-and-generation path. For accepted queries, the system performs query enhancement, dense embedding, vector search, and a relevance check. Generation is permitted only when the retrieval stage returns context judged sufficiently relevant by the application logic. Otherwise, the assistant returns a no-information response instead of fabricating an answer.

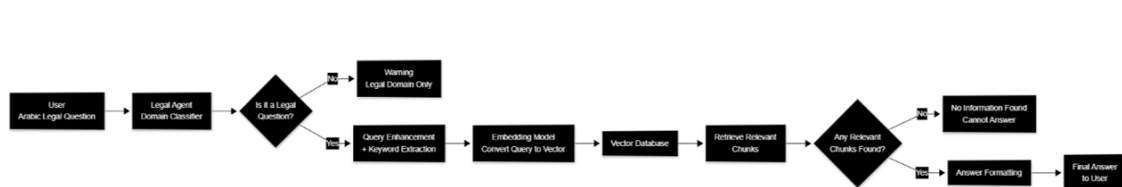


Figure 1. Retrieval-gated RAG workflow used by the Egyptian criminal-law assistant.

Figure 1 makes the system's trust boundary explicit. The LLM is not treated as an autonomous legal authority; it receives a bounded context assembled by the retrieval layer, and the user-facing answer is expected to retain the retrieved source identifiers. This architecture reduces one class of unsupported generation but does not, by itself, prove that every generated legal interpretation is correct. Accordingly, the prototype is positioned as research and professional decision support rather than autonomous legal advice.

3.2 | Application Architecture

Figure 2 shows the broader application architecture. Lawyers, law students, and other users can access the service from mobile, web, or Telegram interfaces. Requests pass through an API gateway and authentication service before entering the query processor. The Python AI service hosts the domain classifier, embedding model, vector search, and locally served LLM through Ollama. A response builder combines the generated answer with conversation context and source information, while SignalR supports real-time delivery. SQL Server stores application state and conversation history. This separation allows the retrieval and generation components to evolve independently from the client interfaces.

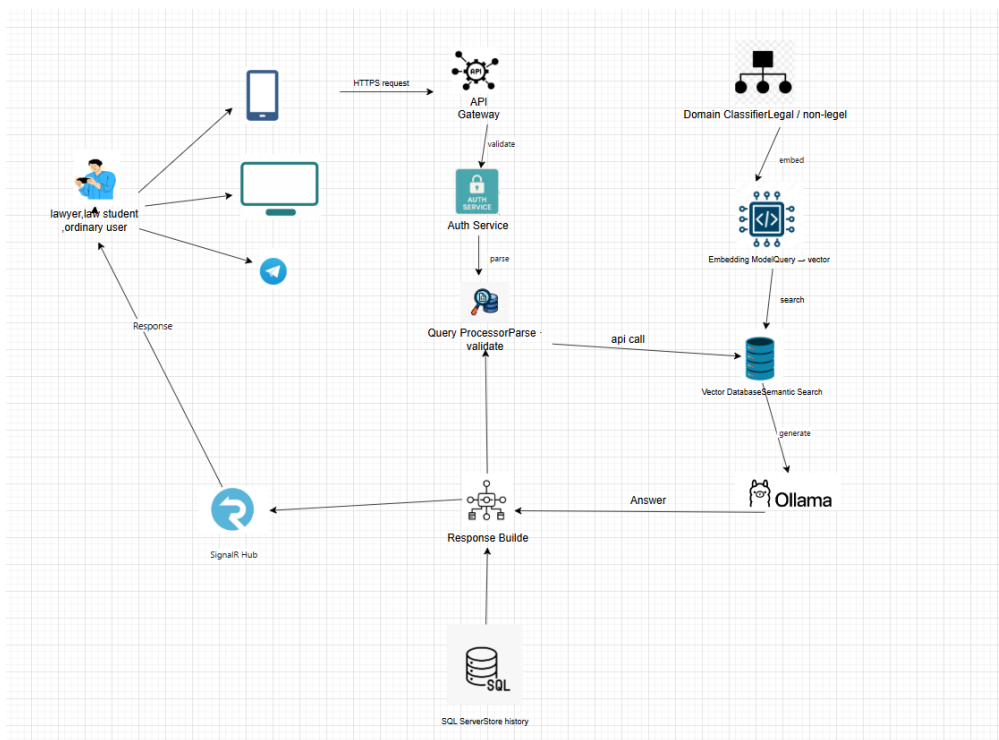


Figure 2. Multi-interface system architecture connecting authentication, query processing, retrieval, generation, history, and real-time response delivery.

The architecture in Figure 2 also separates the legal knowledge index from conversational state. This is important because legal source material and user case data have different provenance and retention requirements. JWT-based authentication and two-factor authentication are used at the account layer in the prototype. Generated pleadings and case analyses remain subject to professional review before use in legal proceedings.

3.3 | Implemented Functional Scope

The prototype supports legal question answering, source retrieval, structured case-issue analysis, and drafting assistance. Drafting functions can produce first-pass structures for lawsuits, memoranda, and pleadings, but the system does not autonomously file documents or make binding legal determinations. Potential weaknesses identified in a case are presented as issues for review rather than as guaranteed 'loopholes.' The web, mobile, and Telegram interfaces provide access to the same underlying retrieval-and-generation service.

4 | Corpus Construction and Retrieval Methodology

4.1 | Legal Source Collection and OCR Extraction

The knowledge base is restricted to Egyptian criminal-law materials. The source manuscript reports two source categories: scanned legal books obtained through the Faculty of Law at Zagazig University, including the Egyptian Penal Code, criminal procedure materials, and related legislation; and supplementary materials collected from legal websites and online legal portals. Because a substantial portion of the primary material existed as scanned pages, the pipeline converts PDF pages to high-resolution images with pdf2image and applies Arabic OCR using pytesseract.

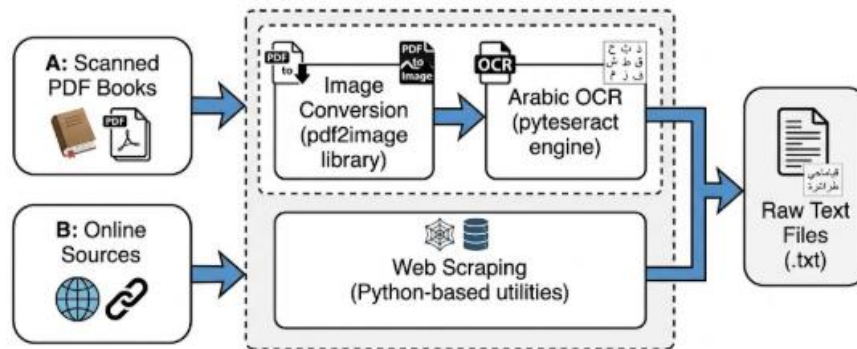


Figure 3. Corpus extraction pipeline for scanned books and supplementary online legal sources.

As illustrated in Figure 3, OCR output and web-extracted text are normalized into plain-text documents before indexing. Keeping the extraction stage separate from indexing makes it possible to inspect and correct OCR artifacts before they become embedded retrieval candidates.

4.2 | Text Cleaning, Chunking, and Metadata

Arabic OCR introduces characteristic errors, including malformed characters, inconsistent spacing, merged or split words, stray Latin symbols, and punctuation noise. The preprocessing stage therefore combines manual inspection with rule-based normalization. After cleaning, documents are segmented into semantically coherent chunks. Each chunk retains source-document metadata and article identifiers when available. This metadata is propagated through retrieval so that user-facing responses can point back to the underlying legal text rather than presenting untraceable prose.

4.3 | Construction of Contrastive Training Records

Fine-tuning requires supervised similarity examples rather than raw legal text alone. Each training record contains an anchor query or legal concept, a positive passage that expresses the relevant legal content, and a negative passage from the same domain. An automated n8n workflow coupled to a Claude-based structuring agent was used to convert cleaned legal text into JSONL triplets. Figure 4 shows the data-generation workflow. The use of in-domain negatives is important because legal passages can be lexically similar while having materially different legal meanings.

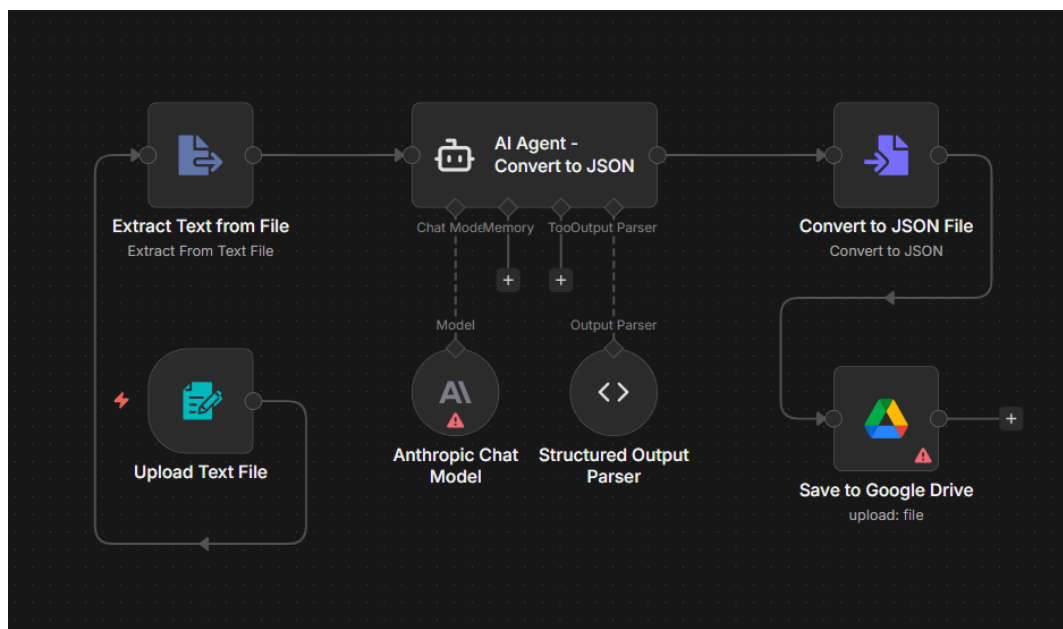


Figure 4. Automated construction of contrastive anchor-positive-negative records from cleaned legal text.

The automated generation process reduces annotation cost, but it also creates a validity boundary: retrieval metrics reflect the quality of these corpus-derived triplets. For that reason, the paper interprets the experiment as an in-domain retrieval study rather than an external legal benchmark.

4.4 | Base Embedding Model

The retriever starts from the Sentence-Transformers model paraphrase-multilingual-mpnet-base-v2, a multilingual sentence-embedding model that produces 768-dimensional vectors and supports semantic similarity across many languages [12]. The model was selected because it provides strong multilingual sentence representations while remaining small enough for task-specific fine-tuning. The legal adaptation is performed at the embedding level; the generator itself is not fine-tuned in the reported experiment.

4.5 | Mean Pooling, Normalization, and Similarity

Let h_t denote the final hidden state for token t and m_t the attention-mask value. Mean pooling followed by L2 normalization is defined as:

$$\mathbf{e}(x) = \frac{\sum_{t=1}^T m_t \mathbf{h}_t}{\sum_{t=1}^T m_t}, \quad \hat{\mathbf{e}}(x) = \frac{\mathbf{e}(x)}{\|\mathbf{e}(x)\|_2} \quad (1)$$

The normalized representations allow cosine similarity to be computed as an inner product:

$$\text{sim}(x, y) = \hat{\mathbf{e}}(x)^\top \hat{\mathbf{e}}(y)$$

This representation is used consistently for anchor queries, positive passages, negative passages, and indexed legal chunks. At inference time, the same encoder transforms the processed user query before vector search.

4.6 | Contrastive Fine-Tuning Objective

The training objective follows the contrastive-learning principle of increasing similarity between an anchor and its positive passage while decreasing similarity to one or more negatives [4, 5]. For anchor a_i , positive passage p_i , negatives n_{ij} , and temperature τ , the InfoNCE-style objective is:

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(a_i, p_i)/\tau)}{\exp(\text{sim}(a_i, p_i)/\tau) + \sum_{j=1}^{K_i} \exp(\text{sim}(a_i, n_{ij})/\tau)} \quad (2)$$

The temperature controls the sharpness of the similarity distribution. In this implementation $\tau=0.07$. The training loop was implemented in PyTorch with AdamW, cosine learning-rate decay, a 10% warm-up phase, mixed-precision training, and a fixed random seed of 8.

4.7 | Indexing, Query Reformulation, and Generation

After fine-tuning, legal chunks are encoded into 768-dimensional normalized vectors and stored in a vector database. At query time, the application performs an enhancement step that reformulates the request into a search-oriented representation; the example workflows include Arabic-to-English reformulation before embedding. Because the retriever is multilingual, the resulting query vector can still be matched against Arabic legal chunks. The independent contribution of translation-based reformulation is not isolated in the reported experiment, so the retrieval gains reported in Section 6 are attributed to the embedding fine-tuning rather than to query reformulation.

The retrieved chunks, original user question, and source metadata are then provided to the local generator. The answer formatter produces Modern Standard Arabic and includes the retrieved legal references. If the retrieval gate finds no sufficiently relevant context, generation is suppressed and the system returns an abstention response.

5 | Experimental Design

5.1 | Training Configuration

Table 2 summarizes the configuration used for the domain-adaptation run. These settings are reported explicitly to make the retrieval experiment reproducible at the model-training level.

Table 2. Fine-tuning configuration used for the legal embedding model.

Hyperparameter	Value	Rationale
Base encoder	paraphrase-multilingual-mpnet-base-v2	Multilingual sentence embeddings; 768-dimensional output
Maximum token length	128	Matches the implementation used in batched tokenization
Batch size	16	Stable mini-batch optimization within available GPU memory
Epochs	10	Observed convergence without divergence in validation loss
Learning rate	2×10^{-5}	Conservative adaptation of pretrained representations
Temperature τ	0.07	Sharpens contrastive similarity distribution
Pooling	Masked mean + L2 normalization	Sentence-level representation with cosine-compatible normalization
Optimizer	AdamW ($\beta_1=0.9$, $\beta_2=0.999$, $wd=0.01$)	Transformer fine-tuning with weight decay
Scheduler	Cosine decay + 10% warm-up	Stable early updates and gradual decay
Precision	fp16 with GradScaler	Reduced GPU memory and increased throughput
Random seed	8	Reproducibility across Python, NumPy, and PyTorch

The configuration in Table 2 is consistent with the stable loss trajectory observed across the ten epochs; the training and validation curves are analyzed in Section 6.2.

5.2 | Evaluation Metrics and Interpretation

The retrieval experiment uses validation loss, mean reciprocal rank (MRR), ranking accuracy, average positive similarity, and average negative similarity. MRR and ranking accuracy are defined as:

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i}, \quad \text{Acc}_{\text{rank}} = \frac{1}{N} \sum_{i=1}^N I\left(s_i^{\text{pos}} > \max_j s_{ij}^{\text{neg}}\right) \quad (3)$$

Here, r_i is the rank of the positive passage for validation item i , s_i^+ is the anchor-positive similarity, and s_{ij}^- denotes the anchor-negative similarity. The reported experiment is based on contrastive candidate sets derived from the constructed triplets. Consequently, ranking accuracy measures discrimination within the evaluated candidate set; it must not be read as corpus-wide top-1 legal retrieval accuracy.

To make embedding separation directly interpretable, the analysis also reports the mean similarity margin:

$$\Delta = \mu_{\text{pos}} - \mu_{\text{neg}} \quad (4)$$

A larger Δ indicates greater average separation between relevant and irrelevant legal passages. This measure is particularly informative here because fine-tuning slightly lowers absolute positive similarity while substantially lowering negative similarity.

5.3 | Scope of the Evaluation

The experiment evaluates the retrieval encoder, not the legal correctness of generated answers. No end-to-end factuality, citation-faithfulness, or lawyer-rated drafting score is inferred from the retrieval metrics. This separation follows current legal-RAG evaluation practice, where retrieval quality and generation reliability are treated as distinct layers [7, 10, 11]. The end-to-end examples in Section 7 are therefore illustrative system traces rather than an independent accuracy benchmark.

6 | Results and Analysis

6.1 | Retrieval Performance Before and After Fine-Tuning

Table 3 reports the evaluation summary from the source experiment. The final column adds the positive-negative similarity margin Δ , derived directly from the reported mean similarities.

Table 3. Retrieval evaluation before and after domain-specific fine-tuning.

Metric	Before fine-tuning	After fine-tuning	Interpretation
Validation loss	0.51	0.04	Substantially lower contrastive error
MRR	0.86	0.99	Positive passage ranked near the top of evaluated candidate sets
Ranking accuracy	0.76	0.98	Improved discrimination of positive over negative candidates
Average positive similarity	0.78	0.75	Absolute positive similarity changes little
Average negative similarity	0.52	0.17	Large reduction in similarity to irrelevant legal passages
Separation margin Δ	0.26	0.58	Mean positive-negative separation more than doubles

Table 3 shows that the main effect of fine-tuning is not a uniform increase in all cosine scores. Average positive similarity decreases modestly from 0.78 to 0.75, while average negative similarity falls much more sharply, from 0.52 to 0.17. Accordingly, the separation margin increases from 0.26 to 0.58. This is the most informative geometric result: the model learns to reject semantically nearby but legally irrelevant passages more strongly.

6.2 | Convergence Behavior

Figure 5 presents the reported training and validation losses across the ten fine-tuning epochs. Training loss falls rapidly during the first three epochs and remains close to zero thereafter. Validation loss decreases from approximately 0.161 in the first epoch to approximately 0.042 by the final epoch and then stabilizes, with no divergence between the training and validation trajectories.



Figure 5. Training and validation loss across ten epochs of domain-specific embedding fine-tuning.

The convergence pattern in Figure 5 supports the decision to stop after ten epochs. It also provides a consistency check against the aggregate value in Table 3: the final logged validation loss is approximately 0.042, which rounds to the reported 0.04.

6.3 | Similarity-Space Separation

Figure 6 compares the positive and negative similarity distributions before and after fine-tuning. Before adaptation, the distributions overlap substantially in the mid-similarity region. After adaptation, the negative distribution shifts toward low similarity values, whereas the positive distribution remains concentrated at high similarity.

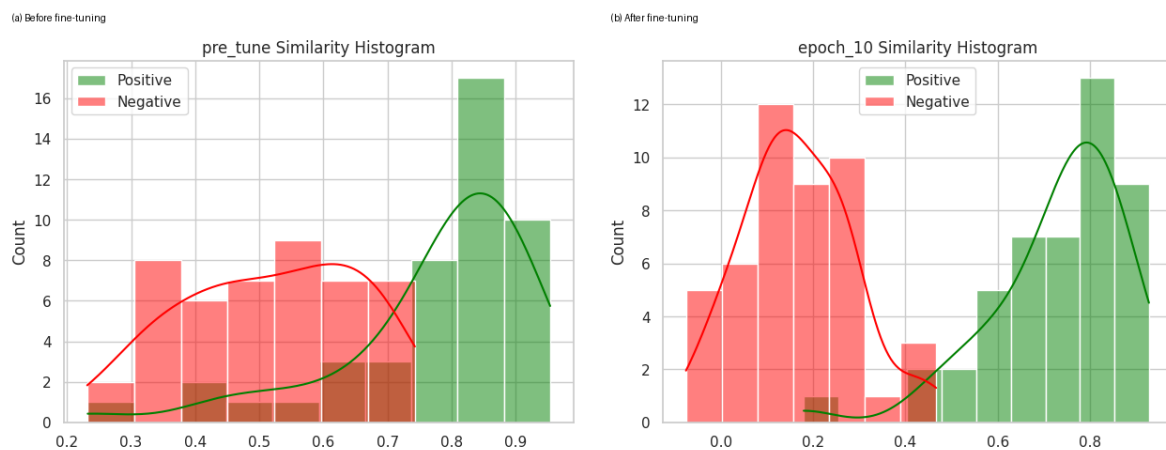


Figure 6. Positive and negative cosine-similarity distributions: (a) before fine-tuning and (b) after fine-tuning.

The distributions in Figure 6 visually confirm the numerical margin change in Table 3. This pattern is preferable to simply increasing all similarity scores because dense retrieval depends on relative ranking: a useful encoder must preserve relevant-query similarity while creating a decisive gap to hard negatives.

6.4 | What the 0.98 Ranking Accuracy Establishes

The 0.98 ranking accuracy is strong evidence that the fine-tuned encoder discriminates the evaluated in-domain positives from their negatives substantially better than the untuned model. It does not establish 98%

end-to-end legal-answer accuracy, nor does it guarantee that the correct authority will always be first in a full vector index containing many near-duplicate statutes, amendments, and commentary passages. That distinction is important for a legal system: retrieval metrics validate the retrieval layer, while answer correctness requires separate evaluation of citation fidelity, legal interpretation, temporal validity of legislation, and abstention behavior.

7 | End-to-End Behavior and Illustrative Legal Queries

The prototype was exercised on representative Arabic criminal-law questions to verify that the application can route a query through reformulation, retrieval, source assembly, and Arabic response generation. These examples demonstrate workflow behavior and citation handling; they are not used as substitutes for the quantitative retrieval evaluation.

7.1 Article-Oriented Query

A representative user asks:

ما هي أنواع الجرائم المذكورة في المادة 9؟

The system reformulates the query for retrieval, searches the legal index, retrieves the highest-scoring passages associated with Article 9, and passes those passages and their source metadata to the generator. The answer formatter then returns a structured Arabic explanation with the retrieved legal references. This example tests precise article-oriented retrieval rather than open-ended legal prediction.

7.2 Accomplice Liability Query

لو حد شارك في جريمة بس مش كان ناوي يحصل اللي حصل بالظبط، هل بيتعاقب؟

This query is intentionally colloquial and underspecified. The value of the pipeline lies in mapping the wording to relevant provisions on participation and liability before generation. The assistant can summarize the retrieved provisions, but the final legal characterization depends on the facts of the case and therefore remains a matter for professional legal review.

7.3 Multiple-Defendant Fine Query

لو حكموا على مجموعة متهمين بنفس الجريمة، الغرامة بتتقسم إزاي؟

The retrieval step searches provisions relevant to joint liability and the treatment of financial penalties. The generated response is structured around the retrieved authority and exposes the source to the user. This design lets the reviewer verify the cited text directly rather than accepting an uncited model statement.

8 | Prototype Implementation and User Interface

The research prototype exposes the same backend through multiple client surfaces. Figure 7 shows representative authentication, registration, mobile-assistant, and web-interface views. These interfaces are deliberately thin: legal retrieval and generation remain centralized in the AI service so that model updates do not require separate logic in each client.

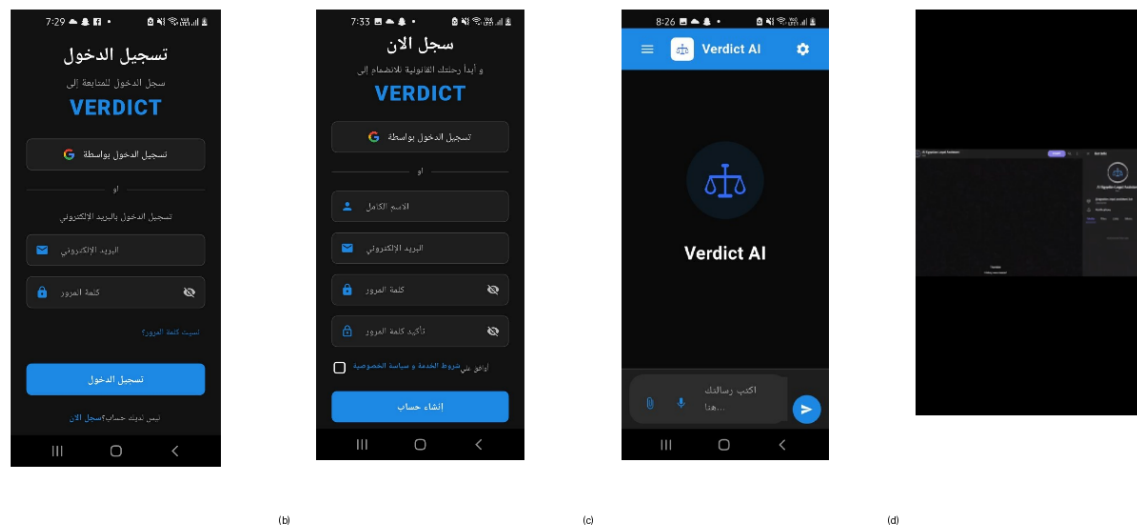


Figure 7. Representative Verdict AI prototype interfaces: (a) sign-in, (b) account registration, (c) mobile legal assistant, and (d) web interface.

Figure 7 demonstrates that the research pipeline was integrated beyond an offline notebook into a usable application. The prototype also includes subscription and notification infrastructure, but commercial features are not treated as scientific contributions in this paper.

9 | Reliability, Governance, and Validity Boundaries

9.1 Reliability Mechanisms

Reliability is implemented through layered controls rather than a claim of hallucination-free generation. First, domain classification restricts the supported scope. Second, the retrieval gate can block generation when no relevant context is found. Third, chunk metadata provides source provenance for the answer. Fourth, the application separates generated explanation from the underlying legal text so that a lawyer or informed user can verify the authority. These controls are consistent with evidence from legal-AI evaluations showing that retrieval reduces risk but does not remove the need for human verification [10, 11].

9.2 Professional Use Boundary

The assistant is designed for legal research support, education, and draft preparation. It is not an autonomous adjudicator and should not be used as a substitute for licensed legal advice in fact-sensitive matters. Generated pleadings, case analyses, and issue lists require human review for jurisdiction, amendments, procedural posture, evidentiary sufficiency, and the current legal status of cited provisions.

9.3 Experimental Validity

The reported metrics are internally consistent with an in-domain contrastive retrieval experiment and demonstrate a large improvement in embedding separation. Two boundaries are material when interpreting the findings. First, the validation examples originate from the same corpus-generation process as the training records; external generalization to unseen legal topics is not measured here. Second, the triplets were generated through an automated AI-assisted workflow rather than independent expert annotation. These boundaries do not invalidate the observed retrieval gain, but they limit the claim to domain-adapted retrieval discrimination on the reported corpus.

A stronger external evaluation would therefore pair the current retrieval study with expert-validated Egyptian legal queries, larger candidate pools, temporal tests across amended legislation, and end-to-end citation-faithfulness assessment. Those evaluations are methodologically distinct from the present encoder study and should be reported separately rather than inferred from embedding metrics.

10 | Conclusion

This paper presented a source-grounded RAG assistant specialized in Egyptian criminal law, with the retrieval layer treated as the central scientific component. The system converts scanned and online Arabic legal material into a metadata-bearing vector index, adapts a multilingual sentence encoder using contrastive fine-tuning, and integrates the resulting retriever with a gated generation workflow that can abstain when evidence is insufficient.

The experiment shows a substantial improvement in retrieval discrimination after fine-tuning: validation loss falls from 0.51 to 0.04, MRR rises from 0.86 to 0.99, ranking accuracy rises from 0.76 to 0.98, and the positive-negative similarity margin widens from 0.26 to 0.58. The distributional analysis shows that this improvement is driven primarily by pushing irrelevant legal passages away in the embedding space rather than artificially inflating positive similarity.

The resulting contribution is therefore a precisely scoped one: an implemented, domain-adapted retrieval backbone and source-gated application architecture for Arabic Egyptian criminal-law assistance. By separating retrieval quality from end-to-end legal correctness and retaining explicit human review, the system provides a technically grounded basis for further evaluation of Arabic legal RAG without overstating what the current experiment proves.

Author Contribution

All authors contributed equally to this study.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” in *Proc. EMNLP*, 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.
- [3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [4] P. Khosla et al., “Supervised Contrastive Learning,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [5] A. van den Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” *arXiv:1807.03748*, 2018.
- [6] N. Guha et al., “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 44123–44279.
- [7] N. Pipitone and G. H. Alami, “LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain,” *arXiv:2408.10343*, 2024.
- [8] F. Hijazi et al., “ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models,” in *Proc. 2nd ArabicNLP Conf.*, 2024, pp. 225–249, doi: 10.18653/v1/2024.arabicnlp-1.20.
- [9] H. Abu Shairah et al., “ALARB: An Arabic Legal Argument Reasoning Benchmark,” in *Proc. 3rd ArabicNLP Conf.*, 2025, pp. 389–406, doi: 10.18653/v1/2025.arabicnlp-main.32.
- [10] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,” *Journal of Empirical Legal Studies*, vol. 22, no. 2, pp. 216–242, 2025, doi: 10.1111/jels.12413.
- [11] A. Figueiredo et al., “Grounded in Law: A Multi-Stage Anti-Hallucination Pipeline for Legal RAG Systems in Brazilian Portuguese,” in *Proc. PROPOR 2026*, vol. 2, 2026, pp. 30–34.

-
- [12] Sentence-Transformers, “paraphrase-multilingual-mpnet-base-v2,” Hugging Face model card, accessed Sep. 2026. Available: <https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>.

Disclaimer/Publisher’s Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.