Section-Aware Retrieval for Turkish Case Law: A Case Study on Scanned Court Decisions

Kübra Kovayçin[1], Ece İrem Şişer[2], Derda Sina Günay[2], Yusuf Evren Aykaç[1,2], and Refik Samet[2]

[1] Ankara Yıldırım Beyazıt University, Ankara, Türkiye
[2] Ankara University, Ankara, Türkiye
yeaykac@aybu.edu.tr

Abstract. Finding relevant precedents often needs more than topical overlap: operative facts and judicial reasoning can be decisive. In Turkey, court decisions are often long and frequently shared as scanned (image-based) PDFs, making extraction and indexing harder. We present a section-aware retrieval case study on 1,000 Turkish court decisions about social media–related disputes. Using a vision–language model, each PDF is transcribed and organized into `Facts`, `Reasoning`, and `Verdict` segments; statute/article mentions are also extracted as lexical anchors. We compare a lexical baseline (BM25) with multilingual and legal bi-encoders adapted via parameter-efficient fine-tuning, and we test lexical–dense hybrid fusion, including a simple section-aware hybrid that scores sections separately under fixed token budgets. Because expert-labeled precedent links are not available, we use a self-alignment known-item diagnostic: an LLM generates a short, query-like summary, and the system tries to retrieve the decision it came from; full-document BM25 is included as a reference baseline. Under this proxy, BM25 over `Reasoning` is the strongest single-section index (R@1=0.744), and section-aware hybrid fusion with MPNet reaches R@1=0.780 and MRR@10=0.810, outperforming naive fusion while preserving section-level interpretability. This proxy setup is intended as a diagnostic rather than a full relevance benchmark; retrieval-time scoring is deterministic, using BM25 and cosine similarity over fixed embeddings.

Keywords: Legal information retrieval · precedent retrieval · section-aware retrieval · BM25 · neural bi-encoder models · sparse–dense hybrid fusion · large language models (LLMs) · Turkish case law

1 Introduction

Finding persuasive prior decisions is central to legal research and case preparation. In practice, a helpful precedent is not only about the same topic; it should also align with key facts and with how the court reasons about the law [7, 19].

Working with Turkish decisions adds practical issues. Turkish is morphologically rich, which can increase lexical mismatch in term-based retrieval [26]. Many

decisions are also distributed as scanned, image-based PDFs. When the source is noisy, straightforward OCR and rule-based section splitting can be brittle, and the resulting text may not be ready for retrieval without additional processing.

Evaluation is another limitation. Benchmarks such as COLIEE provide labeled links between cases [10], but we are not aware of a public Turkish dataset with expert-labeled precedent links for our target domain. For this reason, we use a controlled *self-alignment* proxy: for each decision we generate a short, query-like summary and ask the system to retrieve the same decision from the corpus. We treat this as a diagnostic for section usefulness and fusion choices, not as a substitute for expert relevance judgments.

In this work, we compile 1,000 Turkish judicial decisions involving social media–related disputes. We process the scanned PDFs with a vision–language model and obtain structured `Facts`, `Reasoning`, and `Verdict` segments (Section 3). We then compare sparse retrieval (BM25), dense bi-encoders, and simple lexical–dense hybrid fusion under identical section and token-budget constraints.

We study three *research questions*: (RQ1) Which sections are most useful under the self-alignment proxy? (RQ2) Does section-aware fusion help compared to naive concatenation under fixed token budgets? (RQ3) How do lexical anchors (e.g., statute mentions) and semantic similarity interact in early-rank retrieval?

Contributions. In this case study, we:

(1) **create** a section-structured corpus of 1,000 Turkish decisions by extracting role-labeled text and metadata from scanned PDFs with a VLM pipeline (with schema validation and reprocessing);
(2) **run** controlled comparisons of sparse, dense, and hybrid retrievers across different sections using the self-alignment proxy;
(3) **describe** a simple section-aware hybrid fusion template and report observations that may inform future, expert-labeled evaluations.

The remainder of the paper is organized as follows: Related Works reviews prior research; Method describes corpus construction, the proxy diagnostic, and retrieval models; Results reports quantitative findings; Discussion discusses implications and limitations; and Conclusion summarizes takeaways and future directions.

2 Related Work

We situate our study within prior work on legal case retrieval, long-document retrieval, Turkish legal NLP, and LLM-assisted preprocessing.

2.1 Legal Case Retrieval and Evaluation

Precedent retrieval is often framed as a *query-by-document* task, where both the query and the candidate are long, structured decisions and relevance depends on operative facts and legal reasoning [7, 19]. When expert labels exist,

benchmarks such as COLIEE enable more direct comparison [10]. Some work also uses citation or linkage structure with graph-based or transformer-based models [22, 23], but this typically requires machine-readable links that are not available in many low-resource corpora. Because expert judgments are costly and jurisdiction-specific [7], proxy evaluations (e.g., self-retrieval or weak supervision) are often used for controlled diagnostics, but they do not replace true precedent relevance [24].

2.2 Sparse, Dense, and Hybrid Retrieval for Long Decisions

BM25 is still a common and efficient first-stage baseline in law, partly because legal text contains strong lexical anchors such as citations and terminology [18]. Dense bi-encoders can help with paraphrase and vocabulary mismatch [12], while learned-sparse and late-interaction models provide intermediate trade-offs [8,13]. Hybrid fusion methods (score fusion or rank fusion) combine lexical and semantic signals and can be more robust when scores are not well calibrated [5, 28]. For long, heterogeneous legal decisions, several studies report benefits from passage- or structure-aware aggregation instead of naive full-text concatenation [1, 2, 14, 20].

2.3 Domain Adaptation and Turkish Legal NLP

Domain-adapted encoders (e.g., Legal-BERT) and multilingual backbones such as XLM-R are commonly used starting points for legal retrieval [3,4]. Parameter-efficient fine-tuning (LoRA/QLoRA) can make adaptation feasible under limited compute [6,11]. For Turkish, resources remain more limited than for English, but Turkish legal encoders and prior-case retrieval studies have started to appear [16, 27]. Public corpora and datasets are also emerging [15], although stable expert-labeled precedent links are still uncommon.

2.4 Large Language Models in the Pipeline

LLMs are increasingly used for summarization and extraction in modern IR pipelines [9]. In this paper, we restrict LLM usage to preprocessing steps (PDF transcription/structuring and query-surrogate summary generation). At retrieval time, we do *not* call an LLM; scoring is deterministic via BM25 and cosine similarity over fixed embeddings.

Gap and motivation. Much of the existing literature either assumes clean, machine-readable case text or relies on curated relevance labels/citation graphs. For Turkish case law in scanned PDF form, both assumptions can be fragile. Our goal is therefore modest: we provide a VLM-based preprocessing pipeline and a reproducible diagnostic that make it easier to compare section choices and simple fusion strategies before investing in expert labeling.

```
{
    "file_name": "2022_551.pdf",
    "meta_data": {
        "esas_no": "2021/1168",
        "karar_no": "2024/1919",
        "court_name": "İSTANBUL BÖLGE ADLİYE MAHKEMESİ"
    },
    "raw_text_preview": "T.C.\n\nİSTANBUL\n\nBÖLGE ADLİYE MAHKEMESİ\n\n12. HUKUK DAİRESİ\n\nDOSYA NO: 2022/551\n\nKARAR NO:
2024/1919\n\nTÜRKMİLLETİADINA\n\nİSTİNAFKARARI\n\nİNCELENEN KARARIN\n\nMAHKEMESİ: BAKIRKÖY 2. ASLİYE TİCARET MAHKEMESİ\n"
}
```


Fig. 1. Regex-based segmentation failure on a scanned decision.

3 Method

We next describe corpus construction, section extraction, retrieval models, and the proxy evaluation protocol. LLMs are used only for corpus construction and query-surrogate generation; retrieval itself uses only stored text, indices, and embeddings.

3.1 Data preprocessing

We convert image-based PDFs into structured, retrieval-ready text segments.

Corpus. We curated 1,000 Turkish judicial decisions in the domain of *Social Media Law*. Documents are primarily image-based PDFs with heterogeneous scan quality and layout artefacts (e.g., stamps, headers/footers, skew, and page-break fragmentation), which makes conventional OCR followed by rule-based section parsing brittle.

Why regex segmentation is insufficient. A simple baseline is OCR followed by regex patterns that try to split decisions into rhetorical roles. In practice, layout noise and court-specific templates often cause boundary shifts, missing headers, and merged sections, yielding incomplete or unstructured outputs. Figure 1 shows one typical failure case.

In such cases, the pipeline may not reliably separate `Facts`, `Reasoning`, and `Verdict`, which can introduce noise into downstream indices.

VLM-based Multimodal Extraction and Refinement. For our multimodal extraction pipeline, we utilized `Amazon Nova-2-Lite-V1`, selected for its cost-efficiency—specifically its integrated visual processing without additional surcharges—and its competitive performance with a 1M token context window, comparable to Gemini-class models. The pipeline transforms legal document images into structured text, segmenting each decision into three core components (`Facts`, `Reasoning`, and `Verdict`) and generating dual-purpose summaries (Figure 2). To ensure data integrity, we implemented custom validation algorithms to detect and discard records that were incomplete or truncated due to token limits. For these cases, we utilized a refined inference configuration with `max_tokens` set to 3,000 and a `repetition_penalty` of 1.2 during the re-processing stage. Furthermore, to mitigate "infinite loop" issues observed during legal statute extraction, system prompts were strictly constrained with instructions to mention each law article only once.

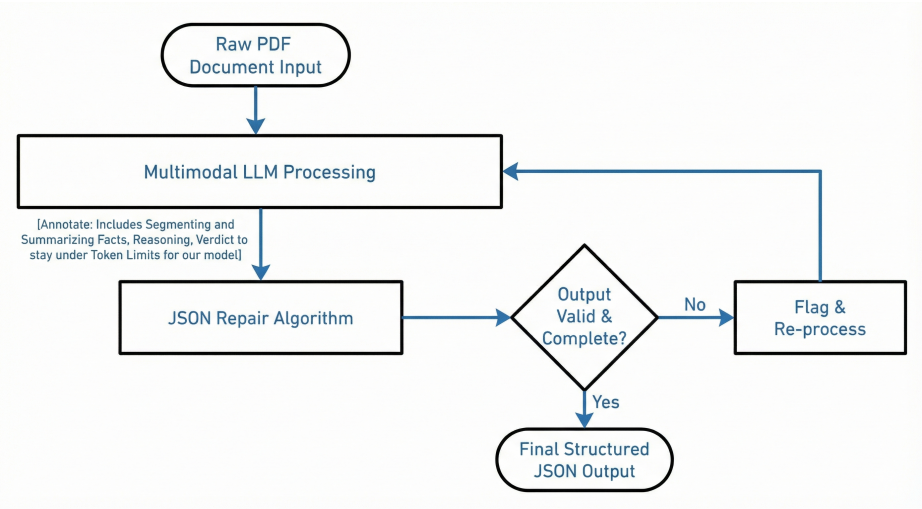

Fig. 2. Preprocessing pipeline for scanned PDFs.


```
{
    "meta_data": {
        "file_name": "2019_1462.pdf",
        "court_name": "İstanbul Bölge Adliye Mahkemesi 16. Hukuk Dairesi",
        "esas_no": "2019/1462",
        "karar_no": "2019/1457",
        "case_subject": "İhtiyati Tedbir",
        "dava_tarihi": "06/03/2019",
        "karar_tarihi": "05/07/2019"
    },
    "rrl_segments": {
        "facts_text": "İddialar, vekilin müvekkilinin ticari markalarını kullanmak için internet sitelerinde reklam ve tanıtım faaliyetleri yapmasıdır. Mahkeme, istinaf dilekçesi ve ek
tedbir taleplerini değerlendirmiştir.",
        "reasoning_text": "Mahkeme, Hükim 353 ve 356 maddeleri çerçevesinde, istinaf ve tedbir taleplerini reddetmiştir. İstinaf istemi reddedilmiş ve dava süreci devam etmiştir.",
        "verdict_text": "Dair, dosya üzerinde yapılan inceleme sonucu 05/07/2019 tarihinde oy birliği ile kesin olarak karar verildi."
    },
    "structural_features": {
        "mentioned_laws": [
            "Hukuk 353",
            "Hukuk 356",
            "Hukuk 353/1-b-1"
        ]
    },
    "summary_for_human": "Mahkeme, vekilin istinaf ve tedbir taleplerini reddetmiştir. İstinaf istemi reddedilmiş ve dava süreci devam etmiştir.",
    "summary_for_model": "Dava, ticari marka ihlali iddiası üzerine açılmıştır. Mahkeme, istinaf ve tedbir taleplerini reddetmiş ve dava süreci devam etmiştir."
}
```


Fig. 3. Example structured JSON record used for indexing and evaluation.

Structured JSON Output. The resulting dataset is stored in a structured JSON format, where each record comprises four key modules: (i) **Metadata**, containing case-specific details such as file name, court, case and decision dates, docket numbers, and subject matter; (ii) **Rhetorical Role Labeling (RRL)**, consisting of the role-labeled text for `Facts`, `Reasoning`, and `Verdict`; (iii) **Structural Features**, capturing extracted statutes (e.g., TCK, FSEK) used as lexical anchors; and (iv) **Summaries**, including a `summary_for_human` for general readability and a `summary_for_model` designed for self-retrieval and proxy evaluation tasks (Figure 3).

Summary generation and overlap. `summary_for_model` is generated using an LLM from the extracted content of the same decision and is used only as a query surrogate for the self-alignment diagnostic (Section 3.2). Because the query is derived from the target decision, lexical overlap is expected; we therefore treat

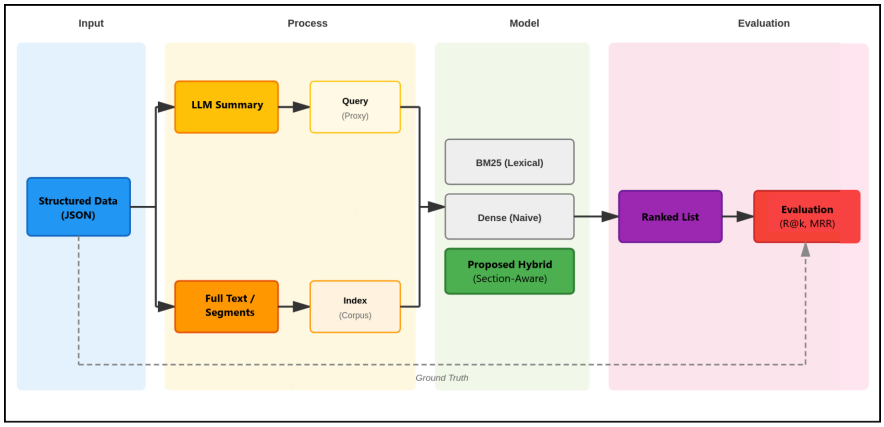

Fig. 4. Self-alignment diagnostic used for section-aware retrieval evaluation.

full-document BM25 as a reference baseline for this proxy and focus on relative section effects and fusion behavior.

Validation and reprocessing. We validate schema completeness and re-run extraction for invalid outputs. About 3% of cases required reprocessing due to segmentation shifts in complex layouts. To reduce repetitive generations, we use deterministic decoding (`temperature=0`) with a mild repetition penalty.

Extraction quality and downstream impact. This relatively small reprocessing rate should not be interpreted as evidence that extraction noise has no effect on retrieval performance. In the current study, we did not separately quantify the downstream impact of such errors, since the issue was identified earlier during segment extraction and completeness checking, before model training. We therefore treat remaining extraction imperfections as part of the practical preprocessing noise inherent to scanned legal documents. A more direct analysis of how extraction quality influences downstream retrieval performance is left for future work.

3.2 Section-aware self-alignment diagnostic

Because no expert-labeled Turkish precedent benchmark is available for our target domain, we use a controlled known-item proxy.

No Turkish benchmark provides expert-labeled precedent links for our target domain. We therefore use a reproducible *self-alignment* known-item proxy to compare section choices and lexical–dense fusion under identical noise and length constraints (Figure 4). This proxy is best read as a diagnostic rather than true precedent relevance; we include full-document BM25 as a reference baseline.

Proxy task. Let $\mathcal{D} = \{d_i\}_{i=1}^{N}$ denote the corpus of $N = 1{,}000$ decisions. For each d_i, we query with $q_i = \texttt{summary\_for\_model}(d_i)$ and treat d_i as the single relevant target. A retriever produces a ranked list over $\mathcal{D}$; success is measured by the rank of d_i.

Metrics. With one relevant target per query, we report Recall@1/5/10 and MRR@10.

3.3 Retrieval models

We evaluate three families of first-stage retrievers: sparse (lexical), dense (semantic), and simple hybrids. None of these models call an LLM at retrieval time.

BM25 (sparse). We use BM25 [18] as the lexical baseline. To isolate section effects (RQ1), we build separate indices over `facts`, `laws`, `facts+laws`, `facts+laws+subject`, `facts+laws+subject+verdict`, `reasoning`, and `full`.

Dense bi-encoders. Dense retrievers encode queries and candidates into a shared vector space and rank by cosine similarity [17]. We evaluate multilingual encoders (MiniLM, MPNet) [21, 25] and legal/multilingual backbones (BERTurk-Legal, XLM-R) [4, 16]. For XLM-R, we compare full fine-tuning and parameter-efficient variants (LoRA and 4-bit QLoRA) [6, 11]. Weak supervision uses $q = \texttt{summary\_for\_model}(d)$ with the positive segment $d^+ = \texttt{reasoning\_text}(d)$. Hard negatives are mined with BM25 within the training split to encourage discrimination among lexically similar decisions.

Hybrid fusion. We combine lexical and dense signals via late fusion:

$$s_{\text{hyb}}(q, d) = \alpha \cdot \tilde{s}_{\text{BM25}}(q, d) + (1 - \alpha) \cdot \tilde{s}_{\text{dense}}(q, d), \tag{1}$$

with per-query min–max normalization

$$\tilde{s}(q, d) = \frac{s(q, d) - \min_{d'} s(q, d')}{\max_{d'} s(q, d') - \min_{d'} s(q, d') + \epsilon}. \tag{2}$$

To reduce semantic dilution under fixed token budgets (RQ2) and to combine complementary evidence (RQ3), we use a simple section-aware hybrid that fuses BM25 over `facts+laws` and `reasoning` with dense similarity over `reasoning`:

$$s_{\text{sec-hyb}}(q, d) = \beta \cdot \tilde{s}_{\text{BM25}}^{(\text{facts+laws})}(q, d) + \gamma \cdot \tilde{s}_{\text{BM25}}^{(\text{reason})}(q, d) + (1 - \beta - \gamma) \cdot \tilde{s}_{\text{dense}}^{(\text{reason})}(q, d), \tag{3}$$

where $\beta, \gamma \geq 0$ and $\beta + \gamma \leq 1$. As an alternative that avoids score calibration, we also evaluate Reciprocal Rank Fusion (RRF) [5]:

$$\text{RRF}(d) = \sum_{m \in \mathcal{M}} \frac{1}{k + \text{rank}_m(d)}. \tag{4}$$

Cross-encoder reranking is included only as an ablation; in our runs, gains were small under our length constraints, so we do not use it in the main pipeline.

3.4 Experimental settings

We summarize the main experimental choices below for reproducibility.

Splits. We use the full corpus of $N = 1{,}000$ structured decisions and a fixed 70/10/20 train/validation/test split (700/100/200) with decision-level separation. Dense training (including BM25-based negative mining) uses only the training split; hyperparameter and fusion-weight selection uses validation; we report metrics on the held-out test set. To control leakage, test decisions are excluded from training both as positives and as mined negatives.

BM25 preprocessing. We build BM25 indices for each segment configuration. Text processing includes Unicode and whitespace normalization, lowercasing, and lightweight stopword filtering tailored to Turkish and common legal boilerplate. To improve citation matching, statutory references (e.g., "TTK 55/1-a-1") are normalized into single tokens (e.g., "ttk_55_1_a_1") while preserving digits. For selected segment configurations, contextual fields (e.g., mentioned laws or subject) may be upweighted via token repetition. Unless otherwise stated, BM25 uses $k_1 = 1.2$ and $b = 0.75$.

Dense training. Dense retrievers are trained as bi-encoders with cosine similarity (dot product of L2-normalized embeddings). We fine-tune using TripletLoss (margin 0.2). For each query, a hard negative is sampled from BM25's top-50 results (excluding the positive), drawn from ranks 6–30 to avoid trivial negatives. Inputs are truncated to a maximum sequence length of 256 tokens. We run an epoch sweep over $\{1, \ldots, 7\}$ with fixed random seed 42, batch size 8, learning rate $2 \cdot 10^{-5}$, and mixed-precision training (AMP). The best checkpoint is selected based on validation MRR@10 of the section-aware hybrid system.

Fusion and reproducibility. Score-level fusion uses per-query min–max normalization (Eq. 2). Unless otherwise stated, we use $\alpha = 0.5$ for the basic hybrid (Eq. 1) and $(\beta, \gamma) = (0.45, 0.20)$ for the section-aware hybrid (Eq. 3). These weights were selected on the validation split of the current dataset and should therefore be interpreted as task-specific settings rather than generally optimal values. We intentionally kept the fusion design simple and interpretable, since the goal of this study is not to propose a heavily tuned or learned fusion mechanism, but to examine section utility and hybrid behavior in a controlled, low-resource diagnostic setting. RRF uses $k = 60$ (Eq. 4) and does not require score calibration. LLM usage is limited to corpus construction and summary generation; retrieval-time scoring is deterministic given fixed indices and embeddings.

4 Results

We report the main quantitative findings under the self-alignment proxy below. As noted above, these results are best read as a diagnostic for recovering the source decision, not as expert-labeled precedent relevance.

Table 1. BM25 across indexed segments under the self-alignment proxy (query=`summary_for_model`). Full-document BM25 is included as a reference baseline for this proxy.

BM25 Index Segment	R@1	R@5	R@10
Facts	0.542	0.687	0.715
Laws	0.111	0.192	0.236
Facts + Laws	0.629	0.757	0.799
Facts + Laws + Subject	0.645	0.778	0.813
Facts + Laws + Subject + Verdict	0.728	0.851	0.894
Reasoning	0.744	0.833	0.865
Full Document	**0.913**	**0.963**	**0.974**

BM25 segment effects (RQ1). Table 1 reports BM25 across section-wise indices under the self-alignment proxy. We include full-document BM25 as a reference baseline for this proxy (R@10=0.974). Among single sections, `Reasoning` gives the highest scores in our setup (R@1=0.744, R@10=0.865). Indexing `Facts + Laws + Subject + Verdict` also helps at early ranks (R@1=0.728, R@10=0.894), bringing performance closer to the full-document reference. In contrast, `Laws`-only indexing is weak, and `Facts`-only remains below `Reasoning`. Overall, this suggests that much of the signal for this proxy task comes from legal reasoning, while contextual fields (subject and verdict) can provide additional lexical cues.

Dense vs. hybrid retrieval (RQ2–RQ3). In this weakly supervised known-item setting, dense-only retrievers are generally weaker than BM25, suggesting that lexical overlap and statutory cues remain highly informative under this proxy. However, combining dense retrieval with BM25 consistently improves performance across models. Table 2 shows representative backbones and indicates that section-aware fusion usually yields the strongest early-rank results. In particular, MPNet with section-aware fusion achieves the best overall R@1 and MRR@10 scores (R@1=0.780, MRR@10=0.810), while the highest R@5 score is 0.870, reached by both XLM-R (FULL) and XLM-R (LoRA) with section-aware fusion, and XLM-R (LoRA) attains the highest R@10 (0.900). Overall, these results suggest that reasoning-focused semantic matching is most effective when combined with explicit lexical anchoring, rather than compressed into a single naive concatenated representation. This brings hybrid performance closer to the full-document BM25 reference while preserving section-level scoring and interpretability.

Training segment pairs. When training dense encoders, using `summary_for_model` as queries and `Reasoning` as positives gave the best dense-only results in our runs. Training on full-text or multi-section concatenations tended to underperform, which is consistent with signal dilution under fixed token budgets (table omitted for space).

Table 2. Dense vs. hybrid retrieval results on the test split (seed 42). Best epochs are selected based on validation MRR@10 of the section-aware hybrid system.

Model / System	**R@1**	**R@5**	**R@10**	**MRR@10**
MiniLM Dense	0.390	0.560	0.620	0.458
MiniLM Hybrid (naive)	0.640	0.780	0.850	0.707
MiniLM Hybrid (section-aware)	0.730	0.860	0.880	0.788
MPNet Dense	0.490	0.620	0.650	0.544
MPNet Hybrid (naive)	0.670	0.800	0.860	0.728
MPNet Hybrid (section-aware)	**0.780**	0.860	0.890	**0.810**
BERTurk-Legal Dense	0.430	0.520	0.590	0.468
BERTurk-Legal Hybrid (naive)	0.640	0.830	0.850	0.717
BERTurk-Legal Hybrid (section-aware)	0.710	0.860	0.880	0.775
XLM-R (FULL) Dense	0.380	0.460	0.460	0.409
XLM-R (FULL) Hybrid (naive)	0.620	0.770	0.820	0.683
XLM-R (FULL) Hybrid (section-aware)	0.690	**0.870**	0.890	0.765
XLM-R (LoRA) Dense	0.380	0.460	0.470	0.413
XLM-R (LoRA) Hybrid (naive)	0.620	0.770	0.820	0.683
XLM-R (LoRA) Hybrid (section-aware)	0.700	**0.870**	**0.900**	0.772
XLM-R (QLoRA, 4bit) Dense	0.415	0.470	0.510	0.442
XLM-R (QLoRA, 4bit) Hybrid (naive)	0.620	0.765	0.810	0.682
XLM-R (QLoRA, 4bit) Hybrid (section-aware)	0.745	0.860	0.895	0.796

5 Discussion

We interpret the proxy results below and highlight their main limitations.

Under the self-alignment proxy, section choice has a clear effect. `Reasoning` tends to be more informative than `Facts` or `Laws` alone, and dense-only models trained with weak supervision do not match lexical BM25 in this setup. However, simple late fusion of lexical and dense signals often improves early ranks, especially when dense similarity focuses on `Reasoning`.

Implications for section choice. Our emphasis on the `Reasoning` section was not chosen arbitrarily. It was also informed by discussions with legal practitioners, who indicated that the reasoning part of a decision is often the most critical component when assessing precedent value. In this sense, our findings do not merely suggest a data-driven pattern, but also provide empirical support for an intuition already present in legal practice. Under our proxy setting, the experimental results reinforce the view that `Reasoning` carries a particularly strong retrieval signal compared with `Facts` or isolated law mentions alone.

Why naive concatenation can hurt. Concatenating heterogeneous sections into a single encoder input can underperform section-wise scoring. Long inputs trigger truncation, boilerplate templates may dominate the representation, and frequent citations can introduce false positives. Section-aware fusion can reduce these issues by anchoring with BM25 over `facts+laws` and focusing dense similarity on `reasoning`.

Limitations of the proxy. The proxy does not measure true precedent relevance: queries are summary surrogates derived from the same decision, and full-document BM25 is naturally advantaged (reported here as a reference baseline). Results may vary with summary style, OCR/segmentation noise, and the small training regime (N=1,000; 700 train), which can limit dense adaptation under domain/style shift. In addition, because the query is derived from the target decision itself, the setup should be interpreted as a controlled known-item diagnostic with expected lexical overlap rather than a realistic substitute for expert-judged precedent search. A necessary next step is expert pooling with graded relevance and standard ranked metrics [16].

6 Conclusion

In this case study, we explored section-aware retrieval for Turkish legal decisions using a reproducible self-alignment proxy. Within this setup, we observed that the `Reasoning` section often carries more retrieval signal than `Facts` or `Laws` alone, and that simple lexical–dense fusion can help under fixed token budgets. We emphasize that our evaluation is a diagnostic rather than expert-labeled precedent retrieval. The corpus was intentionally limited to social-media-related disputes so that section effects and fusion behavior could be studied in a more controlled setting, with less heterogeneity in legal terminology, case structure, and retrieval signals. While this limits direct generalization to other legal domains, it also makes the present study methodologically cleaner as an initial domain-specific diagnostic. Future work will add expert-labeled evaluation with pooling and graded relevance (and/or practitioner user studies), extend the corpus gradually to additional legal domains, explore paragraph-level aggregation baselines, and revisit cross-encoder reranking under tighter evidence selection.

References

1. Althammer, S., Hofstätter, S., Sertkan, M., Verberne, S., Hanbury, A.: Parm: A paragraph aggregation retrieval model for dense document-to-document retrieval. In: European Conference on Information Retrieval. pp. 19–34. Springer (2022)
2. Askari, A., Verberne, S., Abolghasemi, A., Kraaij, W., Pasi, G.: Retrieval for extremely long queries and documents with rprs: a highly efficient and effective transformer-based re-ranker. ACM Transactions on Information Systems **42**(5), 1–32 (2024)

3. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-bert: The muppets straight out of law school. arXiv preprint arXiv:2010.02559 (2020)
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 8440–8451 (2020)
5. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 758–759 (2009)
6. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023)
7. Feng, Y., Li, C., Ng, V.: Legal case retrieval: A survey of the state of the art. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6472–6485 (2024)
8. Formal, T., Piwowarski, B., Clinchant, S.: Splade: Sparse lexical and expansion model for first stage ranking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2288–2292 (2021)
9. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 **2**(1) (2023)
10. Goebel, R., Kano, Y., Kim, M.Y., Rabelo, J., Satoh, K., Yoshioka, M.: Overview and discussion of the competition on legal information, extraction/entailment (coliee) 2023. The Review of Socionetwork Strategies **18**(1), 27–47 (2024)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
12. Karpukhin, V., Oguz, B., Min, S., Lewis, P.S., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: EMNLP (1). pp. 6769–6781 (2020)
13. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 39–48 (2020)
14. Nigam, S.K., Dubey, T., Shallum, N., Bhattacharya, A.: Segment first, retrieve better: Realistic legal search via rhetorical role-based queries. arXiv preprint arXiv:2508.00679 (2025). https://doi.org/10.48550/arXiv.2508.00679, https://arxiv.org/abs/2508.00679
15. Öztürk, C.E., Köksal, Ö., Özçelik, Ş.B., Koç, A.: Prior case retrieval for the court of cassation of turkey. In: 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). pp. 539–544. IEEE (2022)
16. Öztürk, C.E., ÖzçelıK, Ş.B., Koç, A.: A transformer-based prior legal case retrieval method. In: 2023 31st Signal Processing and Communications Applications Conference (SIU). pp. 1–4. IEEE (2023)
17. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
18. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)

19. Sansone, C., Sperlí, G.: Legal information retrieval systems: state-of-the-art and open issues. Information Systems **106**, 101967 (2022)
20. Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., Ma, S.: BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20). pp. 3501–3507 (2020). https://doi.org/10.24963/ijcai.2020/484, https://www.ijcai.org/proceedings/2020/484
21. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems **33**, 16857–16867 (2020)
22. Su, W., Ai, Q., Wu, Y., Ma, Y., Li, H., Liu, Y., Wu, Z., Zhang, M.: Caseformer: Pre-training for legal case retrieval based on inter-case distinctions. arXiv preprint arXiv:2311.00333 (2023). https://doi.org/10.48550/arXiv.2311.00333, https://arxiv.org/abs/2311.00333
23. Tang, Y., Qiu, R., Yin, H., Li, X., Huang, Z.: CaseLink: Inductive graph learning for legal case retrieval. arXiv preprint arXiv:2403.17780 (2024). https://doi.org/10.48550/arXiv.2403.17780, https://arxiv.org/abs/2403.17780
24. Voorhees, E.M.: Report on trec-9. In: ACM SIGIR Forum. vol. 34, pp. 1–8. ACM New York, NY, USA (2000)
25. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems **33**, 5776–5788 (2020)
26. Yılmaz, K.E., Arslan, A., Yilmazel, O.: Turkish text retrieval experiments using Lemur toolkit. arXiv preprint arXiv:1405.1740 (2014), https://arxiv.org/abs/1405.1740
27. Zeidi, F., Amasyali, M.F., Erol, Ç.: Legalturk optimized bert for multi-label text classification and ner. arXiv preprint arXiv:2407.00648 (2024)
28. Zhou, Y., Huang, H., Wu, Z.: Boosting legal case retrieval by query content selection with large language models. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. pp. 176–184 (2023)