

Hybrid Retrieval with Legal Signal Boosting for Indonesia's Job Creation Law and Implementing Regulations

Rheno Septianto^{*1}, Yan Rianto¹

¹ Informatics Master's Program, Faculty of Information Technology, Universitas Nusa Mandiri, Indonesia

Article Info

Keywords:

Hybrid retrieval;
Legal signal boosting;
Retrieval-augmented generation;
Legal information retrieval;
Indonesian Job Creation Law

* Corresponding author

E-mail:

14240011@nusamandiri.ac.id

Abstract

This study evaluated legal signal boosting within a hybrid retrieval system for Indonesia's Job Creation Law and fifteen implementing government regulations. The knowledge base contained 2,914 law chunks and 2,532 regulation chunks, represented with Multilingual E5. BM25 and semantic rankings were combined through reciprocal rank fusion, with an additional ranking derived from document identity, article references, explanatory sections, and legal phrases. Five retrieval methods were compared on fifty questions per scenario. A separate controlled comparison examined the contribution of legal signals. In the controlled comparison, legal signal boosting increased Hit@10 from 94% to 100% and mean reciprocal rank from 0.707524 to 0.850190 in the law corpus. In the regulation corpus, Hit@10 increased from 98% to 100% and mean reciprocal rank from 0.933333 to 0.966667. Dual-corpus evidence coverage increased from 78% to 92%. The selected configuration did not lead on every metric: TF-IDF obtained higher cross-corpus MAP@10. The prototype also supported corpus routing and source-linked answers. A separate quotation check passed 28 of 30 answers on a reused development set. The retrieval comparisons were descriptive and used internally constructed questions and relevance labels; cross-corpus configuration selection reused its evaluation set. The findings supported legal signals as a useful addition to hybrid regulatory retrieval under the tested conditions, while statistical significance and performance on external questions remained unverified.

1. Introduction

The Job Creation framework distributes provisions across a law and its implementing regulations. Similar terms recur in articles, annexes, and explanatory sections. Semantic similarity alone may therefore return text from a different provision than the one requested. This study examined whether signals identifying the document, article, or section could improve hybrid retrieval within that collection.

Retrieval-augmented generation connects a generative model to external evidence [1]. Surveys describe the interaction between retrieval, context construction, and generation as a central design concern [2] [3]. Dense representations can improve semantic matching, but they do not automatically resolve exact legal references. Multilingual E5 provides a multilingual representation suitable for Indonesian text [4], while contrastive dense retrieval offers a complementary alternative to lexical matching [5]. The FAISS library supports efficient vector search and explicit links between indexed vectors and document units [6]. These components are useful foundations, but their integration must account for the structure and coverage of the source collection. A KINETIK study also compared retriever and generator combinations for concept-map assessment [7], illustrating the need to evaluate these components separately.

Legal question-answering research has addressed source attribution, retrieval, and context selection in several ways. Attributed answering links responses to statutory preconditions [8], while multi-phase retrieval and free-format question answering connect questions with supporting legal documents [9] [10]. Hierarchical knowledge integration, case-based retrieval, and parameter-adaptive retrieval offer related approaches to selecting legal context [11] [12] [13]. Indonesian legal consultation research has also combined ensemble retrieval with keyword boosting [14]. The present study evaluated document identity, article references, explanatory sections, and legal phrases across two structurally

different regulatory corpora. Its contribution lies in adapting and evaluating these signals within the selected collection, including a controlled component comparison.

Citation prediction in Dutch case law has been studied using textual and metadata features [15]. SDD-LawLLM combines synthetic-data fine-tuning with retrieval augmentation for legal question answering [16]. Models of law search also represent relationships within the legal corpus [17]. LamBERTa investigates law-article retrieval through adapted language representations [18].

Lawformer addresses long Chinese legal documents [19], while transformer-based approaches have been evaluated for case-law retrieval and statutory entailment [20]. Multi-stage Japanese retrieval [21] and CECRA's structured-evidence reranking [22] explore other approaches to matching legal documents. These tasks motivate attention to structure and evidence, but their scores are not directly comparable with retrieval from Indonesian regulations.

Regulatory RAG has been studied for EU electricity rules [23] and multilingual legal assistance in Kazakhstan [24]. LARIN adapts retrieval triggering, query construction, and evidence fusion for legal tasks [25]; graph-enhanced retrieval has also been evaluated for information-security compliance questions [26]. The present experiments focus on a fixed regulatory collection and a fixed generator, rather than reproducing those systems.

Evaluation must distinguish finding evidence from producing a supported answer. BEIR illustrates variation across retrieval tasks and domains [27], while LegalBench-RAG focuses specifically on legal retrieval [28]. LegalBench and LawBench assess broader legal reasoning or knowledge capabilities [29] [30]. Groundedness evaluation examines the relationship between legal answers and their sources [31]; RAGAs and ARES provide complementary automated evaluation frameworks [32] [33]. Their existence does not justify substituting retrieval hit rate for legal correctness. Empirical legal-AI studies distinguish factual accuracy from support by cited sources and document remaining hallucinations in legal research tools [34] [35].

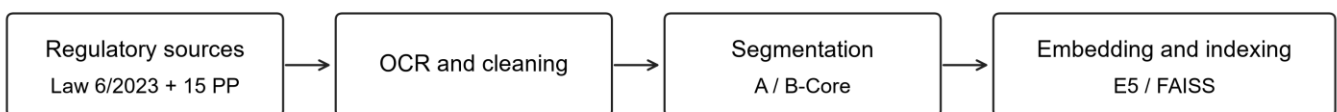
Related computing studies discuss preprocessing and classification of Indonesian text [36] [37], software performance and feature engineering [38] [39], and the evaluation of user-facing systems [40] [41] [42]. They provide methodological background. Evidence of legal retrieval performance in this study comes from the corpus-specific experiments reported below.

This study evaluated legal signal boosting in a regulatory search system that combined lexical and semantic retrieval. Five retrieval configurations were compared, followed by a controlled comparison with and without legal signals. The evaluation covered retrieval within each corpus, evidence retrieval across corpora, query routing, and the quotation and citation of retrieved provisions.

2. Research Method

This study used a comparative experimental design to evaluate legal signal boosting in hybrid retrieval. The knowledge base comprised Law Number 6 of 2023 and fifteen implementing government regulations. Five retrieval configurations were evaluated, and a separate controlled comparison examined the contribution of legal signals. Figure 1 shows corpus preparation and question processing. Prepared indexes supply hybrid retrieval; the user question is the input to routing.

Corpus preparation



Question processing

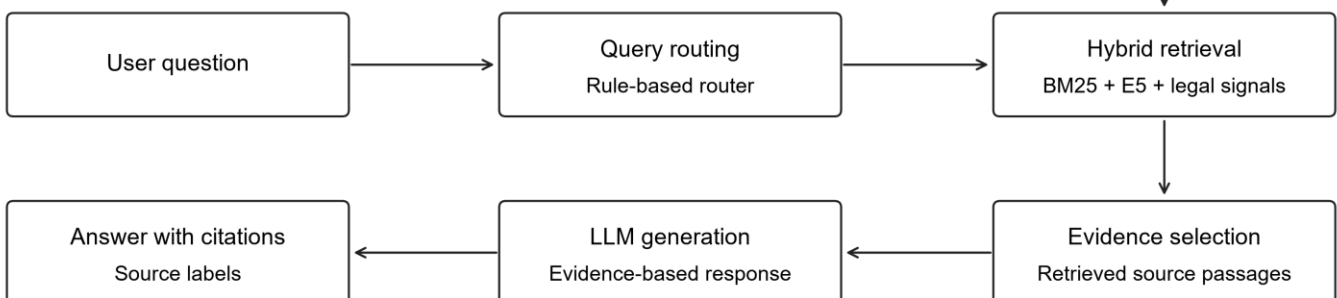


Figure 1. Regulatory Search Workflow

2.1. Source Collection and Corpus Preparation

Corpus A contained Law Number 6 of 2023 [43]. Corpus B-Core contained fifteen government regulations (Peraturan Pemerintah, PP) from 2021: Numbers 5, 6, 7, 18, 19, 21, 22, 23, 26, 27, 28, 29, 35, 36, and 37 [44]–[58]. The original source files were retained unchanged. An inventory recorded document number, year, page count, extraction method, and inclusion status. The collection was a fixed historical snapshot of these document versions; subsequent amendments and replacement regulations were not consolidated. The experiments assessed retrieval from that snapshot rather than the completeness or current legal applicability of the collection.

Source qualification preceded retrieval. Pages 350–739 of Government Regulation Number 5 of 2021 were excluded because the annex contained tables extending across pages and required extraction that preserved table structure. Government Regulation Number 34 of 2021 remained in the archive but was excluded from the final operational scope. These exclusions were recorded explicitly so that incomplete evidence would not be interpreted as proof that a provision did not exist. Work on legislative corpus construction similarly highlights the need to verify machine-readable text after extraction [59].

Text extraction and optical character recognition were followed by removal of recurring headers, footers, and disruptive word breaks. Clean text was retained per page. Corpus A used legal-structure-aware parent and child units; Corpus B-Core used one parent per page and character-based child chunks with a target length of 1,800 characters and an overlap of 220 characters. This distinction retained available legal structure without assuming that all government regulations had equally reliable article metadata. Chunk granularity is a meaningful retrieval design choice [60]. Table 1 summarizes the resulting knowledge base.

Table 1. Validated knowledge-base inventory.

Attribute	Corpus A	Corpus B-Core
Documents	Law 6/2023	15 government regulations
Included pages	1,127	2,486
Parent units	2,803	2,486
Child chunks	2,914	2,532
Embedding vectors	2,914	2,532
FAISS mappings	2,914	2,532

Each child unit retained the source document, page, parent identifier, chunk identifier, and text, together with article metadata where available. Before retrieval, validation checked that text was nonempty, required metadata were present, document exclusions were applied, and the numbers of chunks, embeddings, and indexed vectors agreed.

2.2. Retrieval and Legal Signal Fusion

Five configurations were compared: TF-IDF, BM25, Raw E5, Hybrid RRF, and Hybrid RRF with legal signal boosting. The hybrid configurations combined lexical and semantic rankings, while the boosted configuration also used legal references identified in the question and source metadata. The encoder, intfloat/multilingual-e5-base, produced normalized 768-dimensional vectors stored in FAISS IndexFlatIP. Dense scores in the comparison were calculated by exact search against the fixed embedding matrix. All embedding-based configurations used the same encoder.

Hybrid RRF combined the BM25 and Raw E5 rankings through reciprocal rank fusion [61], as shown in Equation 1. Here, d is a retrieved chunk, m identifies a ranking list, $r_m(d)$ is the rank of d in list m , and w_m is the list weight. The rank constant was fixed at 60.

$$S_{RRF}(d) = \sum_m \frac{w_m}{60 + r_m(d)} \quad (1)$$

The generic legal-signal route used document identity, article references, explanatory sections, and legal phrases found in the question, chunk text, and source metadata. Matching scores were added per chunk: 12.0 for an explicitly named document, 10.0 for an exact article-metadata match, or 6.0 for an article reference in the text when the metadata did not match; a matching explanatory section added 5.0. Each matching legal phrase contributed $3.0 + 0.25 \times \min(\text{number of phrase words}, 5)$. Only chunks with positive legal scores entered the additional candidate list. They were ordered by legal score, then baseline RRF score, and finally chunk index, and the top fifty were used for fusion. A chunk absent from this list could still be retrieved through BM25 or E5.

For B-Core and cross-corpus retrieval, the BM25, E5, and generic legal-signal lists each had weight 1.0 in Equation 1. Corpus A instead used its structure-specific exact route with weight 2.25, alongside BM25 and E5 with weight 1.0, followed by additive metadata signals of 0.025–0.06 per match. Matching metadata signals were summed, so this interval describes individual contributions rather than a cap on the total boost. The controlled Corpus A baseline retained the same lexical implementation and disabled both the exact route and metadata additions. This differs from

the generic lexical baseline in the five-method comparison. Parameters were fixed after development, and ranking did not read relevance labels. Recent work on structured documents also examines retrieval choices for text and tables [62]. Metadata additions can materially affect ranking rather than merely break ties. Their magnitudes and the 2.25 route weight were not tested in a sensitivity analysis, so robustness to other weights remains unverified.

2.3. Query Routing and Answer Generation

The rule-based router used document references and question signals to select A_ONLY, B_ONLY, or CROSS_A_B. Ambiguous questions could use AUTO_ALL as a fallback. Single-corpus modes searched the selected corpus; CROSS_A_B searched both corpora and retained source identities when assembling evidence. Out-of-scope intent detection is a related issue in knowledge-based chatbots [63], but this study evaluated corpus routing rather than a separate out-of-scope classifier.

The Streamlit prototype displayed retrieved evidence with document, page, corpus, and available legal-structure metadata. General search used ten contexts labelled S1–S10. Explicit quotation requests received evidence selected for the requested provision. Improvements separated the main enactment from annexed text, resolved each reference separately, and selected the requested article, paragraph, item, or letter. If extraction did not preserve a recognizable paragraph marker, the system did not invent a new paragraph number. These changes affected context selection without modifying the frozen chunks, embeddings, or FAISS indexes.

The generator was held fixed during the answer evaluation. Comparisons with alternatives such as Mixtral [64] were outside the experiment.

Answers were generated by Qwen through Groq. The recorded API model identifier was qwen/qwen3.8-27b. Temperature was 0, context was limited to 950 characters per source, and the output limit was 1,100 tokens. The prompt requested quotations from the supplied evidence, supporting source labels, and a statement when a requested provision was unavailable. This experiment evaluated generation after evidence selection, rather than the complete interactive application.

2.4. Evaluation Design and Metrics

Each individual-corpus and cross-corpus retrieval scenario contained fifty internally constructed questions. Corpus A and B-Core each used thirty development questions and twenty holdout questions; configurations were fixed before the holdout evaluation. CROSS_A_B used a combined set for comparing retrieval methods and evidence allocation, so its results were exploratory configuration-selection evidence. An additional fifty-question mixed benchmark contained fifteen A_ONLY, fifteen B_ONLY, and twenty CROSS_A_B questions. Some mixed-benchmark questions overlapped with scenario sets, preventing its interpretation as an independent holdout.

Relevance was binary and defined at chunk level using expected documents, article or structural references where available, and important query terms. Corpus A development labels were based on an assessed candidate pool and were not exhaustive. B-Core and cross-corpus labels used document and term criteria. Unlabelled chunks were counted as nonrelevant. Labels remained fixed during ranking but were not assessed by a second annotator or an independent legal expert. All metrics therefore describe agreement with these internal labels. Because document and term criteria also informed the legal signals, label construction could favor the boosted method. Keeping labels outside the ranker prevented direct label access but did not remove this threat to evaluation validity.

Retrieval was evaluated using Hit@K, MRR, Precision@K, Recall@K, MAP@K, and nDCG@K, defined in Equations 2–7. Hit@K measures whether relevant evidence is retrieved within the first K positions, while MRR measures the rank of the first relevant result. Precision and recall measure retrieved relevance and coverage of the labelled relevant set. MAP and nDCG also account for result ordering. Evaluation used K = 1, 3, 5, and 10; the results below report K = 10.

$$Hit@K = \frac{1}{N} \sum_{i=1}^N I(r_i \leq K) \quad (2)$$

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i} \quad (3)$$

$$Precision_q@K = \frac{\sum_{i=1}^K rel_{q,i}}{K} \quad (4)$$

$$Recall_q@K = \frac{\sum_{i=1}^K rel_{q,i}}{|R_q|} \quad (5)$$

$$MAP@K = \frac{1}{N} \sum_{q=1}^N \frac{\sum_{i=1}^K (Precision_q@i \times rel_{q,i})}{|R_q|} \quad (6)$$

$$nDCG_q@K = \frac{DCG_q@K}{IDCG_q@K} \quad (7)$$

N denotes the number of questions, and r_i is the rank of the first relevant result for question i . $I(\cdot)$ is 1 when its condition holds and 0 otherwise; a question with no relevant result contributes 0 to MRR. R_q is the set of chunks labelled relevant to question q , and $rel_{q,i}$ is 1 for a relevant result at position i and 0 otherwise. MAP@K averages truncated average precision using the full labelled-set size $|R_q|$ as denominator. DCG discounts binary relevance by $\log_2(i + 1)$; IDCG is the corresponding ideal score up to K . Precision@K, Recall@K, and nDCG@K were averaged across questions.

Cross-corpus coverage required relevant evidence from both corpora. Local MRR was calculated separately within each corpus; mean local MRR averaged the two values. The original comparison evaluated allocations from A1+B9 to A9+B1 and selected a quota for each method by dual-corpus coverage, mean local MRR, minimum corpus coverage, and then quota balance. Table 5 reports these selected configurations. The controlled comparison in Table 6 and supplemental precision, recall, MAP, and nDCG instead fixed A6+B4 for all methods. Supplemental ordering alternated A and B for the first eight positions, followed by two A results. These recalculated rankings could differ where equal scores were resolved differently; they were not identical to the original quota-selection evaluation.

Controlled legal-signal comparisons activated or deactivated boosting while holding questions, labels, representations, and the base settings constant within each scenario. Their baseline values were reported separately from the earlier five-method comparison. Analyses were descriptive; no statistical-significance or weight-sensitivity conclusion was drawn. Routing used a balanced hundred-question set containing 34 A_ONLY, 33 B_ONLY, and 33 CROSS_A_B questions, plus a hundred-question stress set containing only implicit-signal cross-corpus questions.

Thirty factual quotation questions were selected purposively, ten per scenario. They covered twenty unique provisions from the annex to Law 6/2023 and Government Regulations 5, 7, 18, and 35 of 2021. Paired cross-corpus questions reused those provisions, producing forty quotation requests in total. The reference texts came from the annex to Law 6/2023 and the 2021 regulation texts, without consolidation of subsequent amendments. References were recorded from official PDFs before generation and were not supplied as answers to the model. Matching ignored case, punctuation, and whitespace but retained letters and digits; paraphrases and OCR corrections did not count as exact quotations. A citation passed when its source label was in the same answer section and the cited evidence contained the complete reference quotation. An answer passed only if it completed without truncation and every requested quotation and citation passed. Both quotations were required in cross-corpus answers. Service failures were retained in the denominator. The questions were reused after evidence-selection improvements, so the result measures development-set performance rather than performance on new questions.

3. Results and Discussion

3.1. Retrieval within Each Corpus

Table 2 compares retrieval on fifty questions per corpus. In Corpus A, legal signal boosting reached 100% Hit@10 and MRR 0.850190, compared with 94% and 0.736524 for Hybrid RRF. Raw E5 reached 82% Hit@10. Difficult questions concerned explanatory, transitional, and closing provisions with similar wording. In B-Core, the boosted configuration reached 100% Hit@10 and MRR 0.966667. These results supported separate retrieval configurations for the two corpora.

Table 2. Five-method retrieval comparison on fifty questions per corpus.

Method	A Hit@10 (%)	A MRR	B Hit@10 (%)	B MRR
TF-IDF	94	0.678238	98	0.851389
BM25	84	0.681667	98	0.897333
Raw E5	82	0.591111	100	0.902500
Hybrid RRF	94	0.736524	98	0.933333
Hybrid + legal signals	100	0.850190	100	0.966667

Both holdout sets retained 100% Hit@10 after the configuration was fixed (Table 3). Corpus A MRR increased from 0.838095 on the development questions to 0.868333 on the holdout questions. In B-Core, MRR decreased from 0.983333 to 0.941667. Each holdout contained twenty internally constructed questions. These results provide a limited check beyond the development questions; the aggregate scores do not establish whether one set was systematically easier.

Table 3. Fixed-configuration development and holdout results.

Corpus	Set	Questions	Hit@10 (%)	MRR
A	Development	30	100	0.838095
A	Holdout	20	100	0.868333
B-Core	Development	30	100	0.983333
B-Core	Holdout	20	100	0.941667

The supplemental metrics show that a successful first hit did not imply retrieval of all labelled evidence (Table 4). Corpus A reached 78.33% Recall@10. B-Core reached 78.40% Precision@10 but only 33.79% Recall@10: several relevant chunks were retrieved, while many others remained outside the ten-result limit. Recall therefore needs to be read alongside the size and construction of the relevance labels.

Table 4. Supplemental metrics for the legal-signal configuration.

Scenario	Precision@10 (%)	Recall@10 (%)	MAP@10	nDCG@10
A	35.60	78.33	0.650739	0.853248
B-Core	78.40	33.79	0.303839	0.882390
Cross-corpus	72.00	8.30	0.072196	0.744260

Figure 2 compares Hit@10 and MRR across Corpus A, B-Core, and the mixed benchmark. Legal signal boosting achieved 100% Hit@10 in all three sets. The mixed benchmark partly reused scenario questions, so its results describe performance on the combined set rather than an independent replication.

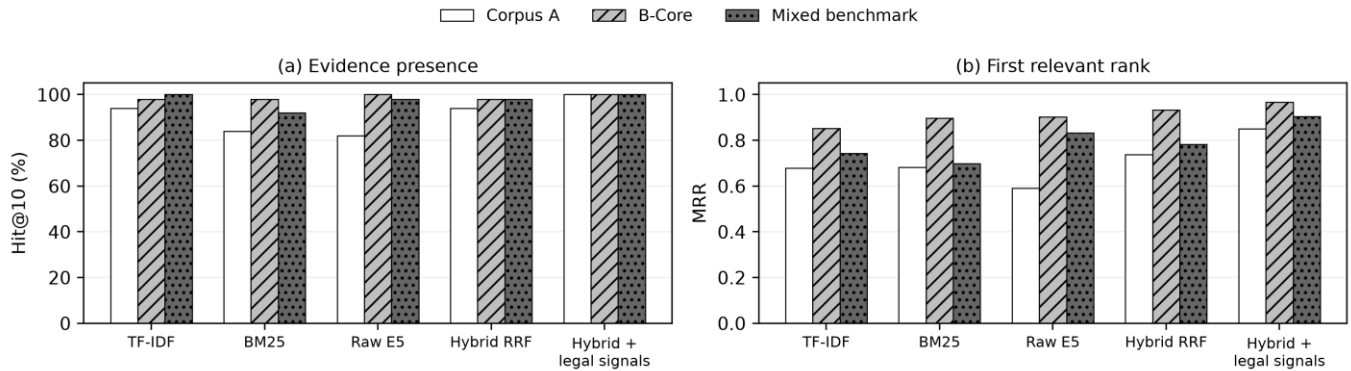


Figure 2. Hit@10 and MRR across Corpus A, B-Core, and the mixed benchmark.

3.2. Cross-Corpus Retrieval and Legal Signal Boosting

Raw E5 and the boosted configuration both retrieved relevant evidence from the two corpora for 92% of the cross-corpus questions (Table 5). The boosted configuration had higher mean local MRR, at 0.878167 compared with 0.837333, and was selected in the retrieval experiment. Four questions still lacked relevant evidence from both corpora. This set was also used for configuration selection, which limits conclusions about performance on new cross-corpus questions. In Table 5, the Hybrid RRF baseline used its selected A7+B3 allocation (80% coverage); Table 6 used the fixed A6+B4 allocation (78%). The two baseline values therefore describe different evaluation configurations. The 92% coverage is an exploratory configuration-selection result on these fifty questions, rather than an estimate for unseen cross-corpus questions.

Table 5. Cross-corpus evidence coverage and local ranking.

Method	A coverage (%)	B coverage (%)	Both (%)	Mean local MRR
TF-IDF	92	96	88	0.827333
BM25	86	84	78	0.674762
Raw E5	92	100	92	0.837333
Hybrid RRF	90	88	80	0.812095
Hybrid + legal signals	92	100	92	0.878167

Cross-corpus Recall@10 was 8.30% against the internal labels. TF-IDF obtained MAP@10 0.074964, above the boosted configuration's 0.072196. MAP accounts for precision across labelled relevant results, whereas dual-corpus coverage requires evidence from both corpora and mean local MRR concerns the first relevant result within each corpus. The different rankings across metrics therefore reflect different evaluation criteria. These aggregate results do not establish which query features caused TF-IDF's MAP advantage. The broad relevance sets and six/four evidence allocation limit interpretation of recall as legal completeness. On the mixed benchmark, boosting reached 100% Hit@10 and MRR 0.903000; this set was partially overlapping.

Table 6 reports the separate controlled comparison with and without legal signals. Corpus A Hit@10 increased from 94% to 100%, MRR from 0.707524 to 0.850190, and Recall@10 from 69.86% to 78.33%. B-Core Hit@10 increased from 98% to 100%, MRR from 0.933333 to 0.966667, and Recall@10 from 31.58% to 33.79%. Dual-corpus coverage increased from 78% to 92%. The Corpus A baseline belongs to this component analysis and differs from the five-method baseline in Table 2. In Table 6, coverage denotes Hit@10 for each individual corpus and dual-corpus coverage for CROSS_A_B; ranking denotes MRR and mean local MRR, respectively.

Table 6. Controlled comparison without and with legal signals.

Scenario	Signals	Hit/both coverage (%)	MRR/local MRR	Recall@10 (%)
A	Off	94	0.707524	69.86
A	On	100	0.850190	78.33
B-Core	Off	98	0.933333	31.58
B-Core	On	100	0.966667	33.79
Cross-corpus	Off	78	0.811167	7.63
Cross-corpus	On	92	0.878167	8.30

Adding legal signals increased successful retrieval from 47 to 50 of 50 questions in Corpus A, from 49 to 50 in B-Core, and from 39 to 46 for dual-corpus coverage in the controlled comparison. The B-Core coverage gain therefore concerned one question, although MRR also increased. The analysis was descriptive and did not test statistical significance.

3.3. Query Routing and Factual Quotation

The router selected the expected search space for all hundred questions in the balanced set (Table 7). Accuracy decreased to 93% on the stress set, which contained cross-corpus questions with implicit signals. The difference indicates that routing depended on how clearly the question identified its document scope. This evaluation measured routing accuracy; it did not measure latency or compare retrieval quality against searching all corpora.

Table 7. Routing and answer-stage evaluation.

Evaluation	Set/scenario	Passed / total	Rate (%)
Routing	Balanced set	100 / 100	100
Routing	Implicit cross-corpus stress	93 / 100	93
Quotation and citation	A	10 / 10	100
Quotation and citation	B-Core	9 / 10	90
Quotation and citation	Cross-corpus	9 / 10	90
Quotation and citation	All scenarios	28 / 30	93.33

After evidence-selection improvements, twenty-eight of thirty answers satisfied both quotation and citation criteria. All responses completed without service failure or output truncation in this test. The two failures concerned Article 2 paragraph (2) of Government Regulation Number 35 of 2021, once in B-Core and once in a paired cross-corpus question. Although the provision text was present in the evidence, extraction had not preserved a recognizable paragraph marker. The model stated that the requested provision could not be found. This failure demonstrates that evidence presence is insufficient when the structural connection between text and provision numbering is unclear.

The 28/30 result applies to explicit quotation requests on the reused development set. It does not establish legal reasoning accuracy or the relationship between paired provisions. The checks also did not assess every additional statement or all potentially applicable legal grounds. Because only one generator was evaluated, the result cannot be attributed to its performance relative to other models.

3.4. Practical Implications and Limitations

Source control and structural metadata mattered alongside retrieval. Corpus A supported explicit article signals; B-Core required different segmentation. Source-linked presentation enabled inspection but did not establish that a citation fully supported a legal conclusion.

The collection excludes part of a table annex and one archived regulation outside the final scope. Questions and relevance labels were constructed internally, without independent annotation or legal-expert validation. Corpus A development labels were not exhaustive, while B-Core and cross-corpus labels used document and term criteria. Cross-corpus selection and quotation improvements reused their evaluation sets. The legal-signal analysis was descriptive, without significance or weight-sensitivity tests. Latency, usability, and full-application performance were not measured. Retrieval benchmarks and the quotation test evaluated defined components and do not establish that every current interactive-application path implements the same ranking procedure.

Further work should address structural OCR, independently annotated queries, legal-expert assessment, and new cross-corpus and answer holdouts. Weight-sensitivity, routing/latency, and real-user experiments should test benefits beyond internally designed questions. A future sensitivity analysis should vary the exact-route weight and metadata scale while keeping the questions, labels, and representations fixed; it was not performed in this study.

4. Conclusion

Legal signal boosting improved the measured coverage and ranking of hybrid retrieval in the controlled comparisons. Corpus A Hit@10 increased from 94% to 100% and MRR from 0.707524 to 0.850190. B-Core Hit@10 increased from 98% to 100% and MRR from 0.933333 to 0.966667. Dual-corpus coverage increased from 78% to 92%. The effect was not uniform across metrics: TF-IDF obtained higher cross-corpus MAP@10 in the five-method comparison. The system also supported corpus routing and evidence-grounded answers; 28 of 30 quotation answers passed the development-set checks after evidence-selection improvements. The findings support the tested combination of lexical, semantic, and legal signals for this regulatory collection. Independent annotations, external questions, and legal-expert assessment remain necessary to extend the conclusions.

References

- [1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [2] P. Zhao et al., "Retrieval-Augmented Generation for AI-Generated Content: A Survey," Jun 2024, [Online]. Available: <https://arxiv.org/abs/2402.19473>.
- [3] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar 2024, [Online]. Available: <https://arxiv.org/abs/2312.10997>.
- [4] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," Feb 2024, [Online]. Available: <https://arxiv.org/abs/2402.05672>.
- [5] G. Izacard et al., "Unsupervised Dense Information Retrieval with Contrastive Learning," Aug 2022, [Online]. Available: <https://arxiv.org/abs/2112.09118>.
- [6] M. Douze et al., "The Faiss library," Oct 2025, [Online]. Available: <https://arxiv.org/abs/2401.08281>.
- [7] M. Maskur, D. D. Prasetya, T. Widiyaningtyas and A. M. Zain, "Comparative Evaluation of BM25–FAISS and Small-LLM–GPT in Retrieval-Augmented Generation Concept Map Assessment," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 11, no. 1, pp. 173–182, Feb. 2026, doi: 10.22219/kinetik.v11i1.2594.
- [8] F. Redelaar, R. Van Drie, S. Verberne, and M. De Boer, "Attributed Question Answering for Preconditions in the Dutch Law," in *Proceedings of the Natural Legal Language Processing Workshop 2024*, Nov 2024, pp. 154–165, doi: 10.18653/v1/2024.nllp-1.12.
- [9] H.-L. Nguyen, D.-M. Nguyen, T.-M. Nguyen, H.-T. Nguyen, T.-H.-Y. Vuong, and K. Satoh, "Enhancing Legal Document Retrieval: A Multi-Phase Approach with Large Language Models," Mar 2024, [Online]. Available: <https://arxiv.org/abs/2403.18093>.
- [10] S. Khazaeli et al., "A Free Format Legal Question Answering System," in *Proceedings of the Natural Legal Language Processing Workshop 2021*, Nov 2021, pp. 107–113, doi: 10.18653/v1/2021.nllp-1.11.
- [11] Y. Jiang, Z. Guan, J. Zhao, W. Zhao, and J. Yang, "H-LegalKI: A Hierarchical Legal Knowledge Integration Framework for Legal Community Question Answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Nov 2024, pp. 14614–14625, doi: 10.18653/v1/2024.findings-emnlp.856.
- [12] N. Wiratunga et al., "CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering," Apr 2024, [Online]. Available: <https://arxiv.org/abs/2404.04302>.
- [13] R. Kalra et al., "HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications," 2024, [Online]. Available: <https://arxiv.org/abs/2409.09046v1>.
- [14] Suharyadi and I. Saputra, "Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting," *Journal of Novel Engineering Science and Technology*, vol. 4, no. 02, pp. 71–85, Jul 2025, doi: 10.56741/jnest.v4i02.1042.
- [15] I. Schepers, M. Medvedeva, M. Buijn, M. Wieling and M. Vols, "Predicting citations in Dutch case law with natural language processing," *Artificial Intelligence and Law*, vol. 32, no. 3, pp. 807–837, 2024, doi: 10.1007/s10506-023-09368-5.
- [16] H. Ma, Y. Lu, Z. Xiao, J. Feng, H. Zhang and J. Yu, "SDD-LawLLM: Advancing Intelligent Legal Systems Through Synthetic Data-Driven Fine-Tuning of Large Language Models," *Electronics*, vol. 14, no. 4, Art. no. 742, 2025, doi: 10.3390/electronics14040742.
- [17] F. Dadgostari, M. Guim, P. A. Beling, M. A. Livermore and D. N. Rockmore, "Modeling law search as prediction," *Artificial Intelligence and Law*, vol. 29, no. 1, pp. 3–34, 2021, doi: 10.1007/s10506-020-09261-5.
- [18] A. Tagarelli and A. Simeri, "Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code," *Artificial Intelligence and Law*, vol. 30, no. 3, pp. 417–473, 2022, doi: 10.1007/s10506-021-09301-8.
- [19] C. Xiao, X. Hu, Z. Liu, C. Tu and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021, doi: 10.1016/j.aiopen.2021.06.003.

- [20] M. Kim, J. Rabelo, H. K. B. Babiker, M. A. Rahman and R. Goebel, "Legal Information Retrieval and Entailment Using Transformer-based Approaches," *The Review of Socionetwork Strategies*, vol. 18, no. 1, pp. 101–121, 2024, doi: 10.1007/s12626-023-00153-z.
- [21] T. Q. Hoang, H. L. Trung, P. N. V. Hoang and H. Q. Huu, "Optimizing Multi-Stage Language Models for Effective Japanese Legal Document Retrieval," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27672–27681, 2025, doi: 10.48084/etasr.12111.
- [22] H. Ma and Y. Sun, "CECRA: Structured Evidence-Aware Reranking for Long-Document Legal Case Retrieval," *Applied Sciences*, vol. 16, no. 20, Art. no. 10037, 2026, doi: 10.3390/app162010037.
- [23] Ş. Ali, S. Oprea and A. Bâra, "Engineering Trustworthy Retrieval-Augmented Generation for EU Electricity Market Regulation," *Electronics*, vol. 15, no. 4, Art. no. 749, 2026, doi: 10.3390/electronics15040749.
- [24] N. Mukazhanov, Z. Alibiyeva, A. Akhmediyarova, B. Ashirbekov, N. Yerbolat and M. Kanat, "Building a Multilingual AI Legal Assistant Using Retrieval-Augmented Generation: A Case Study on the Legal System of Kazakhstan," *Computers*, vol. 15, no. 9, Art. no. 556, 2026, doi: 10.3390/computers15090556.
- [25] H. Qian et al., "Lightweight Adapted Retrieval-Augmented Inference Network for Legal Issues," *Electronics*, vol. 15, no. 14, Art. no. 2996, 2026, doi: 10.3390/electronics15142996.
- [26] D. Jovanovski, M. Stojcheva, M. Dodevska, P. Lameski, I. Mishkovski and D. Gjorgjevikj, "An Empirical Study of Knowledge Graph-Enhanced RAG for Information Security Compliance," *Information*, vol. 17, no. 4, Art. no. 389, 2026, doi: 10.3390/info17040389.
- [27] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1, 2021.
- [28] N. Pipitone and G. H. Alami, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," Aug 2024, [Online]. Available: <https://arxiv.org/abs/2408.10343>.
- [29] N. Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models," Aug 2023, [Online]. Available: <https://arxiv.org/abs/2308.11462>.
- [30] Z. Fei et al., "LawBench: Benchmarking Legal Knowledge of Large Language Models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Nov 2024, pp. 7933–7962, doi: 10.18653/v1/2024.emnlp-main.452.
- [31] D. Trautmann et al., "Measuring the Groundedness of Legal Question-Answering Systems," in *Proceedings of the Natural Legal Language Processing Workshop 2024*, Nov 2024, pp. 176–186, doi: 10.18653/v1/2024.nllp-1.14.
- [32] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "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*, Mar 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16.
- [33] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Jun 2024, pp. 338–354, doi: 10.18653/v1/2024.naacl-long.20.
- [34] M. Dahl, V. Magesh, M. Suzgun and D. E. Ho, "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models," *Journal of Legal Analysis*, vol. 16, no. 1, pp. 64–93, 2024, doi: 10.1093/jla/lae003.
- [35] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning and D. E. Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," *Journal of Empirical Legal Studies*, vol. 22, no. 2, pp. 216–242, 2025, doi: 10.1111/jels.12413.
- [36] W. Gata and A. Bayhaqy, "Analysis sentiment about islamophobia when Christchurch attack on social media," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1819–1827, 2020, doi: 10.12928/TELKOMNIKA.V18I4.14179.
- [37] W. Gata, F. Amsury, N. K. Wardhani, I. Sugiyarto, D. N. Sulistyowati, and I. Saputra, "Informative Tweet Classification of the Earthquake Disaster Situation In Indonesia," in *2019 5th International Conference on Computing Engineering and Design (ICCED)*, 2019, pp. 1–6, doi: 10.1109/ICCED46541.2019.9161135.
- [38] L. P. Manik, "Performance Factors Effect on the Performance Metrics of the Enterprise Service Bus," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 107–115, 2022, doi: 10.12785/ijcds/110108.
- [39] L. P. Manik, "FEATURES ENGINEERING WITH GENERATING GENETIC ALGORITHM FOR SOFTWARE EFFORT ESTIMATION," *ICIC Express Letters, Part B: Applications*, vol. 13, no. 1, Jan 2022, doi: 10.24507/icicelb.13.01.1.
- [40] R. Shinta Oktaviana et al., "Search Recommendations and Their Impact on Gen Z's Shopping Journey on e-Commerce," in *2025 16th International Conference on E-Education, E-Business, E-Management and E-Learning, IC4e 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 491–495, doi: 10.1109/IC4e65071.2025.11075451.
- [41] R. Shinta Oktaviana, E. H. Hermaliani, N. R. R. Saelan, and E. P. Tamara, "User Satisfaction Factors Self-Service Technology in Fast Food Restaurants Using The Heuristic Evaluation Method," in *2023 8th International Conference on Informatics and Computing, ICIC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, doi: 10.1109/ICIC60109.2023.10382056.
- [42] L. P. Manik et al., "Usability and acceptance of crowd-based early warning of harmful algal blooms," *PeerJ*, vol. 11, Mar 2023, doi: 10.7717/peerj.14923.
- [43] Republik Indonesia, "Undang-Undang Republik Indonesia Nomor 6 Tahun 2023 tentang Penetapan Peraturan Pemerintah Pengganti Undang-Undang Nomor 2 Tahun 2022 tentang Cipta Kerja Menjadi Undang-Undang," 2023.
- [44] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 37 Tahun 2021 tentang Penyelenggaraan Program Jaminan Kehilangan Pekerjaan," 2021.
- [45] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 36 Tahun 2021 tentang Pengupahan," 2021.
- [46] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 35 Tahun 2021 tentang Perjanjian Kerja Waktu Tertentu, Alih Daya, Waktu Kerja dan Waktu Istirahat, dan Pemutusan Hubungan Kerja," 2021.
- [47] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 29 Tahun 2021 tentang Penyelenggaraan Bidang Perdagangan," 2021.
- [48] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 28 Tahun 2021 tentang Penyelenggaraan Bidang Perindustrian," 2021.
- [49] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 27 Tahun 2021 tentang Penyelenggaraan Bidang Kelautan dan Perikanan," 2021.
- [50] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 26 Tahun 2021 tentang Penyelenggaraan Bidang Pertanian," 2021.
- [51] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 23 Tahun 2021 tentang Penyelenggaraan Kehutanan," 2021.
- [52] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 22 Tahun 2021 tentang Penyelenggaraan Perlindungan dan Pengelolaan Lingkungan Hidup," 2021.
- [53] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 21 Tahun 2021 tentang Penyelenggaraan Penataan Ruang," 2021.
- [54] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 19 Tahun 2021 tentang Penyelenggaraan Pengadaan Tanah bagi Pembangunan untuk Kepentingan Umum," 2021.
- [55] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 18 Tahun 2021 tentang Hak Pengelolaan, Hak Atas Tanah, Satuan Rumah Susun, dan Pendaftaran Tanah," 2021.

- [56] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 6 Tahun 2021 tentang Penyelenggaraan Perizinan Berusaha di Daerah," 2021.
- [57] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 7 Tahun 2021 tentang Kemudahan, Pelindungan, dan Pemberdayaan Koperasi dan Usaha Mikro, Kecil, dan Menengah," 2021.
- [58] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 5 Tahun 2021 tentang Penyelenggaraan Perizinan Berusaha Berbasis Risiko," 2021.
- [59] M. Lasandi and N. Jayatilleke, "SinhaLegal: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts," Mar 2026, [Online]. Available: <https://arxiv.org/abs/2603.04854>.
- [60] Z. Zhong, H. Liu, X. Cui, X. Zhang, and Z. Qin, "Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation," Jan 2025, [Online]. Available: <https://arxiv.org/abs/2406.00456>.
- [61] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "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, New York, NY, USA: ACM, Jul 2009, pp. 758–759. doi: 10.1145/1571941.1572114.
- [62] M. Akarsu, R. K. Karaman, and C. Mierbach, "From BM25 to Corrective RAG: Benchmarking Retrieval Strategies for Text-and-Table Documents," Apr 2026, [Online]. Available: <https://arxiv.org/abs/2604.01733>.
- [63] L. P. Manik et al., "Out-of-Scope Intent Detection on A Knowledge-Based Chatbot," International Journal of Intelligent Engineering and Systems, vol. 14, no. 5, pp. 446–457, 2021, doi: 10.22266/ijies2021.1031.39.
- [64] A. Q. Jiang et al., "Mixtral of Experts," Jan 2024, [Online]. Available: <https://arxiv.org/abs/2401.04088>.