

Benchmarking Embedding Models for POJK Regulatory Retrieval in Indonesian Insurance Documents

Daeng Ahmad Nurdin¹, Bambang Jokonowo^{2*}, Bagus Priambodo³

¹⁻³Data Science Department, Faculty of Computer Science, Mercu Buana University, Kota Jakarta Barat, Indonesia

Received:

June 24, 2026

Revised:

July 20, 2026

Accepted:

September 7, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Bambang Jokonowo

Email*:

bambang.jokonowo@mercubuana.ac.id

DOI:

10.63158/journalisi.v8i5.1781

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. Indonesian insurance regulations are clause-dense and terminologically specific, making manual retrieval of applicable provisions difficult. This study presents a retrieval-only benchmark comparing embedding and lexical methods on POJK 26/2025 and POJK 36/2025. The corpus contains 181 context chunks and 181 synthetic Indonesian queries, each assigned one primary relevant chunk. Queries were generated in JSON mode using Gemma 4 E4B via Ollama (ollama:gemma4:e4b), with malformed outputs retrieved and heuristic fallbacks prohibited. Evaluation comprised a four-model local dense comparison and a separate native BGE-M3/BM25 ablation using exact exhaustive ranking. The local comparison reports MRR and Hit rates, while the native ablation additionally reports NDCG@10, MAP@10, bootstrap confidence intervals, paired tests, and rank bins. Qwen3 Embedding achieved the highest local MRR (0.8075). Under a prespecified descriptive composite rule combining MRR, Hit rates, and measured latency, BGE-M3 ranked first. In the native ablation, BGE-M3 combined retrieval achieved NDCG@10 of 0.8810. A practitioner-validated subset audit and artifact-backed qualitative analysis supplemented the quantitative results. Findings are limited to this two-document benchmark and do not evaluate answer generation, citation correctness, or end-to-end RAG faithfulness.

Keywords: regulatory retrieval, legal information retrieval, embedding benchmark, POJK, BGE-M3, BM25

1. INTRODUCTION

Large language models are increasingly used in question answering, but hallucination and weak source grounding remain persistent risks in knowledge-intensive tasks [1]. Retrieval-augmented generation (RAG) addresses this problem by bringing relevant passages into the prompt before generation, which improves factuality and traceability [2]. In such systems, the retrieval layer is often decisive because a generative model can only answer well if the retrieved context is semantically aligned with the user's query.

This study focuses on regulatory documents issued by the Indonesian Financial Services Authority, also known as *Otoritas Jasa Keuangan* (OJK). Specifically, it examines Financial Services Authority Regulations, also known as *Peraturan Otoritas Jasa Keuangan* (POJK). These POJK documents are highly complex because they combine formal legal language, dense clause structure, and domain-specific financial terminology. While digital document portals provide public access to statutory texts, searching long-form regulatory corpora manually remains inefficient when compliance officers need to identify fine-grained statutory obligations or explanatory clauses across complex regulatory structures [3].

The urgency of automating this process is particularly evident in the Indonesian insurance and reinsurance sector, where POJK provisions are routinely consulted for compliance interpretation, product review, and operational alignment. The intended users are compliance units, internal legal teams, and insurance-company risk or product-review staff rather than the public. Faster and more reliable retrieval of POJK chunks can support internal review, but the present study measures retrieval quality only and does not measure institutional or policy outcomes.

For example, a compliance officer reviewing a proposed investment may ask which POJK 26/2025 provision defines admissible assets or limits exposure to a particular instrument. A retrieval system should return the relevant regulatory chunk near the top of the ranking so that the officer can inspect the original wording and associated conditions. This scenario motivates the present task definition: retrieval of the primary relevant chunk, not generation of a legal interpretation. POJK 26/2025 and POJK 36/2025 were selected because they cover two distinct operational domains (asset-liability

management and the health-insurance ecosystem) while maintaining an auditable corpus size for exact exhaustive ranking.

Within the broader literature on information retrieval, dense semantic retrieval has established itself as a practical foundation for question answering. REALM and RAG demonstrated the value of external retrieval for knowledge-intensive language tasks, while DPR and Sentence-BERT shifted the paradigm toward continuous vector matching [4], [5], [6], [7]. Furthermore, benchmark efforts such as BEIR, MTEB, and RAGAs have shown that embedding quality must be interpreted under a clearly controlled task design rather than through one universal score, confirming that retrieval configuration remains central to factuality and domain adaptation [8], [9], [10], [11].

However, recent studies also suggest that single-vector semantic matching is not always sufficient for clause-dense corpora. Multi-signal retrieval models, such as SPLADE (which strengthens lexical precision through sparse expansion) and ColBERTv2 (which preserves token-level late interaction), indicate that balancing semantic understanding with exact-term matching yields superior results [12], [13]. Building on these concepts, BGE-M3 integrates dense, sparse, and multi-vector outputs within a single multilingual architecture. This combination is especially relevant for regulatory retrieval, where users may search with exact legal vocabulary, paraphrased obligations, or fine-grained clause distinctions [14].

Domain-oriented legal retrieval studies reinforce the necessity of specialized retrieval configurations. Recent work in Arabic and Vietnamese legal retrieval shows that chunking quality, domain-aware negative mining, and the balance between semantic and lexical signals all affect ranking reliability [15], [16]. Similarly, legal information retrieval research that combines BM25, transformer methods, and semantic support concludes that hybrid approaches often outperform monolithic dense models [17]. Even so, empirical evidence remains limited regarding how these advanced retrieval paradigms translate to Indonesian regulatory corpora.

A growing body of research has explored general Indonesian legal retrieval, but POJK retrieval has a narrower operational character: the target text contains dense financial terminology, article references, numerical thresholds, and repeated formal clauses. These

properties make exact lexical matching important alongside semantic similarity. There is therefore value in evaluating chunk retrieval on a small, auditable POJK corpus before considering broader legal or end-to-end RAG claims.

To bridge this gap, this study evaluates how lexical, dense, sparse, multi-vector, and fused retrieval methods perform on Indonesian POJK insurance regulations under a controlled retrieval-only benchmark. Rather than proposing an entirely new neural architecture, the contribution lies in constructing a reproducible POJK retrieval benchmark and systematically comparing BGE-M3 against a spectrum of alternatives, including qwen3-embedding, nomic-embed-text-v2-moe, and nomic-embed-text. This methodological approach provides controlled evidence about retrieval-quality and computational-efficiency trade-offs under the specified evaluation protocols.

In summary, this research provides a reproducible retrieval benchmark and compares dense, sparse, multi-vector, and lexical methods under the same two-document corpus. Its contribution is methodological rather than a claim of broad POJK coverage: it documents a fixed query/qrel construction, exact ranking protocol, statistical uncertainty, practitioner validation of a subset, and latency decomposition that can be extended with larger corpora and human judgments.

2. METHODS

This study employs a controlled comparative experimental design grounded in quantitative evaluation logic for information retrieval systems. The corpus, query pool, top-k retrieval depth, and ranking procedure are held constant within each comparison, while the candidate dense model or retrieval configuration is varied [18]. This design limits confounding from document distribution and reranking policies, although workstation latency remains configuration-dependent [19], [20].

2.1. Research Design and Workflow

To systematically evaluate the models against regulatory text, the research procedure was structured into a five-stage workflow, as illustrated in Figure 1. This workflow was designed to emulate a realistic RAG deployment pipeline while maintaining strict experimental control over the retrieval phase.

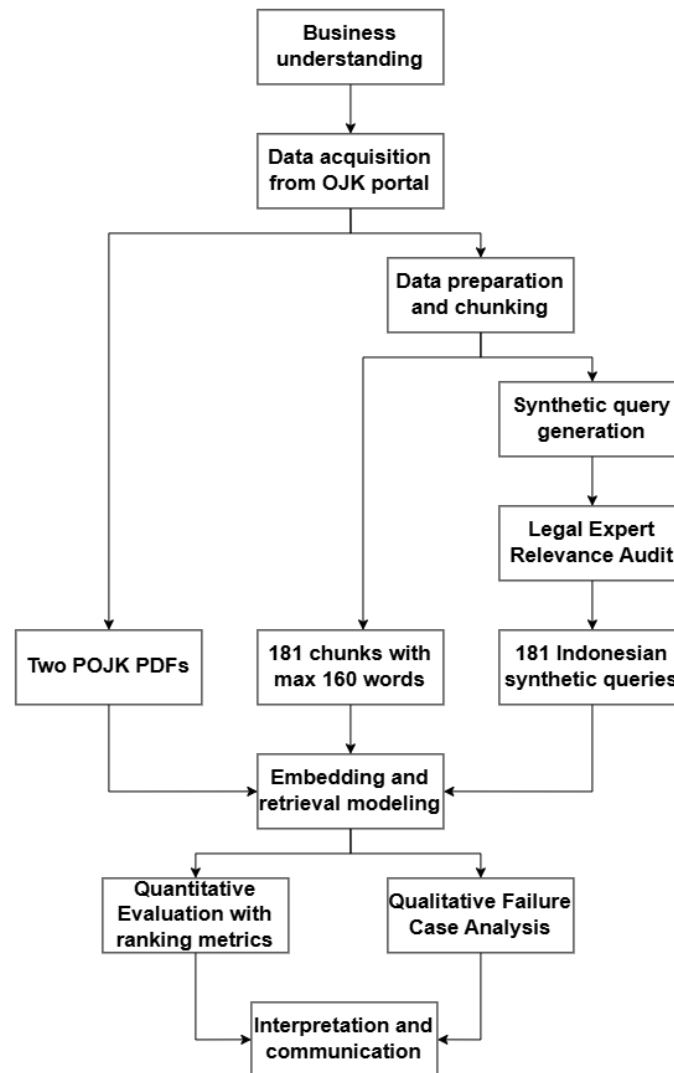


Figure 1. Research Design and Evaluation Workflow

The stages implemented in the research design are as follows:

- 1) Data Acquisition and Normalization: Raw PDF documents of POJK regulations were collected from the official OJK portal, structurally normalized, and segmented into fixed, non-overlapping text windows of at most 160 words.
- 2) Synthetic Query Generation & Validation: To construct a domain-specific evaluation pool, a localized LLM (ollama:gemma4:e4b) generated one contextual question for each chunk in JSON mode. The authors prepared a stratified 30-query audit worksheet by randomly sampling 15 queries from each regulation with seed 2025, and an insurance compliance practitioner

validated primary relevance, context dependency, and terminology against the source chunks.

- 3) **Corpus and Query Encoding:** Both the regulatory chunks and the synthetic queries were mapped into a shared vector space using each candidate embedding model.
- 4) **Similarity Search:** For each query, the top-k relevant contexts were retrieved from the corpus using exact exhaustive scoring without secondary reranking.
- 5) **Evaluation & Analysis:** The retrieval outputs were quantitatively evaluated using ranking metrics, statistical tests for the native ablation, and computational efficiency measurements. This was supplemented by qualitative analysis of artifact-backed divergence cases.

2.2. Data Gathering and Preprocessing

The primary data source consisted of two regulatory documents retrieved directly from the public OJK portal: POJK 26 of 2025 on Asset and Liability Management in Insurance and Reinsurance Companies, also known as *POJK 26 Tahun 2025 Pengelolaan Aset dan Liabilitas Perusahaan Asuransi dan Perusahaan Reasuransi*, and POJK 36 of 2025 on Strengthening the Health Insurance Ecosystem, also known as *POJK 36 Tahun 2025 Penguatan Ekosistem Asuransi Kesehatan*. After whitespace normalization and fixed non-overlapping windows of at most 160 words, the pipeline yielded 181 context chunks and 36,023 source tokens. Each window serves as an independent retrieval unit without enforced sentence or article boundaries, preserving structural layout only as captured during PDF text extraction. The final evaluation uses one synthetic Indonesian query per chunk. The exact query-token total is reported from the regenerated manifest. The detailed corpus specifications, extraction parameters, and query generation settings are summarized in Table 1.

Table 1. Corpus and preprocessing settings

Component	Value	Interpretation
Source files	POJK 26 of 2025 (POJK 26 Tahun 2025), POJK 36 of 2025 (POJK 36 Tahun 2025)	Public regulatory corpus retrieved from the OJK portal

Component	Value	Interpretation
Source mode files	Static files	The benchmark was built from static document files rather than live web scraping
LLM for synthetic query generation	Ollama Gemma 4 E4B (gemma4:e4b, model ID c6eb396dbd59), 8.0B parameters, Q4_K_M, in JSON mode	Each chunk requires one valid model-generated JSON pair. Malformed responses are retried up to 3 times, and execution fails without heuristic fallbacks
Chunk limit and overlap	160 words (0-word overlap)	The source pipeline uses fixed non-overlapping windows without enforced sentence or article boundaries
Chunk count	181	Number of indexed context units used in retrieval
Query count	181	One Gemma-generated Indonesian query per chunk
Source token count	36,023	Indicates a long-form regulatory corpus with dense clause structure
Query token count	3,807	Shows the compact size of the evaluation questions compared with the source text

The LLM prompt is: *"Buat tepat satu pasangan pertanyaan dan jawaban berbahasa Indonesia untuk benchmark retrieval dari potongan POJK berikut. Pertanyaan harus spesifik, dapat dijawab sepenuhnya dari potongan, dan tidak boleh generik. Jawaban harus ringkas, faktual, dan hanya memakai informasi dari potongan. Keluarkan tepat satu JSON object valid tanpa Markdown atau teks tambahan dengan skema ini: {"question": "...", "answer": "..."}"* The prompt is followed by the chunk identifier and chunk text.

Generation uses gemma4:e4b (model ID c6eb396dbd59), temperature 0, and seed 2025, validates the JSON schema and content, retries malformed responses up to three times, and aborts execution if generation fails, strictly prohibiting heuristic fallbacks. Ground truth is binary: each query has one primary relevant chunk, so Recall@10 and Hit@10 are numerically equivalent in this setup. Table 2 presents representative examples of the generated Indonesian queries alongside their corresponding source chunks and ground-truth legal focus. To ensure full auditability and experimental reproducibility, the complete corpus, extracted chunks, generated query pool, and ground-truth qrels are provided as supplementary materials submitted to the journal editorial office.

Table 2. Representative examples of corpus-query pairs

Source File	Chunk ID	Example query	Ground-truth focus
POJK 26 of 2025 (POJK 26 Tahun 2025)	chunk-1	What is the legal basis considered in the enactment of Financial Services Authority Regulation Number 26 of 2025? (Apa dasar hukum yang menjadi pertimbangan penetapan Peraturan Otoritas Jasa Keuangan Nomor 26 Tahun 2025?)	Statutory considerations underlying POJK 26/2025 (Pertimbangan hukum pembentukan POJK 26/2025)
POJK 26 of 2025 (POJK 26 Tahun 2025)	chunk-8	What is the minimum solvency level that must be maintained by the Company at all times under POJK 26 of 2025? (Berapa Tingkat Solvabilitas paling rendah yang wajib dipenuhi Perusahaan setiap saat menurut POJK 26 Tahun 2025?)	Minimum solvency-level requirement (Kewajiban tingkat solvabilitas minimum)
POJK 26 of 2025 (POJK 26 Tahun 2025)	chunk-10	What types of securities can qualify as investment assets placed abroad? (Jenis surat berharga apa saja yang dapat menjadi aset investasi yang ditempatkan di luar negeri?)	Eligible securities for overseas investment placement (Surat berharga yang memenuhi syarat investasi luar negeri)

2.3. Embedding Architecture and Task Mechanics

BGE-M3 is used as the focal model because its published architecture exposes dense, sparse lexical, and multi-vector outputs [14]. The local dense-model comparison evaluates four Ollama embedding models through their dense paths, while the native BGE-M3 ablation reported here separately evaluates sparse and multi-vector modes. This separation avoids attributing ablation results to the dense-only local comparison.

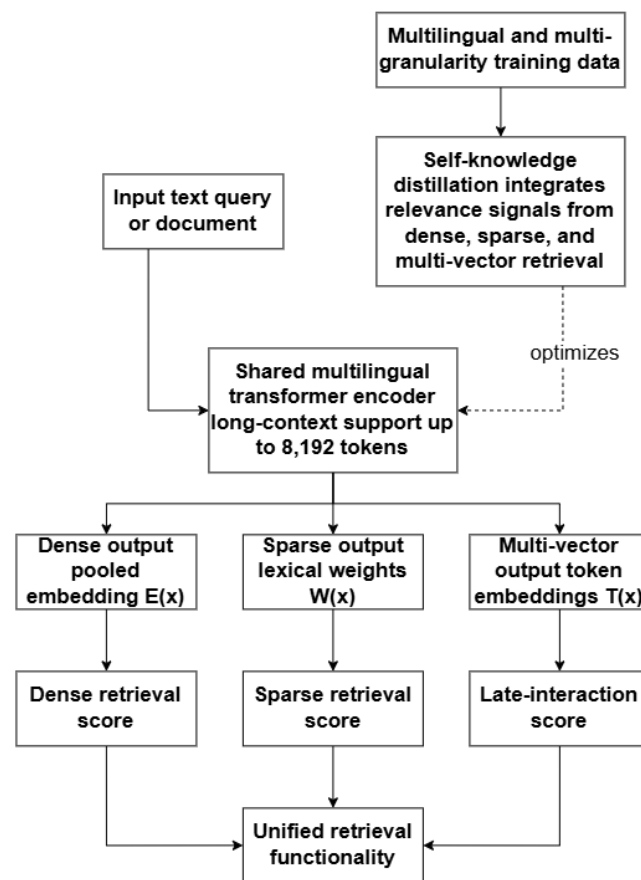


Figure 2. Architectural overview of the BGE-M3 model. Adapted from [14]

From an algorithmic perspective, as pictured in Figure 2, BGE-M3 follows a shared-encoder pipeline where both POJK chunks and user queries are passed through the same transformer encoder, producing a pooled dense vector $E(x)$, lexical weights $W(x)$, and token-level embeddings $T(x)$. Self-knowledge distillation acts as the unifying training mechanism, enabling the model to balance semantic flexibility with exact legal-term precision [14]. In this benchmark, BGE-M3 is evaluated off-the-shelf in a zero-shot setting without domain fine-tuning.

The formal mathematical formulations governing the retrieval paths are specified as follows. In the dense path, the model compares a query vector $E(q)$ with a chunk vector $E(d)$ via cosine similarity, as defined in Equation (1):

$$\text{Score}_{\text{dense}}(q, d) = \cos(E(q), E(d)) = \frac{E(q) \cdot E(d)}{\|E(q)\|_2 \|E(d)\|_2} \quad (1)$$

In the sparse lexical path, token-level importance weights $w_q(t)$ and $w_d(t)$ computed by the encoder are matched across overlapping vocabulary according to Equation (2):

$$\text{Score}_{\text{sparse}}(q, d) = \sum_{t \in q \cap d} w_q(t) \cdot w_d(t) \quad (2)$$

In the multi-vector (ColBERT-style late interaction) path, max-similarity token alignments are aggregated across all query token vectors $T_q^{(i)}$ and candidate chunk token vectors $T_d^{(j)}$ as formulated in Equation (3):

$$\text{Score}_{\text{multi}}(q, d) = \frac{1}{|T_q|} \sum_{i=1}^{|T_q|} \max_{j=1}^{|T_d|} \cos(T_q^{(i)}, T_d^{(j)}) \quad (3)$$

In the hybrid combined path, dense, sparse, and multi-vector scores are independently normalized to [0,1] by per-query min-max normalization and then combined through linear convex fusion with empirical weights ($\alpha = 0.4$, $\beta = 0.2$, $\gamma = 0.4$), as expressed in Equation (4):

$$\text{Score}_{\text{final}}(q, d) = \alpha \widehat{\text{Score}}_{\text{dense}}(q, d) + \beta \widehat{\text{Score}}_{\text{sparse}}(q, d) + \gamma \widehat{\text{Score}}_{\text{multi}}(q, d), \quad \alpha + \beta + \gamma = 1 \quad (4)$$

All dense vectors in the native ablation are L_2 -normalized before exhaustive cosine scoring. While the four-model local comparison evaluates dense retrieval exclusively, the native BGE-M3 ablation investigates sparse and multi-vector outputs.

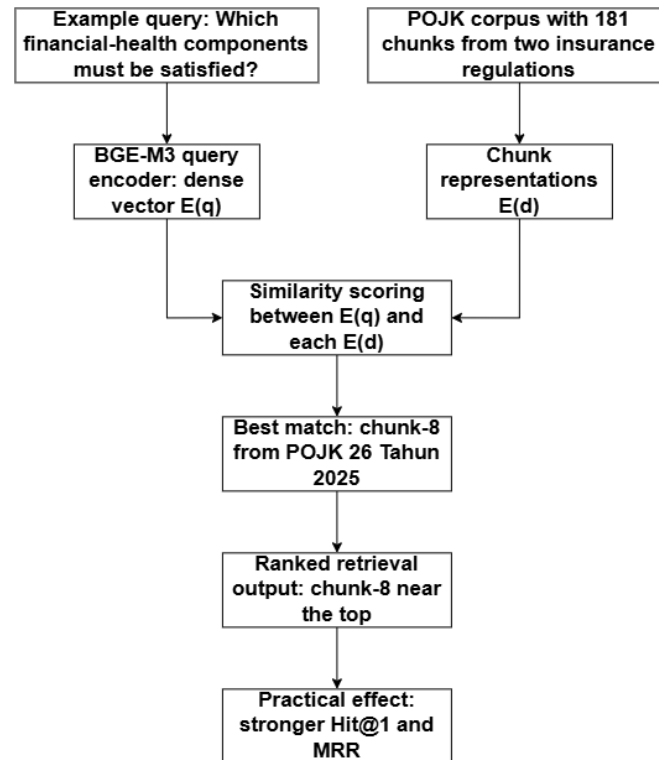


Figure 3. Query-to-chunk mapping visualization. Source: Authors.

The flowchart in Figure 3 illustrates how BGE-M3 maps a financial-health query to its corresponding regulation chunk by encoding both into a comparable space and computing relevance. The distinct functional roles of these three retrieval outputs and their specific operational implications for POJK regulatory text are summarized in Table 3.

Table 3. Output modes of BGE-M3 and implications for POJK retrieval

Output mode	Role in retrieval	Implication for POJK documents
Dense retrieval	Single-vector semantic matching	Captures paragraph-level meaning and paraphrased questions
Sparse lexical matching	Token-weighted exact term matching	Useful for article numbers, legal terms, and formal phrases that should not be softened
Multi-vector interaction	Late interaction at token granularity	Supports clause-level discrimination when similar wording appears across different sections

2.4. Comparative Protocol and Multi-Criteria Scoring

The local dense-model comparison includes four embeddings: BGE-M3, qwen3-embedding, nomic-embed-text-v2-moe, and nomic-embed-text. BGE-M3 is a multilingual multi-function embedding model, qwen3-embedding is a modern multilingual open model [21], nomic-embed-text-v2-moe is the middle-weight general multilingual baseline [22], and nomic-embed-text is the lightweight local baseline utilizing Matryoshka representation learning [23]. Holding the evaluation pool, corpus, and top-k depth constant enables controlled comparison under this specific configuration without implying universal model superiority.

The evaluation protocol uses exact exhaustive ranking over all 181 chunks, with no HNSW index, approximate nearest-neighbor search, or reranking. Dense vectors are L_2 -normalized before cosine scoring. The local dense comparison reports MRR, Hit@1, Hit@3, Hit@10, and latency. The native BGE-M3 ablation additionally measures sparse lexical and ColBERT-style multi-vector scores, plus a normalized three-way fusion, and reports NDCG@10, Recall@10, MAP@10, and mean first relevant rank. For the native ablation, bootstrap confidence intervals are computed from 2,000 query-resampling replicates (seed 2025), complemented by paired t-tests and Wilcoxon signed-rank tests for key comparisons. The exact model checkpoints, dimensionalities, execution modes, and environment specifications for all evaluated models are detailed in Table 4.

The native ablation was executed on a dedicated Windows workstation environment equipped with an AMD Ryzen 5 CPU, NVIDIA GeForce GTX 3050 (4 GB VRAM) GPU, and 16 GB system RAM. It used batch size 4 with the `--no-fp16` command flag, resulting in FP32 execution under the available GPU-memory constraint. The ablation explicitly separates corpus encoding, query encoding, and exhaustive score computation. The generation and statistical-resampling procedures use seed 2025, and cached data are invalidated whenever source document fingerprints change. The local models were served sequentially by a GPU-enabled Dockerized Ollama 0.32.11 instance. The benchmark submitted one text per embedding request and performed one measured run per model, so its latency results are descriptive rather than repeated-run estimates. Table 4 records the exact published aliases and immutable 12-character Ollama model identifiers used in this environment. Because nomic-embed-text-v2-moe has no versioned public Ollama alias beyond latest, its immutable model ID is reported.

Table 4. Model configuration and execution context

Model	Exact checkpoint	Dimension	Execution mode	Tokenizer/max input	Quantization and batching	Input/search notes
BGE-M3	BAAI/bge-m3 through native FlagEmbedding and bge-m3:567m-fp16 (790764642607) in the dense comparison	1024	Native local ablation / Dockerized Ollama dense comparison	1024 tokens in the native ablation and 8192 tokens in Ollama	Native FP32 with batch size 4. Ollama F16 with one text per request	Native dense, sparse, and multi-vector outputs. Ollama dense-only output. L_2 -normalized dense vectors
qwen3-embedding	qwen3-embedding:8b-q4_K_M (64b933495768)	4096	Dockerized Ollama dense comparison	40,960 tokens	Q4_K_M, one text per request	Dense-only comparison with configuration-dependent latency
nomic-embed-text-v2-moe	nomic-embed-text-v2-moe:latest (ff9c2f10ef5e, the only published Ollama tag)	768	Dockerized Ollama dense comparison	512 tokens	F16, one text per request	Dense-only comparison
nomic-embed-text	nomic-embed-text:137m-v1.5-fp16 (0a109f422b47)	768	Dockerized Ollama dense comparison	2048 tokens	F16, one text per request	Dense-only comparison

To compare retrieval quality and operational cost descriptively, this study applies a multi-criteria decision logic that prioritizes top-ranked precision while penalizing latency [10], [20]. The multi-criteria decision logic is formalized as a composite evaluation function $\mathcal{F}(m)$ in Equation (5) for each candidate model m :

$$\mathcal{F}(m) = w_1 Q_{\text{primary}} + w_2 Q_{\text{secondary}} - w_3 L_{\text{norm}} \quad (5)$$

where:

$$Q_{\text{primary}} = \text{MRR}$$

$$Q_{\text{secondary}} = \frac{\text{Hit@1} + \text{Hit@3} + \text{Hit@10}}{3}$$

$$L_{\text{norm}} = \frac{\text{Latency}_m}{\max_k \text{Latency}_k}$$

The weights ($w_1 = 0.6$, $w_2 = 0.3$, $w_3 = 0.1$) represent a prespecified descriptive preference for early-rank precision over loose top-ten coverage, with latency treated as a secondary operational penalty. They are not learned from production click logs. The manuscript therefore does not use the composite score to declare a universal winner. Disaggregated metrics remain the primary basis of evaluation.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation

The final ablation evaluates 181 query-chunk pairs. Because each query has one primary relevant chunk, Hit@10 and Recall@10 are identical for each method, while MAP@10 equals MRR when the first relevant chunk is the only relevant target. The primary empirical findings of the native retrieval ablation are presented in Table 5. BGE-M3 combined achieves the highest overall retrieval scores across NDCG@10, MRR, Hit@3, and Hit@10, closely followed by sparse BGE-M3 and ColBERT-style late interaction. BM25 is also a strong baseline, whereas BGE-M3 dense-only is lower on this extractive query set.

The BGE-M3 dense-versus-combined difference in NDCG@10 is 0.0546 and is statistically distinguishable under both the paired t-test ($p = 3.10 \times 10^{-6}$) and Wilcoxon test ($p = 4.36 \times 10^{-6}$). BGE-M3 dense is higher than BM25 by 0.0306 NDCG points, but this difference is not statistically distinguishable under the paired t-test ($p = 0.156$) or Wilcoxon test ($p = 0.217$). These tests quantify differences under the current synthetic qrels and should not be interpreted as evidence of performance on unseen regulatory documents.

Table 5. Retrieval ablation results on 181 query-chunk pairs

Method	NDCG@10 (95% CI)	MRR (95% CI)	Hit@1 (95% CI)	Hit@3 (95% CI)	Hit@10 / Recall@10 (95% CI)	MAP@10	Mean First rank
BGE-M3 combined	0.8810 [0.8456, 0.9135]	0.8488 [0.8056, 0.8917]	0.7624 [0.7017, 0.8232]	0.9392 [0.9006, 0.9724]	0.9779 [0.9558, 0.9945]	0.8488	1.4199
BGE-M3 sparse	0.8797 [0.8438, 0.9126]	0.8473 [0.8021, 0.8862]	0.7680 [0.7072, 0.8287]	0.9227 [0.8785, 0.9558]	0.9779 [0.9558, 0.9945]	0.8473	1.4309
BGE-M3 ColBERT	0.8702 [0.8317, 0.9051]	0.8382 [0.7957, 0.8799]	0.7569 [0.6961, 0.8177]	0.9006 [0.8508, 0.9392]	0.9669 [0.9392, 0.9890]	0.8382	1.4088
BM25+dense hybrid	0.8657 [0.8264, 0.9036]	0.8350 [0.7897, 0.8808]	0.7680 [0.7072, 0.8233]	0.8840 [0.8343, 0.9282]	0.9613 [0.9282, 0.9890]	0.8350	1.5083
BGE-M3 dense	0.8264 [0.7868, 0.8633]	0.7805 [0.7318, 0.8262]	0.6685 [0.6022, 0.7348]	0.8840 [0.8343, 0.9282]	0.9669 [0.9392, 0.9890]	0.7805	1.6851
BM25	0.7958 [0.7441, 0.8413]	0.7584 [0.7022, 0.8100]	0.6740 [0.6022, 0.7403]	0.8398 [0.7845, 0.8895]	0.9116 [0.8674, 0.9503]	0.7584	1.5470

Table 6. Rank bins for the native ablation on 181 query-chunk pairs

Method	Rank 1	Ranks 2-3	Ranks 4-10	Outside top 10
BGE-M3 combined	138	32	7	4
BGE-M3 sparse	139	28	10	4
BGE-M3 ColBERT	137	26	12	6
BM25+dense hybrid	139	21	14	7
BGE-M3 dense	121	39	15	6
BM25	122	30	13	16

Table 6 provides the complete rank-bin distribution for all native-ablation methods. The native BGE-M3 run required 193.68 seconds for corpus encoding, 4.39 seconds for query encoding, and 15.38 seconds for exhaustive score computation, for 213.45 seconds in total. BM25 indexing and scoring required 0.03 and 0.23 seconds respectively.

3.2. Practitioner Validation of a Benchmark Audit Subset

Following evaluation-validation practices in domain-specific legal information retrieval benchmarks [15], [16], the authors prepared an audit worksheet for a stratified random sample of 30 queries ($\approx 16.5\%$ of the 181-query pool), selected as 15 queries from POJK 26/2025 and 15 queries from POJK 36/2025 using seed 2025. An insurance compliance practitioner then validated primary relevance, context dependency, terminology authenticity, and qualitative notes against the supplied source chunks.

The evaluation assessed: (i) *Primary Ground-Truth Accuracy* (whether the single assigned chunk directly answers the regulatory question), and (ii) *Contextual Completeness* (whether adjacent chunks or general definitions in Article 1 are necessary to fully resolve the query).

The audit found that 28 of the 30 sampled queries (93.33%) had their primary legal ground truth accurately and uniquely identified by the synthetic pipeline. For 4 of the 30 queries (13.33%), while the primary chunk contained the core operative rule, an adjacent clause provided complementary context. These subset results support the assigned targets within the audited sample, while indicating that production compliance systems would benefit from multi-chunk context expansion and graded relevance labels. Table 7 presents representative audited queries and their validation outcomes.

Table 7. Representative sample of synthetic queries in the practitioner-validated audit

Query ID	Regulation (Target Article)	Synthetic Question	Practitioner Assessment	Audit Justification
POJK26- chunk- 45	POJK	How long must the	Valid	The core
	26/2025	Company submit an action	(Requires	timeline is
	Article 21	plan for resolving violations	Adjacent	correct. The
	(POJK	of maximum investment	Context)	definition of

Query ID	Regulation (Target Article)	Synthetic Question	Practitioner Assessment	Audit Justification
	26/2025 Pasal 21)	limits to the Financial Services Authority after such violation is determined? (Berapa lama Perusahaan wajib menyampaikan rencana tindak penyelesaian pelanggaran batasan maksimum investasi kepada Otoritas Jasa Keuangan setelah penetapan pelanggaran tersebut?)		"batasan maksimum" requires preceding article context.
POJK26- chunk-1	POJK 26/2025 Preamble (POJK 26/2025 Pembukaan)	What is the legal basis considered in the enactment of Financial Services Authority Regulation Number 26 of 2025? (Apa dasar hukum yang menjadi pertimbangan penetapan Peraturan Otoritas Jasa Keuangan Nomor 26 Tahun 2025?)	Valid (Standalone)	Directly quotes and accurately captures the legislative preamble.
POJK26- chunk- 96	POJK 26/2025 Article 86 (POJK 26/2025 Pasal 86)	What is the status of Financial Services Authority Regulation Number 71/POJK.05/2016 concerning Financial Health of Insurance and Reinsurance Companies following the enactment of POJK 26 of 2025? (Bagaimana status Peraturan Otoritas Jasa	Valid (Standalone)	Accurately tests the transitional status of the prior regulation.

Query ID	Regulation (Target Article)	Synthetic Question	Practitioner Assessment	Audit Justification
		Keuangan Nomor 71/POJK.05/2016 tentang Kesehatan Keuangan Perusahaan Asuransi dan Perusahaan Reasuransi setelah berlakunya POJK 26 Tahun 2025?)		

3.3. Qualitative Retrieval Error and Failure Case Analysis

To examine the practical mechanics behind the ranking distributions in Table 6, qualitative error analysis was conducted on queries where distinct retrieval paradigms diverged. Table 8 details four representative cases, including both successful and failed combined retrieval.

Table 8. Artifact-backed qualitative retrieval divergence cases

Case ID	Query & Target Regulation	Ground-Truth Clause Focus	BGE-M3 Combined Rank	BM25 Rank	BGE-M3 Dense Rank	Observed Ranking Behavior
Case 1	What administrative sanctions are imposed on Companies violating the provisions of Article 14 paragraph 2 and/or Article 15? (Apa sanksi administratif yang dikenakan	Article 17: administrative sanctions for violations of Article 14(2) and/or Article 15 (Pasal 17: sanksi administratif atas pelanggaran Pasal 14 ayat 2 dan/atau Pasal 15)	1	1	10	BM25, sparse BGE-M3, ColBERT, and combined retrieval rank the target first, whereas dense retrieval places it at Rank 10. The explicit article references coincide with

Case ID	Query & Target Regulation	Ground-Truth Clause Focus	BGE-M3 Combined Rank	BM25 Rank	BGE-M3 Dense Rank	Observed Ranking Behavior
	kepada Perusahaan yang melanggar ketentuan Pasal 14 ayat 2 dan/atau Pasal 15?) [POJK 36/2025, chunk-14]					strong lexical and token-level rankings in this case.
Case 2	What is meant by 'Company' in this Financial Services Authority Regulation? (Apa yang dimaksud dengan 'Perusahaan' dalam Peraturan Otoritas Jasa Keuangan ini?) [POJK 26/2025, chunk-2]	Assigned definition-related source window (Potongan definisi 'Perusahaan')	1	>10	1	Dense and combined retrieval rank the assigned target first. BM25 does not retrieve it in the top 10. This case illustrates divergence between semantic and lexical ranking for a short definitional query.
Case 3	According to Article 9 Paragraph 1, what is the impact of	Article 9(1) explanation: non-compliant investment placement is	1	>10	9	Combined and ColBERT retrieval rank the target first, sparse

Case ID	Query & Target Regulation	Ground-Truth Clause Focus	BGE-M3 Combined Rank	BM25 Rank	BGE-M3 Dense Rank	Observed Ranking Behavior
	violating the specified provisions? (Menurut Pasal 9 Ayat 1, apa dampak pelanggaran terhadap ketentuan yang disebutkan?) [POJK 26/2025, chunk-103]	excluded from admissible assets (Penjelasan Pasal 9 ayat 1: penempatan investasi tidak patuh dikeluarkan dari aset yang diperkenankan)				retrieval ranks it second, dense retrieval ranks it ninth, and BM25 misses the top 10. The spread indicates that an explicit cross-reference alone does not ensure consistent ranking across retrieval modes.
Case 4	What is meant by strategic policy according to POJK 26 of 2025? (Apa yang dimaksud dengan kebijakan strategis menurut POJK 26 Tahun 2025?) [POJK 26/2025, chunk-112]	Definition of strategic policy (<i>kebijakan strategis</i>) as a policy concerning business direction and objectives with significant impact (Definisi kebijakan strategis terkait arah dan	>10	8	>10	Combined, dense, ColBERT, and BM25+dense retrieval miss the target in the top 10. Sparse retrieval ranks it third and BM25 ranks it eighth. This combined-retrieval

Case ID	Query & Target Regulation	Ground-Truth Clause Focus	BGE-M3 Combined Rank	BM25 Rank	BGE-M3 Dense Rank	Observed Ranking Behavior
		sasaran usaha berdampak signifikan)				Failure shows that score fusion can suppress a target supported primarily by an exact definitional phrase.

3.4. Local Dense Model Comparison

Beyond the native ablation, a separate comparison evaluated four locally served embedding models over the same corpus, query pool, and retrieval depth. Its underlying values are the current 181-query, zero-fallback run described in Table 9 and the accompanying report artifact. Reported latencies are descriptive workstation measurements comprising corpus encoding, query encoding, and exact scoring rather than end-to-end production timings.

Table 9. Local dense-model comparison on 181 query-chunk pairs

Model	MRR	Hit@1	Hit@3	Hit@10	Latency (s)
qwen3-embedding:8b-q4_K_M	0.8075	0.7182	0.8674	0.9779	430.00
bge-m3:567m-fp16	0.7918	0.6740	0.8950	0.9890	141.36
nomic-embed-text-v2-moe:latest (ff9c2f10ef5e)	0.7748	0.6575	0.8729	0.9834	120.58
nomic-embed-text:137m-v1.5-fp16	0.4660	0.3591	0.5304	0.7072	37.14

The local dense-model comparison presented in Table 9 ranks Qwen3 first by MRR, followed by BGE-M3 and Nomic Embed v2 MoE. BGE-M3 has the highest Hit@3 and

Hit@10 of the local models, while Nomic Embed Text is fastest but substantially lower in retrieval quality. No inferential tests were conducted for this local comparison, so these differences are reported descriptively. The comparison is also separate from the native BGE-M3 ablation in Table 5, which uses different output modes and latency accounting.

Table 10. Descriptive composite-score calculation for the four-model local comparison

Model	Q_{primary}	$Q_{\text{secondary}}$	L_{norm}	$\mathcal{F}(m)$
bge-m3:567m-fp16	0.7918	0.8527	0.3287	0.6980
nomic-embed-text-v2-moe:latest (ff9c2f10ef5e)	0.7748	0.8379	0.2804	0.6882
qwen3-embedding:8b-q4_K_M	0.8075	0.8545	10.000	0.6409
nomic-embed-text:137m-v1.5-fp16	0.4660	0.5322	0.0864	0.4306

Table 10 applies the prespecified composite function to the same local comparison, with Q_{primary} defined as MRR. BGE-M3 has the highest descriptive composite score because its near-leading quality is paired with lower measured runtime than Qwen3. This ranking reflects the prespecified evaluation weights (0.6, 0.3, 0.1) and sequential local execution path rather than a generalized deployment recommendation.

3.5. Discussion

The ablation results support a clear, bounded conclusion. BGE-M3 combined retrieval achieved the highest NDCG@10 (0.8810), followed closely by sparse BGE-M3 (0.8797) and ColBERT-style retrieval (0.8702), while the BM25+dense hybrid (0.8657) markedly outperformed dense-only BGE-M3 (0.8264). This pattern shows that lexical and token-level signals effectively complement a single dense vector on regulatory text. Crucially, BGE-M3 is evaluated in an off-the-shelf zero-shot setting and is not domain-fine-tuned on Indonesian legal text. These findings do not establish universal superiority, because the query set is synthetic, the corpus contains two regulations, and ground truth was assigned through an automated single-target protocol.

BM25's strong result is substantively important for regulatory retrieval. Exact article references, defined statutory terms, financial instrument names, and numeric thresholds represent critical signals that dense-only vector spaces frequently smooth over. Combining normalized lexical and dense scores improves retrieval robustness, but native

multi-vector late interaction yields superior clause discrimination when similar statutory phrasing appears across different chapters.

This pattern is consistent with SPLADE's evidence that learned sparse representations preserve lexical discrimination [12] and ColBERTv2's finding that token-level late interaction can improve fine-grained ranking [13]. It also accords with legal retrieval results reported by Kim et al. [17], where BM25 and transformer-based semantic methods provide complementary evidence rather than a universally dominant single method. The present study extends that comparison to Indonesian POJK clauses, but its smaller corpus and synthetic single-target qrels prevent direct performance equivalence with those broader studies.

The timing results should also be interpreted locally. The native BGE-M3 ablation separates corpus encoding, query encoding, and exhaustive score computation, whereas the local dense comparison measures a sequential one-text-at-a-time provider path used to avoid an Ollama batch timeout. Qwen3's higher measured latency therefore characterizes this workstation and execution path. The single run does not isolate the effects of embedding dimension, model size, provider implementation, or hardware utilization.

Under the prespecified descriptive composite rule in Table 10, BGE-M3 ranks first in the local comparison because its near-leading quality is balanced against lower measured runtime than Qwen3. This preference is conditional on the selected weights and sequential local execution path, so it complements rather than replaces the metric-specific result that Qwen3 leads dense MRR.

Retrieval effectiveness is only one operational component of a regulatory compliance pipeline. In production settings, isolated retrieval of statutory chunks entails significant domain risks:

- 1) *Outdated or Amended Regulations:* A retriever may return an obsolete provision if regulatory amendments or circular letters, also known as *Surat Edaran OJK*, are not indexed with active temporal metadata.
- 2) *Cross-Reference Truncation:* Indonesian POJK regulations heavily utilize cross-referencing, such as statutory citations reading "as referred to in Article 4,

paragraph (2)" (Indonesian: "*sebagaimana dimaksud dalam Pasal 4 ayat (2)*").

Fragmented fixed-window chunking risks severing these operative links.

- 3) *Snippet Overreliance & Generation Hallucination*: Users or generative LLMs presented with isolated chunks may draw inaccurate legal conclusions without consulting full statutory context.

A production-grade regulatory RAG system requires document-version monitoring, automated citation verification, hierarchical cross-reference expansion, explicit refusal mechanisms for ambiguous legal queries, and continuous legal expert oversight. The present benchmark deliberately scopes its evaluation to chunk retrieval precision.

The main limitations of this study are the two-document corpus, fixed 160-word non-overlapping chunks, one primary relevant chunk per query, synthetic LLM-generated questions, and lack of end-to-end generation metrics. These constraints bound the interpretation of every metric reported above: the results characterize clause-level retrieval on a small, auditable POJK sample and do not support claims about broader regulatory coverage, graded relevance judgments, or answer generation quality.

4. CONCLUSION

This study demonstrates that multi-signal retrieval improves regulatory search performance on Indonesian insurance regulations. In the native ablation, BGE-M3 combined retrieval achieved the highest retrieval quality (NDCG@10 = 0.8810), while Qwen3 Embedding led the separate local dense comparison (MRR = 0.8075), and under the prespecified descriptive composite rule BGE-M3 ranked first in that local comparison because its near-leading precision was paired with lower measured latency. BM25 and the BM25+dense hybrid confirmed that lexical matching remains essential for exact statutory numbers and defined legal terms. These conclusions are specific to the tested two-document POJK benchmark, the synthetic query set, and the exact exhaustive ranking protocol, and they do not constitute a general deployment recommendation. The methodological contribution is an auditable evaluation framework combining statistical significance testing, practitioner validation of an audit subset, artifact-backed qualitative analysis, and latency decomposition. Future research will extend this benchmark through compliance-officer testing with real-world insurance queries, multi-annotator legal

expert graded relevance judgments, document-version monitoring, broader corpus coverage spanning banking, capital markets, and non-bank financial regulations, neural cross-encoder reranking, and end-to-end generative evaluation of answer faithfulness and citation correctness.

REFERENCES

- [1] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.
- [2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 9459–9474. doi: 10.48550/arXiv.2005.11401.
- [3] H.-T. Nguyen, M.-K. Phi, X.-B. Ngo, V. Tran, L.-M. Nguyen, and M.-P. Tu, "Attentive deep neural networks for legal document retrieval," *Artif. Intell. Law*, vol. 32, no. 1, pp. 57–86, Mar. 2024, doi: 10.1007/s10506-022-09341-8.
- [4] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval Augmented Language Model Pre-Training," in *Proc. 37th Int. Conf. Mach. Learn. (ICML), Proc. Mach. Learn. Res.*, vol. 119. PMLR, Jul. 2020, pp. 3929–3938. doi: 10.48550/arXiv.2002.08909.
- [5] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.
- [6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. and 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.
- [7] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, 2021. doi: 10.48550/arXiv.2104.08663.

- [8] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive Text Embedding Benchmark," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL)*, 2023, pp. 2014–2037. doi: 10.18653/v1/2023.eacl-main.148.
- [9] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Syst. Demonstrations (EACL)*, St. Julians, Malta: Association for Computational Linguistics, 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.
- [10] Y. Huang and J. X. Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," *ACM Comput. Surv.*, vol. 58, no. 12, pp. 1–38, Sep. 2026, doi: 10.1145/3805774.
- [11] L. Wang et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *arXiv preprint arXiv:2212.03533*, 2022, doi: 10.48550/arXiv.2212.03533.
- [12] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking," in *Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval (SIGIR)*, Virtual Event Canada: ACM, Jul. 2021, pp. 2288–2292. doi: 10.1145/3404835.3463098.
- [13] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction," in *Proc. 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (NAACL)*, Seattle, United States: Association for Computational Linguistics, 2022, pp. 3715–3734. doi: 10.18653/v1/2022.naacl-main.272.
- [14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," *arXiv preprint arXiv:2402.03216*, 2024, doi: 10.48550/arXiv.2402.03216.
- [15] J. Alsubhi et al., "Optimizing RAG Pipelines for Arabic: A Systematic Analysis of Core Components," *arXiv preprint arXiv:2506.06339*, 2025, doi: 10.48550/arXiv.2506.06339.
- [16] V.-H. Le, D.-V. Nguyen, V. K. Nguyen, and N. L.-T. Nguyen, "Optimizing Legal Document Retrieval in Vietnamese with Semi-hard Negative Mining," in *Computational Collective Intelligence, Lect. Notes Comput. Sci.*, vol. 16138, Cham: Springer Nature Switzerland, 2026, pp. 269–283. doi: 10.1007/978-3-032-09318-9_19.
- [17] M.-Y. Kim, J. Rabelo, K. Okeke, and R. Goebel, "Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods," *Rev. Socionetwork Strateg.*, vol. 16, no. 1, pp. 157–174, Apr. 2022, doi: 10.1007/s12626-022-00103-1.

- [18] E. M. Voorhees, D. Samarov, and I. Soboroff, "Using Replicates in Information Retrieval Evaluation," *ACM Trans. Inf. Syst.*, vol. 36, no. 2, pp. 1–21, Apr. 2018, doi: 10.1145/3086701.
- [19] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, Dense, and Attentional Representations for Text Retrieval," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 329–345, Apr. 2021, doi: 10.1162/tacl_a_00369.
- [20] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury, "Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling," in *Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval (SIGIR)*, Virtual Event Canada: ACM, Jul. 2021, pp. 113–122. doi: 10.1145/3404835.3462891.
- [21] Y. Zhang et al., "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models," *arXiv preprint arXiv:2506.05176*, 2025, doi: 10.48550/arXiv.2506.05176.
- [22] Z. Nussbaum and B. Duderstadt, "Training Sparse Mixture Of Experts Text Embedding Models," *arXiv preprint arXiv:2502.07972*, 2025, doi: 10.48550/arXiv.2502.07972.
- [23] A. Kusupati et al., "Matryoshka Representation Learning," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 30233–30249, doi: 10.48550/arXiv.2205.13147.