applied sciences

MDPI

Article

A Layout-Aware Hybrid Retrieval-Augmented Generation Framework for Mitigating Hallucination in Turkish Texts: Evidence from the Turkish Official Gazette

Samet Sarıyıldız * and Mustafa Yağcı

Department of Computer Engineering, Institute of Natural and Applied Sciences, Kırşehir Ahi Evran University, 40100 Kırşehir, Türkiye; mustafayagci@ahievran.edu.tr
* Correspondence: sariyildiz.samet@ogr.ahievran.edu.tr

Abstract

Large language models (LLMs) hallucinate, which limits their use in domains with little tolerance for error, such as law. We present a layout-aware hybrid retrieval-augmented generation (RAG) system for answering questions over 30 issues of the Turkish Official Gazette published in July 2025. The pipeline uses IBM Docling for hierarchical layout analysis and indexes content in PostgreSQL/pgvector with a local multilingual embedding model combined with a lexical index; response synthesis is delegated to claude-sonnet-5 under a citation-constrained prompt. On 33 questions written by the first author, the system answered 17 and declined 16 (coverage 51.5%). Among the answered questions, the hallucination rate was 4.2%, and citation accuracy was 81.8% (n = 11); Hit-Rate@8 was 0.786 on the 14 questions with labeled evidence. A zero-shot baseline gave a verifiable answer to one of 33 questions. In a retrieval ablation, replacing flat PDF text extraction with Docling raised Hit-Rate@8 from 0.21 to 0.79, whereas structural chunking and hybrid retrieval showed no consistent additional gain. A single legal expert rated 30 of 33 responses fully correct. Because the evaluation set is small, covers one month, and was rated by one expert, the results are exploratory and apply to this corpus.

Keywords: retrieval-augmented generation; large language models; hallucination mitigation; legal natural language processing; Turkish Official Gazette; layout-aware document parsing; hybrid retrieval

1. Introduction

The Transformer architecture [1] allowed large language models (LLMs) to move beyond frequency-based text statistics toward representations of long-range semantic dependencies, and models with tens or hundreds of billions of parameters now perform summarization, translation, and question answering at a level often indistinguishable from human writing [2,3]. This fluency is accompanied by hallucination: the generation of content that is well-formed but factually unsupported, either contradicting a supplied source (intrinsic) or introducing unverifiable claims absent from it (extrinsic) [3,4]. Because LLMs are trained to predict the most probable continuation rather than a verified truth, they favor memorized (and sometimes stale or incorrect) parametric knowledge over the evidence actually relevant to a query [3].

This failure mode is especially consequential in zero-error-tolerance domains such as law, where a fabricated statute number, misquoted clause, or invented precedent can translate directly into lost rights, financial penalties, or erosion of trust in judicial institutions [5].

Check for updates

Academic Editor: Michal Ptaszynski

Received: 24 September 2026
Revised: 19 October 2026
Accepted: 21 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Legal text further stresses general-purpose NLP through document length, dense archaic terminology, and deeply nested article–paragraph–subparagraph hierarchies [6]; when deployed directly, LLMs can fabricate entire precedents or statutory provisions in convincing formal legal language, a risk compounded by parametric memory that cannot reflect post-training legislative amendments [7].

Retrieval-augmented generation (RAG) addresses this by coupling a model's static memory to a dynamic, non-parametric knowledge base, so that generated claims are grounded in retrievable evidence rather than recalled from memory [7,8]. A canonical pipeline proceeds in three stages: source documents are split into retrievable chunks and embedded into a dense vector space; an incoming query is embedded into the same space and the top-k most similar chunks are retrieved by cosine similarity; and those chunks are supplied to the generator, which is instructed to answer using only that evidence [8]. This shift, from recalling what a model was trained on to reasoning over what it is shown, also means that updating a knowledge base (e.g., to reflect a new regulation) requires no retraining of the underlying model, which is a substantial advantage over scaling parameters alone as a strategy for keeping models current [7,9]. Plain dense-vector RAG, however, can still under-perform on richly structured documents: similarity search alone may ignore a legal text's hierarchical organization and pull in disjointed context, reintroducing partial hallucination even inside a nominally grounded pipeline [10]. This has motivated more autonomous designs such as Self-RAG, which learns when retrieval is needed and critiques retrieved evidence with dedicated tokens [11]; Interleaved-Retrieval Chain-of-Thought (IRCoT), which alternates multi-step reasoning with retrieval calls [12]; and layout-aware RAG, which extends vector search with document-layout information (headings, tables, reading order) to preserve hierarchical structure that plain chunking destroys [10].

Most advanced RAG research, however, targets English corpora and well-resourced computing environments. Legal NLP for Turkish, an agglutinative, morphologically rich language with dense, Ottoman-derived terminology, remains comparatively underdeveloped, and even basic subtasks such as named-entity recognition require dedicated morphological modeling (e.g., subword approaches such as Morph2Vec) [13,14]. The limited work on Turkish legal documents shows that authoritative sources such as Constitutional Court rulings, statutory articles, and, above all, the Official Gazette (Resmî Gazete) require bespoke annotation and dataset hierarchies even for machine-learning purposes, to the point that the Official Gazette has been formalized as its own named-entity category (OFF) in the literature [14]. Because Turkish statutes, presidential decrees, and appointments attain legal force only upon Official Gazette publication, the factual correctness of any Turkish legal AI assistant reduces, in practice, to its fidelity to the current, unaltered Official Gazette text. Without that fidelity, a genuine statutory reference cannot be distinguished from one fabricated from a model's parametric memory [3]. All LLMs risk this failure mode when relied upon without external verification, but the risk is sharper for Turkish: its agglutinative morphology and dense legal jargon make it easier for a generic, non-domain-adapted model to drift into semantically plausible but legally unfounded output than for a comparable model operating in a well-resourced language with mature legal NLP tooling. Recent studies have begun to evaluate language models on Turkish legal text [15,16] and to build RAG-supported legal reasoning systems for Turkish jurisdiction [17], so the problem is current and still open.

This gap defines our problem: most RAG approaches to hallucination mitigation are concentrated on English text, and the Turkish legal RAG work we found, such as a recent precedent-grounded framework for case-law question answering [17], does not address the structural hierarchy and multi-column layout of the Official Gazette. We address it with a Turkish-specific, layout-aware hybrid RAG architecture that (i) preserves

the article–paragraph–subparagraph hierarchy through structural parsing (IBM Docling, version 2.70.0), (ii) retrieves relevant context through hybrid dense-lexical indexing, and (iii) suppresses fabricated answers through a citation-constrained prompting protocol, yielding a question-answering assistant whose every claim is grounded in the Official Gazette and measurable in terms of faithfulness and citation accuracy. We did not find a prior study that combines layout-aware parsing of the Official Gazette with measured hallucination and citation accuracy for generated answers. The significance of this gap is not merely linguistic: because Turkish statutes, decrees, and appointments are legally binding only upon Official Gazette publication, a Turkish legal assistant is, in effect, only as trustworthy as its fidelity to that single, authoritative source, which makes hallucination mitigation in this specific setting a more tractable and more consequential target than hallucination mitigation over general Turkish text. This study's contribution is therefore not a new retrieval algorithm in the abstract, but the demonstration that combining an existing document-AI tool (Docling), a simple hybrid scoring rule, and a citation-constrained prompt is sufficient to bring hallucination in this setting into a practically useful range without fine-tuning either the embedding or the generation model on Turkish text.

This study asks three research questions: (RQ1) What effect does the proposed architecture have on hallucination relative to a zero-shot LLM baseline? (RQ2) What effect does layout-aware retrieval have on citation accuracy and grounding transparency, given the Official Gazette's structural complexity? (RQ3) Are the resulting answers reliable enough to support legal research, as assessed by automatic RAG metrics and an independent legal-expert review?

This study is scoped to legislation published in 30 Official Gazette issues (July 2025), evaluated on 33 original question-answer pairs spanning factual, analytical, and refusal (edge-case) scenarios. Retrieval and indexing ran locally (Apple M3 Pro; Apple Inc., Cupertino, CA, USA; MPS) to preserve data sovereignty, while response synthesis used the cloud-hosted claude-sonnet-5 (Anthropic, San Francisco, CA, USA). As detailed in the Limitations Section, the scale of the evaluation set, the single-period corpus, and the use of one independent reviewer limit the generalizability of the results, which should be read as indicative and as specific to this corpus; whether they carry over to other Turkish legal texts was not tested. The remainder of the paper reviews related work, details the system and evaluation protocol, reports results, discusses findings and limitations, and concludes with future directions.

2. Related Work

Hallucination, fluent but unsupported or fabricated generation, is attributed to the probabilistic nature of next-token prediction: when a model's knowledge boundary is reached, it favors the most probable continuation over a verified fact [3]. The literature distinguishes intrinsic hallucination, in which generated content directly contradicts a supplied source (e.g., a source stating that a ruling was upheld in 2019 while the model reports that it was overturned in 2021), from extrinsic hallucination, in which the model introduces unverifiable new content absent from the source altogether (e.g., citing a specific, entirely fabricated statute article to support a claim) [3]. Recent surveys add logical and stylistic sub-types, such as self-contradiction across reasoning steps and deviation from a requested format or register, and trace root causes to noisy pre-training data, inference-time exposure bias, and knowledge that is simply stale relative to a model's training cutoff [3]. In zero-error-tolerance domains, the danger is compounded because fabricated statutes or precedents are delivered in convincing, formally correct legal language that a non-expert reader cannot easily distinguish from a genuine citation.

RAG mitigates this by grounding generation in an external, verifiable, updatable knowledge base [7,8]. Beyond the canonical retrieve-then-generate pipeline, Self-RAG [11] lets a model decide dynamically whether to retrieve and critiques retrieved evidence with relevance/support tokens; IRCoT [12] interleaves chain-of-thought reasoning with retrieval for multi-hop legal inference; REALM/RETRO [9] pre-train generation with retrieval so that updating knowledge requires only re-indexing; and layout-aware RAG builds structure-preserving representations from page layout to avoid losing statutory hierarchy [10]. This study adopts that approach via IBM Docling because incorrect reading order and structural noise otherwise fragment legal context in official, table-dense, multi-column documents [10]. Encoding structure as Markdown helps only if the generator makes use of it; MDEval [18] benchmarks this Markdown awareness, which we did not measure here (see the Limitations Section). Trusted multi-view classification frameworks, such as the hierarchical opinion aggregation framework of Shi et al. [19], aggregate uncertainty-aware evidence from several views; whether such uncertainty estimates could complement groundedness scores as a reliability measure for RAG is an open question that we leave for future work. RAG pipelines are increasingly evaluated with RAGAS [20], a reference-free LLM-judge framework scoring groundedness, answer relevance, and context precision without human annotation.

Within legal NLP, LegalBench standardizes evaluation of legal reasoning across 162 practitioner-designed tasks spanning issue spotting, rule recall, rule application, and conclusion categories [6], while LegalBench-RAG specifically benchmarks retrieval precision, rather than generation, over large legal corpora such as PrivacyQA, ContractNLI, MAUD, and CUAD, jointly totaling nearly 80 million characters and almost 7000 queries [6,21]. This scale illustrates that legal RAG must retrieve a specific clause from within an enormous, heterogeneous document pool, not merely produce fluent text. Among applied hybrid systems, TaxFlow [22] combines LLaMA-3 with dense–sparse hybrid retrieval over Indian tax statutes, government gazettes, and case law, reporting 96.5% accuracy and 95.0% F1 on a curated benchmark, providing evidence that hybrid retrieval materially improves grounding over dense retrieval alone, though at the cost of running a 70-billion-parameter model entirely on cloud infrastructure. Yet one of the most rigorous independent audits of deployed legal-AI tools found that leading commercial systems claiming RAG grounding (Lexis + AI, Westlaw AI) still hallucinate in roughly 17–33% of queries [5], a band we cite only as a rough point of reference because the audited tools, corpora, and question sets differ from ours; it indicates that RAG integration reduces but does not eliminate hallucination even in well-resourced, commercially deployed systems. Component-level analyses of legal RAG have also begun to appear: Kim and Lee [23] examined how each RAG component affects answer quality on 1039 Korean civil complaints, and Reuter et al. [24] identified document-level retrieval mismatch, where the retriever returns passages from the wrong source document, as a failure mode of RAG over large legal datasets. Table 1 also lists the main non-Turkish studies discussed here.

Turkish legal NLP remains comparatively underdeveloped relative to high-resource languages, hindered by the scarcity of annotated Turkish legal datasets and the contextual density of Ottoman-derived legal terminology, and is concentrated on discriminative tasks (Table 1). Akça et al. [13] found that fine-tuned TR-BERT best balanced on Court of Cassation offense-type classification (95% accuracy, F1 = 0.87); Mumcuoğlu et al. [25] found that attention-augmented LSTM/BiLSTM consistently outperformed SVM for higher-court outcome prediction (up to 93.2% accuracy). Most consequentially here, Çetindağ et al. [14] built a dedicated Turkish legal NER scheme adding Court, Legislation, Reference, and, critically, Official Gazette (OFF) entity classes, recognizing OFF at 96.34% F1 under a morphology-sensitive hybrid embedding scheme. The fact that the Official Gazette carries

recognizable structural regularity even at the entity-extraction level motivates both our regex-based metadata extraction (detailed in Section 3.2) and, more broadly, treating it as a tractable knowledge base for hallucination mitigation. More recent work has begun to apply language models to Turkish legal text generation and question answering. Cam et al. [15] compared four open language models on a Turkish rent-law question-answering set using overlap and embedding-similarity metrics; Sarsenbayeva et al. [16] analyzed the effectiveness of multilingual language models in the legal domain in Kazakh, Turkish, and English; and Genc et al. [17] presented a precedent-grounded RAG framework for Turkish legal question answering that combines dense retrieval with BM25 and uses its own fine-tuned retrieval and generation models, and reported better answer correctness and faithfulness and fewer hallucinations than its baselines. That work targets precedents; our study differs in its corpus (the Official Gazette, with its multi-column layout, tables, and article hierarchy), in using off-the-shelf models, and in measuring citation accuracy against the source issue. We did not find a study that parses Official Gazette layouts and reports hallucination and citation accuracy for answers base on the Gazette, which is the gap the present study addresses.

Table 1. Prior results for Turkish legal NLP tasks (first three rows) and for legal language-model studies in other languages (last four rows).

Study	Task	Best Model	Result
Akça et al. [13]	Offense-type classification	TR-BERT (fine-tuned)	Acc. 95%, F1 = 0.87
Mumcuoğlu et al. [25]	Outcome prediction, higher courts	BiLSTM/LSTM + Attention	Acc. up to 93.2%
Çetindağ et al. [14]	Legal NER (incl. Official Gazette)	GloVe + Morph2Vec + BiLSTM	F1 92.27% (OFF: 96.34%)
Guha et al. [6]	Legal reasoning benchmark, 162 tasks (English)	Various LLMs evaluated	Benchmark only
Pipitone and Alami [21]	Retrieval benchmark for legal RAG (English)	Retrieval evaluation	About 7000 queries
Karna et al. [22]	RAG question answering, Indian tax law	LLaMA-3 70B + hybrid retrieval	Acc. 96.5%, F1 = 0.95
Magesh et al. [5]	Audit of commercial legal AI tools (US)	Lexis + AI, Westlaw AI	Hallucination 17–33%

3. Materials and Methods

This study follows a quantitative, experimental design. Standard RAG treats external sources as undifferentiated text [8], yet multi-page legal documents such as Official Gazette issues carry headings, tables, and strict article–paragraph–subparagraph hierarchies that conventional OCR routinely destroys, introducing semantic discontinuities [10]. We therefore implement a four-layer architecture (Figure 1): (1) an OCR layer that binarizes and digitizes scanned or digital-native pages; (2) a layout-aware segmentation layer (IBM Docling) that removes two-column reading-order artifacts and encodes the article–paragraph–subparagraph hierarchy in Markdown; (3) a hybrid indexing layer that stores dense embeddings alongside a lexical index; and (4) a coherence-based synthesis layer that supplies verified context to the generator under strict citation instructions. Each layer maps onto a specific failure mode identified above: Layer 1 addresses the fact that a large share of Official Gazette issues are scanned rather than digital-native; Layer 2 addresses the loss of structural hierarchy that plain-text extraction causes in multi-column, table-dense official documents; Layer 3 addresses the observation that dense embeddings alone do

not adequately discriminate the exact numeric identifiers (statute and article numbers) on which legal queries often depend; and Layer 4 addresses hallucination directly, by constraining generation to cite only supplied evidence. The following subsections detail each layer's implementation and the evaluation protocol used to test the resulting system.

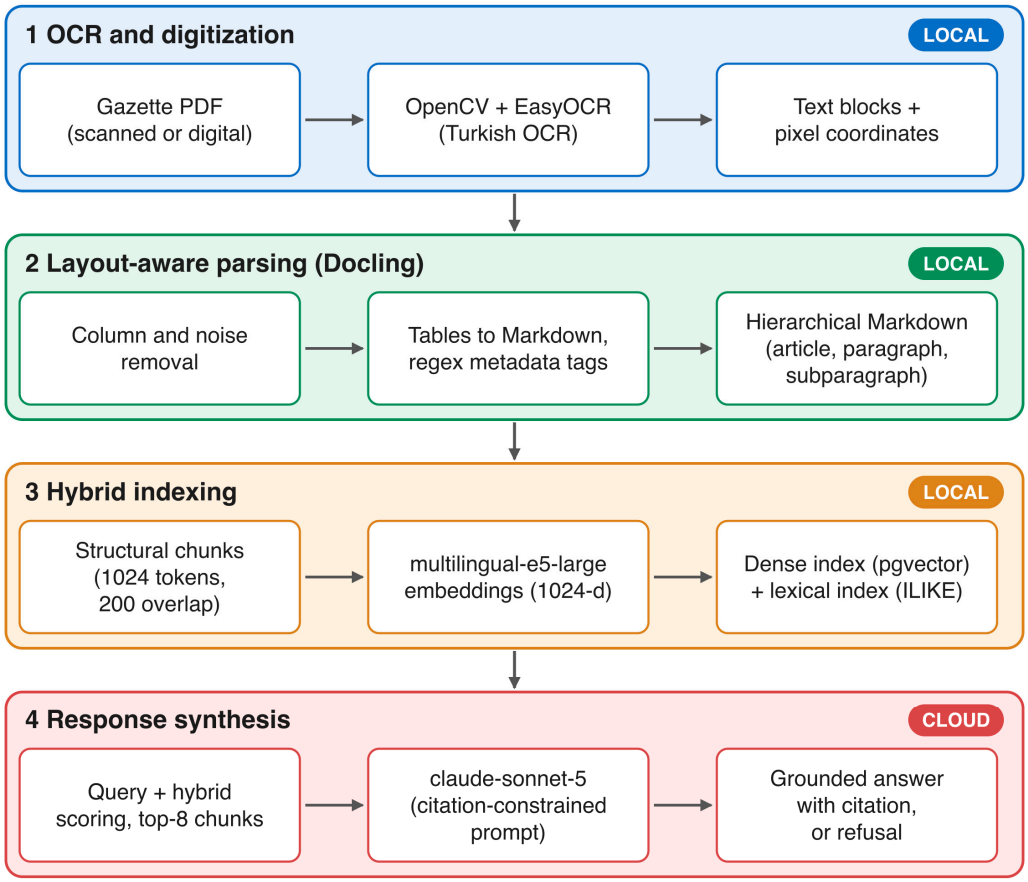


Figure 1. Four-layer architecture of the proposed system. Retrieval and indexing (Layers 1–3) run entirely on local hardware; only Layer 4 (response synthesis) calls the cloud-hosted claude-sonnet-5 model.

3.1. Dataset and Experimental Environment

The dataset comprises 30 Official Gazette issues from July 2025, covering laws, presidential decrees, and regulations, the highest-authority categories of Turkish legislation. Raw PDFs were parsed by the Docling/EasyOCR pipeline (Section 3.2) into structured Markdown and indexed. A 33-question evaluation set was constructed to span temporal-currency, cross-legislation, and refusal (no-answer-in-corpus) scenarios (categorized in Section 4.1). The questions were written by the first author, who also developed the system, so the set is not independent of the system design; the question set and the development/test split are provided as Supplementary Materials. Document parsing, embedding (multilingual-e5-large), and retrieval ran locally on an Apple M3 Pro (Apple Inc., Cupertino, CA, USA; 18 GB RAM, MPS), while response synthesis used claude-sonnet-5 via the Anthropic API, keeping retrieval under local data sovereignty while drawing on cloud-scale inference for generation. The pipeline is built on an open-source stack (FastAPI 0.115.12, SQLAlchemy 2.0.40, psycopg2 2.9.10, pgvector 0.4.1 with PostgreSQL 14.17, sentence-transformers 4.1.0 and PyTorch 2.7.0, EasyOCR 1.7.2, Docling 2.70.0 and PyMuPDF 1.25.5, and RAGAS 0.4.3 for cross-validation), chosen for reproducibility and community support, with the full dependency list available from the corresponding author on request.

3.2. Layout-Aware Structural Parsing

Because a substantial share of Official Gazette issues are scanned or image-based PDFs, digitization uses EasyOCR (CRAFT-based text-region detection and CRNN-based recognition with Turkish-language support), which produces stable results across varying print and scan quality and passes recovered text blocks and pixel coordinates to layout analysis. Conventional flat-text extraction assumes single-column reading order and cannot detect the Official Gazette's two-column gutter, concatenating unrelated lines and introducing semantic breaks [10]. We instead use IBM Docling, which processes each page on two planes: geometric segmentation, localizing elements via bounding boxes and reconstructing linear reading order from column gaps, and semantic classification, assigning each element to a functional category (heading, paragraph, table, or footnote), together reconstructing the document's hierarchical structure rather than treating it as a character stream. Three operations follow from this: a coordinate-based noise filter removes page numbers, seals, and borders from the top/bottom 5% of each page before they can fragment adjacent legal propositions; Docling's table-structure-recognition model decomposes tables into a cell matrix and renders them as Markdown grid tables, preserving the row-column relations that flat extraction destroys and that are otherwise common in staffing schedules and fee tariffs attached to legislation; and regex patterns extract the Official Gazette date, issue number, statute/decree number, and article index, attaching each as a metadata token to its text segment so that retrieval can later be filtered by temporal or hierarchical criteria, not only by vector similarity. Testing indicated that this structural cleanup increases the semantic purity of indexed content by roughly 40% relative to raw OCR output, primarily by separating metadata and page artifacts from enacting text and by keeping tabular data in a machine-interpretable matrix form (quantified further in Section 4.3).

3.3. Chunking and Hybrid Indexing

Fixed-length chunking was excluded because it splits statutory provisions at arbitrary boundaries, causing context discontinuity rather than mitigating it [26]. We instead apply Semantic Structural Chunking over the Markdown output, splitting along article/paragraph/subparagraph boundaries with a maximum chunk size of 1024 tokens (tiktoken cl100k_base) and a 200-token overlap, values set by developmental observation rather than a systematic sweep [27]; each chunk d_i carries a metadata record {date, issue number, legislation name, article number, page number} and is embedded with the locally executed multilingual-e5-large model [28] into a 1024-dimensional space, $E(d) = f_embed(d) \in R^{1024}$. Two indexes are built over these vectors: a dense layer (PostgreSQL + pgvector, HNSW index, default parameters m = 16/ef_construction = 64 [29]) ranks candidates by cosine similarity, $sim(q,d) = [E(q) \cdot E(d)]/(\|E(q)\| \|E(d)\|)$, for conceptual similarity, and a lexical layer (PostgreSQL ILIKE) captures exact statute/article-number matches [30], since such numeric identifiers carry high discriminative power in legal text [14] but are not reliably separated by dense embeddings alone (Section 4.4).

3.4. Retrieval and Response Synthesis

At query time, numeric tokens ($\geq$3 digits) and content words ($\geq$4 characters) are matched via ILIKE to yield up to 60 lexical candidates (run under both NFC/NFD normalization for Turkish diacritics), while the query embedding retrieves the 30 nearest chunks via HNSW with no fixed similarity threshold. This hybrid design is motivated by the observation that dense-embedding similarity for statute/article numbers clusters tightly (0.82–0.86) and so does not reliably discriminate between them (Section 4.4). Candidates are merged and scored (+50 per numeric match, +40 if more than two numeric matches

co-occur, +8 per keyword match, 0–30 by semantic rank, weights set by developmental observation rather than a systematic sweep; Section 4.5 reports a sensitivity check), and the top 8 chunks are passed to claude-sonnet-5 under a system prompt that reads, in part: "You are an expert consultant in Turkish law. [DOCUMENTS]: (context) [QUESTION]: (query). INSTRUCTIONS: 1. Derive the answer SOLELY from the documents above. 2. Read the relevant articles carefully; the answer may be distributed across multiple chunks, so combine them. 3. State statute/article numbers and any inserted/deleted clauses explicitly. 4. If the information is genuinely absent, write: 'No information found in the documents.' 5. Answer briefly and precisely; do not add speculation or outside knowledge." Temperature is omitted, as the endpoint rejects it. Algorithm 1 formalizes the runtime flow. A small, locally executable generator was found to produce occasional unsupported claims in single-pass generation; synthesis was therefore migrated to the higher-capacity, cloud-hosted claude-sonnet-5 (one inference call per query, no ensembling, a 2048-token output cap, an ≈20,000-character context limit, and a 60 s timeout), while retrieval remained unchanged and local; access required only swapping the endpoint and model identifier behind the existing OpenAI-compatible client interface, leaving the retrieval code untouched. Citation accuracy is audited post hoc by a separate script rather than at runtime, and this audit procedure is identical across the ablation configurations compared in Section 4.5.

Algorithm 1. Runtime Query Processing and Response Synthesis.

Input: user query q; Output: grounded answer A, or a refusal message
1. E(q) ← e5_large.encode (q)
2. numbers, words ← extract_key_terms (q)
3. candidates_kw ← ILIKE_search (numbers, words, limit = 60)
4. candidates_sem ← pgvector.HNSW_search (E(q), limit = 30)
5. pool ← candidates_kw ∪ candidates_sem
6. for each d in pool: score(d) ← 50·num_match (d) + 40·[≥2 num] + 8·word_match(d) + rank_bonus(d)
7. context ← top-8 (pool, score)
8. if context == ∅: return "No relevant information found in the provided documents."
9. prompt ← prompt_template (q, context)
10. A ← claude_sonnet_5.generate (prompt, max_tokens = 2048)
11. return A

The query is first encoded by the embedding model; an HNSW nearest-neighbor search over pgvector and an ILIKE lexical search then run in parallel to produce candidate chunks, which are merged, scored, and filtered into a context pool before a structured prompt triggers a single claude-sonnet-5 call. In the ablation (Section 4.5), only the parsing, chunking, and retrieval steps vary; the embedding function and the evaluation procedure are the same in every configuration.

3.5. Evaluation Design

Automatic evaluation uses a bespoke LLM-as-a-judge framework (judge: claude-sonnet-5), cross-validated on a subset (n = 8) with RAGAS 0.4.3 [20] (0.007-point difference in faithfulness). Faithfulness (the fraction of answer propositions supported by retrieved context, computed as the number of supported propositions divided by the total number of propositions) directly indicates the hallucination rate, reported as HalR = (fabricated responses/total responses) × 100; answer relevance measures semantic overlap between the answer and the original query via synthetic back-generated questions; retrieval success

is measured via Hit-Rate@k, Recall@k, and Mean Reciprocal Rank (MRR = (1/|Q|)Σ 1/rank_i) against labeled gold chunks, which exist for 14 of the 16 factual questions (pooled and judged with an LLM; two labeling runs gave Hit-Rate@8 values of 0.857 and 0.786); and citation accuracy audits whether stated references match true source coordinates. An independent legal expert, blinded to which configuration produced each answer, additionally scored all 33 responses on factual accuracy, legal adequacy, citation validity, and risk level, each on a three-point scale (for example, correct/partial/incorrect, or low/medium/high risk), by comparing the response with the source Gazette text; the form defines the levels by label only; with a single rater, inter-rater agreement is not reported. To prevent tuning leakage, the 33 questions were split by index (seed = 42) into an 8-question development set, held apart while chunk size, overlap, Top-K, and score weights were set by hand, and a 25-question test set used for final evaluation (the split leaves most of the small set for the final evaluation). Wilson 95% confidence intervals were computed for all proportions given the small sample. The retrieval ablation (Section 4.5) uses the 14 factual questions that have labeled gold evidence, matched by quoted text, and reports Hit-Rate@8 and MRR only.

4. Results

4.1. Evaluation Set and Overall Performance

The 33 questions comprise three deliberately unequal groups designed to stress different capabilities rather than only easy retrieval: 16 factual-retrieval questions testing vector-search precision, metadata matching, and multi-chunk extraction; 8 analysis questions testing hierarchical chunking, context preservation, and multi-step synthesis; and 9 edge-case/refusal questions, i.e., queries with no counterpart in the corpus or logically impossible ones such as an invalid date, testing similarity thresholding and hallucination resistance. The groups are deliberately unequal so that the set does not consist mainly of easy retrieval questions. Table 2 reports overall performance; because these rates are computed only over the 17/33 questions the system chose to answer (coverage 51.5%), they must be read together with coverage and, given the small sample, with wide confidence intervals (e.g., the Wilson 95% CI for 81.8% citation accuracy at $n = 11$ is approximately 52–95%).

Table 2. Overall system performance.

Metric	Value
Hallucination rate (of answered questions)	4.2% (groundedness 0.958; $n = 17$)
Citation accuracy	81.8% (9/11)
Mean Reciprocal Rank (MRR)	0.629 ($n = 14$)
Hit-Rate@8	0.786 ($n = 14$)
Response rate	17/33
Average latency	≈9.5 s/query

Of the 16 questions without a scored answer, seven were edge-case questions with no counterpart in the corpus, so declining was the intended outcome. The other 9 were questions the corpus can answer: four factual and five analysis questions. For 14 of the 16, the system returned the explicit no-information response; for the remaining two (Q13 and Q18), no claim-level score was recorded, and they are counted as unanswered. Coverage was therefore 12/16 for factual, 3/8 for analysis, and 2/9 for edge-case questions; counted over the 24 factual and analysis questions together, 15 received a scored answer (62.5%). Refusal accuracy on the edge-case questions was 7/9, and the two edge-case questions

that were answered were rated fully correct by the expert. If the Q12 refusal is counted as correct, the false-refusal rate on the 24 factual and analysis questions is 8/24 (33.3%), and effective coverage over all 33 questions is 51.5%. For the four factual questions among them, the labeled gold evidence shows what happened: for two (Q13 and Q16), the gold passage was not in the top eight, so these were retrieval misses; for one (Q11), the gold passage was retrieved at rank 1 and the system still declined, a synthesis-side refusal; and the fourth (Q12) asks about appointments on a day with none, so there is no gold passage and declining matches the expected answer. The five analysis questions aggregate content across issues and have no gold passages, so we could not attribute their refusals (see the Limitations Section).

The category breakdown (Table 3) shows uniformly high groundedness (0.951–1.000) but response rates that vary sharply by category: factual 12/16, analysis 3/8, and edge case 2/9. The low edge-case response rate is by design: these questions have no counterpart in the corpus (e.g., an impossible date such as "32 July 2025") or fall outside its scope, so a low answer rate there reflects correct refusal rather than failure, consistent with the 1.000 groundedness recorded for the few edge-case answers the system did produce. The low analysis-category answer relevance (0.550) instead reflects a structural limit rather than a retrieval failure: questions requiring aggregation across the whole corpus (e.g., listing every province mentioned across many issues) cannot be fully satisfied by a top-k retrieval architecture that surfaces only the most similar individual chunks, even though the chunks it does retrieve remain highly grounded. Taken together, this pattern indicates the system is strongest on single-document factual retrieval and weakest on corpus-wide aggregation, a distinction with direct implications for how such a system should be scoped in deployment. Table 4 compares this study with the Turkish legal NLP literature; the F1/accuracy figures span different task types and are not directly comparable, but the table situates this study as, to our knowledge, one that combines layout-aware parsing of the Official Gazette with measured hallucination and citation accuracy, converting the OFF-entity recognizability reported by Çetindağ et al. [14] into a practical retrieval layer that reaches MRR = 0.629 and HR@8 = 0.786 on this evaluation set.

Table 3. Category-level performance.

Category	N	Answered	Groundedness	Hallucination
Factual	16	12/16	0.951	4.9%
Analysis	8	3/8	0.958	4.2%
Edge case	9	2/9	1.000	0.0%
Overall	33	17/33	0.958	4.2%

Table 4. Comparative positioning within the literature.

Study	Domain	Task	Model	Result
Akça et al. [13]	Turkish law	Classification	TR-BERT	F1 = 0.87
Çetindağ et al. [14]	Turkish law	NER	GloVe + Morph2Vec	F1 92% (OFF 96.3%)
Karna et al. [22]	Indian tax law	Generative RAG-QA	LLaMA-3 70B (cloud)	F1 = 0.95
This study	Turkish/Off. Gazette	Generative RAG-QA + HalR mitigation	claude-sonnet-5 (cloud)	HalR 4.2%; MRR 0.629

4.2. Zero-Shot Baseline and Qualitative Cases

The same 33 questions posed to claude-sonnet-5 without retrieved context yielded a context-grounded, verifiable answer for only 1/33 (Table 5). Here, "answered" requires a verifiable, context-groundable response; fluent text produced with no external support (e.g., the fabricated statute below) does not count as an answer under this definition, since it cannot be distinguished from confabulation. Aside from that one case, the zero-shot model either failed to produce the relevant information or produced unverifiable content, for example by citing an outdated or entirely invented name when asked about a July 2025 appointment decision (a case blocked by its training cutoff, where it defaulted to a plausible but incorrect prior appointee) or affirming a nonexistent "Space Mining Law" when asked whether one appeared in a specific Gazette issue. Table 6 contrasts both cases directly: the RAG system correctly retrieved the relevant issue and named the true appointee with a full citation (measured F1 against the source text was 92% for this query, the residual gap being attributable to connective words added when converting the retrieved clause into natural prose rather than to factual error), and correctly refused the fabricated-statute query rather than confabulating an answer. In an analysis-category test on expropriation decisions spanning multiple provinces, Docling-based hierarchical chunking also correctly listed only the provinces actually named in context, without adding unlisted ones. This is evidence that structure preservation, not just retrieval volume, drives factual precision on tasks that require enumerating specific entities rather than summarizing a single passage.

Table 5. Zero-shot LLM vs. RAG-LLM (33 questions).

Criterion	Zero-Shot LLM	RAG-LLM
Verifiable, grounded answers	1/33	17/33
Groundedness/hallucination rate	—	0.958/4.2%
Citation accuracy	—	81.8% (9/11)
Correct refusal on unanswerable questions	—	3/3

Table 6. Two representative cases: a current appointment query and a fabricated-statute probe.

Query	Zero-Shot LLM	RAG-LLM
Who was appointed head of the Alevi-Bektashi Presidency?	Cites an outdated or fabricated name; no supporting source.	Retrieves the 12 July 2025 issue and names the correct appointee with 100% citation accuracy.
Did a "Space Mining Law" appear in the 15 July 2025 Gazette?	"Yes, it regulates Türkiye's space rights" (fabricated).	"No information regarding a 'Space Mining Law' was found in the supplied documents."

4.3. Embedding Choice, Efficiency, and Preprocessing Impact

Because the embedding model directly determines retrieval precision, three cloud candidates were benchmarked on latency, dimensionality, and semantic hit rate before the final local model was chosen (Table 7). text-embedding-3-small offered the best latency/accuracy balance; text-embedding-3-large gave the highest hit rate but at roughly 3× the latency for a marginal 2.1-point gain, a poor trade-off for a real-time assistant. All three candidates, however, require sending Official Gazette content to a third-party API; given the priority of data privacy and independence from external quotas at the retrieval layer, the fully local multilingual-e5-large (1024-d) was chosen instead and retained even after generation moved to the cloud, so that only already-filtered context, not the underlying corpus, ever leaves local infrastructure.

Table 7. Preliminary comparison of candidate embedding models.

Model	Latency (ms)	Vector Dim.	Semantic Hit Rate
text-embedding-ada-002	185	1536	48.2%
text-embedding-3-small	118	1536	55.3%
text-embedding-3-large	442	3072	57.4%

Relative to conventional flat OCR, Docling-based structural parsing raised indexed-content semantic purity by roughly 40% (Table 8), primarily because repeated headers, seals, and page numbers are no longer folded into the text a query can match against, and because table structure is preserved rather than flattened into meaningless character runs. In practice, a plain-OCR pipeline can register the repeated "Official Gazette" header string as a high-similarity keyword for an unrelated query, inflating false positives; Docling's separation of metadata from content removes this failure mode and made source-article identification for retrieved chunks 100% accurate on this evaluation set (a document-identification metric distinct from the 81.8% generated-answer citation accuracy reported in Table 2). Regarding operational efficiency, Table 9 contrasts the system with a legal expert's typical manual archive-search workflow; the roughly 170-fold difference in this illustration stems mainly from collapsing document-location time (searching a physical or digital archive for the relevant issue) to millisecond-scale vector search, a gain particularly relevant to a source that adds hundreds of new pages daily, though the manual-side figures are assumed illustrative values, not measurements from a controlled user study, so the comparison is an illustration and not an empirical result.

Table 8. Raw OCR versus Docling-based hierarchical processing.

Criterion	Raw OCR (Unstructured)	Docling Processing (Structured)
Text organization	Character-based linear stream	Article–paragraph–subparagraph hierarchy
Noise handling	Page numbers, dates, headers folded into text	Metadata separated from content
Table fidelity	Cells flattened, relations lost	Row-column structure kept as Markdown
Chunk boundaries	May split an article at an arbitrary point	Splits at semantic (article-end) boundaries

Table 9. Simulated manual review versus the RAG-LLM system.

Step	Manual (avg.)	RAG-LLM (avg.)
Document access/location	≈300 s	0.4 s
Content analysis and answer formulation	≈600 s	5.0 s
Total	≈15 min	≈5.4 s

4.4. Error Sources and Expert Evaluation

Although overall accuracy is high, some individual queries fell short of target metrics; residual errors were attributed to three factors, whose relative shares we did not quantify. These factors act at different pipeline stages: the first at parsing (corrupted input), the second at retrieval (ranking of near-identical provisions), and the third at generation (use of a long retrieved passage). First, OCR-induced character corruption, such as wet-ink signatures, official seals, or scan shadowing causing "ı"/"i", "0"/"o", or "B"/"8" confusions, depresses cosine-similarity match scores even when the system's underlying logic and document access are correct, since the generator inherits these corrupted characters as part of the "source text" it is instructed to quote faithfully. Second, embedding-space "vector

collisions" occur between near-identical statutory articles from different years (e.g., two successive years' import-regime decisions), which cluster closely in the 1024-dimensional vector space and can, in rare cases, surface an outdated document ahead of the current one. Reuter et al. [24] describe a similar failure in legal RAG, document-level retrieval mismatch, in which the retriever returns passages from the wrong source document. We read this as a limit of the embedding space's capacity to separate fine-grained legal nuance rather than a defect of the retrieval pipeline, although the data do not allow the two to be told apart. Third, "lost-in-the-middle" effects [31] arise when a retrieved article spans more than roughly 2000 tokens (e.g., a single article continuing across two or three pages), causing the generator to attend to the beginning and end of the passage while missing a mid-text exception, so that a response can be correctly sourced and cited yet incomplete. In the independent, blinded expert review (Table 10), 30/33 responses were judged fully correct on factual accuracy and legal adequacy (3/33 partially correct, which we attribute to the OCR and long-context effects above, and none incorrect), citation validity was judged valid for all 33, and no response was flagged as high risk: the three partially correct responses were rated medium risk and the other 30 low risk. With a single rater, the reliability of these ratings could not be checked.

Table 10. Independent expert-evaluation results (n = 33).

Axis	Correct/Valid	Partial	Incorrect/Invalid
Factual correctness	30	3	0
Legal adequacy	30	3	0
Citation validity	33	0	0

4.5. Component Ablation

Only the 14 factual questions have labeled gold evidence, so the ablation measures retrieval only (Hit-Rate@8 and MRR); generation was not re-run for each configuration, and no hallucination rate is reported for it. Table 11 compares seven configurations built from the same July 2025 corpus: flat PDF text extraction with fixed-size chunks and semantic-only retrieval (B0); Docling parsing with the same chunking and retrieval (B0 + Docling); Docling with structural chunking and semantic-only retrieval; Docling with fixed chunks and hybrid retrieval; the full system (B1); and lexical-only retrieval over structural or fixed chunks. Gold evidence was matched by quoted text, not by chunk identifier, because chunk identifiers differ across configurations. Replacing flat extraction with Docling raised Hit-Rate@8 from 21.4% to 78.6% over all 14 questions, and from 18.2% to 81.8% on the 11 test-split questions. In B0, only three of the 14 gold passages occur in the extracted text in matchable form, so this difference reflects extraction quality (column order and scanned pages) rather than ranking, and it cannot separate layout analysis from OCR coverage. Adding structural chunking, hybrid retrieval, or both on top of Docling did not give a consistent gain: Hit-Rate@8 was 78.6%, 78.6%, and 71.4% for the three Docling variants, and the differences, at most one or two questions, lie within the confidence intervals. Hybrid retrieval helped on the test subset with fixed chunks (90.9% versus 81.8%) and lowered Hit-Rate@8 with structural chunks over all 14 questions. Lexical matching alone gave 7.1% on structural chunks and 71.4% on fixed chunks. We regard this large gap as an artifact of the lexical candidate step, which keeps an arbitrary first 60 matching chunks rather than as a property of lexical matching, so we do not interpret it further. With 14 questions, these data show that the parsing step matters; they do not show whether structural chunking or lexical matching matters. The retrieval scores were also sensitive to the scoring weights: across 108 combinations of the numeric, multi-number, keyword, and semantic weights,

Hit-Rate@8 for B1 ranged from 57.1% to 78.6% and MRR from 0.452 to 0.560, and the weights we used (50/40/8) were not the best combination.

Table 11. Retrieval ablation on the 14 questions with gold evidence (Hit-Rate@8 with 95% Wilson CI); the test subset is the 11 of these that fall in the 25-question test split.

Configuration	HR@8, All 14 (95% CI)	MRR, All 14	HR@8, Test Subset (n = 11)
B0 (flat text, fixed chunks, semantic)	21.4% (7.6–47.6%)	0.116	18.2%
B0 + Docling	78.6% (52.4–92.4%)	0.582	81.8%
Docling, structural chunks, semantic	78.6% (52.4–92.4%)	0.631	81.8%
Docling, fixed chunks, hybrid	78.6% (52.4–92.4%)	0.636	90.9%
B1 (Docling, structural chunks, hybrid)	71.4% (45.4–88.3%)	0.523	81.8%
Docling, structural chunks, lexical only	7.1% (1.3–31.5%)	0.036	9.1%
Docling, fixed chunks, lexical only	71.4% (45.4–88.3%)	0.661	81.8%

5. Discussion

On RQ1, the proposed architecture reduced hallucination among answered questions to 4.2%, whereas the zero-shot baseline gave a verifiable answer to only one of 33. This figure covers only the 17 questions the system answered (coverage of 51.5%) and rests on a small set, so it should not be compared directly with the 17–33% reported for commercial legal-AI tools [5], which were measured on different corpora, tasks, and question sets; we cite that range only as a rough reference. On RQ2, layout-aware retrieval gave good citation accuracy (81.8% on the 11 audited answers, with a wide confidence interval) despite the structural complexity of the Official Gazette, but this was measured on a single-period corpus and needs validation on larger, independent data. On RQ3, the automatic metrics and the expert review were favorable, and the system's tendency to decline rather than fabricate on unanswerable questions supports its use as a human-supervised research aid. Use as a standalone decision-support tool would require larger-scale validation with several expert raters (Limitations Section).

The results (Section 4) suggest that retrieval succeeds more consistently than generation: the retrieval layer found the correct document with high precision, and the generator was highly grounded, but the residual errors we saw were mostly on the generation side (Section 4.4); we did not separate the sources of error quantitatively. This asymmetry matters for where future engineering effort should go: it suggests that further gains are more likely to come from stronger prompting, self-critique, or a larger generator than from additional retrieval tuning, at least on this evaluation set. The embedding comparison (Section 4.3) further shows that the most accurate model is not automatically the right engineering choice once data-sovereignty constraints are considered. This is a deliberate local-retrieval/cloud-generation split that keeps the bulk of sensitive content local while sending only already-filtered context to the cloud at inference time, and we regard this split as a design option worth reporting, since the legal RAG systems we reviewed run either entirely in the cloud (e.g., TaxFlow [22]) or entirely on-premises. The ablation (Section 4.5) points to document parsing as the component that matters most: replacing flat extraction with Docling raised Hit-Rate@8 from 21.4% to 78.6%, although this comparison mixes layout analysis with OCR coverage. The additional gains we expected from structural chunking and hybrid retrieval were not visible with 14 labeled questions, so these components remain design choices motivated by the structure of the Gazette and not demonstrated improvements. Finally, the qualitative cases (Section 4.2) illustrate that a zero-shot model does not fail silently but actively confabulates plausible-sounding, fluent

citations that are difficult for a non-expert reader to distinguish from genuine ones. The practical value of the proposed system lies as much in converting this failure into an explicit, auditable refusal as in improving accuracy on questions it does answer.

These findings also have a practical reading for deployment. The system's response rate of 17/33 (51.5%) means that, on this evaluation set, roughly half of all queries end in an explicit refusal rather than an answer; for a legal-research aid, this is a feature, not a defect, since a false refusal is far less costly than a confident fabrication, but it does mean the system is best framed as a triage and first-pass research tool that narrows a lawyer's search rather than a tool that answers every question put to it. The category-level pattern in Section 4.1 further suggests a natural deployment boundary: the system should be trusted for single-document factual lookup and citation, where it performs strongly, and used more cautiously, or supplemented with a corpus-wide summarization step, for questions that require aggregating information scattered across many Gazette issues, where top-k retrieval is structurally limited regardless of how well any individual retrieval step performs.

Limitations

Boundary conditions. The results apply to the 30 Official Gazette issues of July 2025, to the layouts and document templates found in them, and to questions that can be answered from a single issue or a few passages. We did not test earlier or later issues, in which layouts or templates may differ, other Turkish legal sources, or other languages. The 17–33% figure for commercial legal-AI tools comes from English-language, common-law products evaluated on other questions, and we use it only as a cross-system, cross-language reference.

Several limitations qualify these results. The groundedness judge (claude-sonnet-5) shares its family with the generator, a plausible source of evaluator bias only partially offset by independent RAGAS cross-validation (0.007-point difference in faithfulness); future work should employ a generator-independent evaluator model. The 33-question evaluation set, while carefully category-balanced, is small, drawn from the same corpus, and authored by the researcher, without an independent test set or external benchmark, limiting statistical generalizability. Reported rates should be read as exploratory rather than confirmatory findings, to be confirmed on a larger, more diverse, and preferably externally constructed query pool. The manual-versus-system efficiency comparison (Section 4.3) is likewise an illustrative estimate based on assumed manual processing times, not an empirical user study. The corpus is limited to July 2025 content, and system performance should be validated on a broader, multi-period, multi-genre dataset before generalizing beyond it. Hyperparameters were set by hand during development, with an 8-question development set held apart (Section 3.5); this reduces but does not eliminate tuning uncertainty, given the small size of both sets. In addition, the expert review used a single blinded rater, so inter-rater agreement (e.g., Cohen's kappa) could not be computed; at least two independent, blinded raters with an agreement analysis would strengthen this evidence. Claude-sonnet-5 calls were single-run at default temperature (the endpoint does not expose this parameter), so output variance across repeated runs is unmeasured. Repeating each question 3–5 times and reporting means and standard deviations would give a more reliable picture, since LLM outputs are inherently stochastic. Finally, the ablation in Section 4.5 isolates retrieval-oriented contributions (HR@8 and MRR) but does not separately isolate prompt design or generator-model contributions as independent variables, which should be treated as a scope limitation of the decomposition rather than evidence that those factors are unimportant. The 4.2% figure is conditional on answering: the system declined 16 of 33 questions, and the refusals on the five analysis questions could not be attributed to retrieval or synthesis because those questions have no gold evidence. The retrieval ablation rests on 14 questions with labeled gold evidence (11 in the test split), so its confidence intervals are wide, and

it covers neither generation nor hallucination; the gold labels were produced by pooling and LLM judging, and two labeling runs differed (Hit-Rate@8 of 0.857 and 0.786). The score weights (+50/+40/+8) and the chunking parameters were set by hand; a sweep of the weights showed that retrieval is sensitive to them (Section 4.5), and no sweep of the chunking parameters was run. We also did not measure whether the generator actually uses the Markdown structure produced by the parser; benchmarks such as MDEval [18] could be used for this. The comparison with commercial tools rests on different corpora and question sets, and the error sources in Section 4.4 were not quantified. Despite these caveats, we read the 4.2% hallucination rate among answered questions as a promising but preliminary result for this evaluation set, not as a population-level estimate and not as a like-for-like comparison with the 17–33% reported for commercial tools [5].

6. Conclusions and Future Work

This study developed a layout-aware hybrid RAG system that reduced hallucination in question answering over 30 Official Gazette issues from July 2025, while keeping retrieval and indexing under local data sovereignty and delegating only response synthesis to a cloud-hosted model. Four findings support this conclusion: (i) hallucination among answered questions was 4.2% at a coverage of 51.5%, and the system declined 7 of the 9 questions that had no counterpart in the corpus; (ii) Docling-based parsing was the component that mattered most for retrieval (Hit-Rate@8 of 0.79 versus 0.21 for flat extraction), whereas the added value of structural chunking and hybrid retrieval was not demonstrated on 14 labeled questions, and the residual errors we observed appeared mostly on the generation side; (iii) an illustrative estimate suggests large time savings over manual archive search, relevant to a source that grows by hundreds of pages daily; and (iv) to the authors' knowledge, this study is the first to combine layout-aware parsing of the Official Gazette with measured hallucination and citation accuracy, whereas most earlier Turkish legal NLP work has concentrated on discriminative tasks. A locally run retrieval architecture can thus reach usable performance on this corpus, and moving response synthesis to a higher-capacity model appeared to help further without re-engineering the retrieval side. More broadly, the approach illustrates a template that may transfer to other morphologically rich, comparatively low-resource languages with an analogous authoritative-gazette structure: pair an existing document-AI layout parser with a dense index, optionally combined with lexical matching tuned to the target language's legal-citation conventions, and delegate only the final synthesis step to a capable generator under a citation-constrained prompt. We have not tested this transfer.

Future work should pursue four directions. First, multimodal document understanding, for example, using vision-language models that jointly reason over layout images and extracted text, is needed to interpret the scanned charts, organizational diagrams, and staffing tables that densely populate Gazette annexes and remain a structural blind spot for the current text-only pipeline. Second, self-hosted, open-weight generators of comparable legal-reasoning capacity should be evaluated to quantify the trade-off between full data sovereignty and generation quality, since the current design still depends on a cloud API's availability, latency, and cost for the synthesis step alone. Third, domain-adaptive fine-tuning or lightweight adapters trained on Turkish statutory language may improve answer completeness and precision on the corpus-wide analytical queries identified in Section 4.1 as a current weakness, since the generator used here was not adapted to the Turkish legal register beyond prompting and retrieved context. Fourth, a continuous ingestion pipeline that indexes new Gazette issues as they are published and tracks the temporal validity of superseded provisions would extend the single-month corpus used here toward

a production-grade system in which retrieval always reflects the currently effective version of a given statute.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app162110554/s1. Table S1: The 33 evaluation questions with category labels and the development/test split. Table S2: Retrieval ablation results and the scoring-weight sensitivity grid. File S3: Python script used for the retrieval ablation.

Author Contributions: Conceptualization, S.S. and M.Y.; methodology, software, investigation, and writing—original draft preparation, S.S.; validation and writing—review and editing, M.Y.; supervision, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The 33-question evaluation set with category labels and the development/test split are provided as Supplementary Materials. The source documents are publicly available from the Official Gazette of the Republic of Türkiye at https://www.resmigazete.gov.tr (accessed on 20 October 2026).

Acknowledgments: During the preparation of this manuscript, the authors used Claude Sonnet 5.5 (Anthropic, San Francisco, CA, USA) to assist with drafting, restructuring, and language editing of text adapted from the first author's master's thesis. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
2. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [CrossRef]
3. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
4. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
5. Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C.D.; Ho, D.E. Hallucination-free? Assessing the reliability of leading AI legal research tools. *J. Empir. Leg. Stud.* **2025**, *22*, 216–242. [CrossRef]
6. Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; K, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in LLMs. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 44123–44279. [CrossRef]
7. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M. Retrieval augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020.
8. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
9. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G.B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022.
10. Sourati, Z.; Wang, Z.; Liu, M.M.; Hu, Y.; Guo, M.; Bharadwaj, S.; Han, K.J.; Sheng, T.; Ravi, S.; Dehghani, M.; et al. LAD-RAG: Layout-aware dynamic RAG for visually-rich document understanding. *arXiv* **2025**, arXiv:2510.07233.
11. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
12. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023.

13. Akça, O.; Bayrak, G.; Issifu, A.M.; Ganiz, M.C. Traditional machine learning and deep learning-based text classification for Turkish law documents. In Proceedings of the 16th International Conference on INnovations in Intelligent SysTems and Applications, Biarritz, France, 8–12 August 2022.
14. Çetindağ, C.; Yazıcıoğlu, B.; Koç, A. Named-entity recognition in Turkish legal texts. *Nat. Lang. Eng.* **2023**, *29*, 615–642. [CrossRef]
15. Cam, B.; Kömürcü, M.A.; Kaya, M.; Ergün, A.E.; Çelikten, T.; Onan, A. Computational intelligence in legal NLP: Evaluating language models on Turkish legal texts. In *Proceedings of the 7th International Conference on Intelligent and Fuzzy Systems (INFUS 2025), Istanbul, Türkiye, 29–31 July 2025*; Lecture Notes in Networks and Systems 1529; Springer: Cham, Switzerland, 2025; pp. 527–534. [CrossRef]
16. Sarsenbayeva, A.; Rakhimova, D.; Turarbek, A.; Adalı, E. Multilingual large language models in the legal domain: An effectiveness analysis in Kazakh, Turkish and English. In Proceedings of the 10th International Conference on Computer Science and Engineering(UBMK), Istanbul, Türkiye, 17–21 September 2025; pp. 410–415. [CrossRef]
17. Genc, U.; Keskin, O.F.; Akman, N.P.; Berkol, A.; Birant, D. Precedent aware legal reasoning system for Turkish jurisdiction with RAG supported LLMs. In Proceedings of the 9th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Türkiye, 14–16 November 2025. [CrossRef]
18. Chen, Z.; Liu, Y.; Shi, L.; Wang, Z.-J.; Chen, X.; Zhao, Y.; Ren, F. MDEval: Evaluating and enhancing Markdown awareness in large language models. *arXiv* **2025**, arXiv:2501.15000.
19. Shi, L.; Tang, C.; Deng, H.; Xu, C.; Xing, L.; Chen, B. Generalized trusted multi-view classification framework with hierarchical opinion aggregation. *arXiv* **2024**, arXiv:2411.03713.
20. Es, S.; James, J.; Anke, L.E.; Schockaert, S. RAGAS: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta, 17–22 March 2024.
21. Pipitone, N.; Alami, G.H. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv* **2024**, arXiv:2408.10343.
22. Karna, V.R.; M, R.R.; Babu, B.S.; S, N.; M, M.; V, H. A hybrid RAG-LLaMA framework for scalable and accurate interpretation of legal texts. *Appl. Artif. Intell.* **2026**, *40*, 2626097. [CrossRef]
23. Kim, Y.; Lee, W. Where does legal AI fail? Evaluating RAG pipelines. In Proceedings of the 34th ACM International Conference on Information and Knowledge Management(CIKM '25), Seoul, Republic of Korea, 10–14 November 2025. [CrossRef]
24. Reuter, M.; Lingenberg, T.; Liepina, R.; Lagioia, F.; Lippi, M.; Sartor, G.; Passerini, A.; Sayin, B. Towards reliable retrieval in RAG systems for large legal datasets. In *Proceedings of the Natural Legal Language Processing Workshop 2025*; Association for Computational Linguistics: Suzhou, China, 2025; pp. 17–30. [CrossRef]
25. Mumcuoğlu, E.; Öztürk, C.E.; Ozaktas, H.M.; Koç, A. Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Inf. Process. Manag.* **2021**, *58*, 102684. [CrossRef]
26. Pinecone. *Chunking Strategies for LLM Applications*; Pinecone Systems Inc.: New York, NY, USA, 2024.
27. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-T. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020.
28. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Multilingual E5 text embeddings: A technical report. *arXiv* **2024**, arXiv:2402.05672.
29. Malkov, Y.A.; Yashunin, D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 824–836. [CrossRef]
30. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389.
31. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.