

Multilingual-E5 Size Comparison on Indonesian Public Nutrition Regulation: A Single-Document Case Study

Annisa Desianty^{1*)}, Yulianti²⁾, Hertanto Suryoprayogo³⁾

¹⁾Departement of Informatics, Telkom University, Jakarta, Indonesia

^{2,3)}Departement of Information Technology, Telkom University, Jakarta, Indonesia

¹⁾Center of Excellence for Technological Society, Universitas Telkom, Kampus Jakarta

²⁾Center of Excellence of Artificial Intelligence for Learning and Optimization, Telkom University, Bandung, Indonesia

³⁾PUI-PT Intelligent Sensing-IoT, Research Institute of Sustainable Society, Universitas Telkom, Kampus Jakarta

¹⁾annisadesianty@telkomuniversity.ac.id, ²⁾yulianti@telkomuniversity.ac.id,
³⁾hertantosp@telkomuniversity.ac.id

Submitted: Aug 31, 2026 | **Accepted:** Oct 11, 2026 | **Published:** Oct 13, 2026

Abstract: Rigid structure and complex legalese frequently make it difficult to access Indonesian legal documents, such as Ministry of Health rules, and traditional Large Language Models are still prone to hallucinations. This problem is mitigated by RAG, but its accuracy is highly dependent on the semantic quality of the dense retriever, an aspect that is rarely investigated for Indonesian regulatory text, where previous studies either tested different model families on non-Indonesian corpora or used generic embeddings without efficiency trade-off analysis. In order to ascertain whether model size influences retrieval accuracy and answer quality, this study empirically benchmarks two variants of the same embedding family, multilingual-e5-large and multilingual-e5-base, within a RAG pipeline for question answering on Indonesian public nutrition law across three retrieval depths ($k = 3, 5, 7$). In addition to Context Relevance, Faithfulness, and Answer Relevance assessed by an LLM, 22 test questions were assessed across 132 combinations utilizing Precision, Recall, and F1-Score against ground truth. Wilcoxon signed-rank, Friedman, post-hoc, and non-inferiority tests were then used for analysis. Multilingual-e5-large and multilingual-e5-base did not differ significantly on any of the six metrics ($p > 0.05$), although k had a significant impact on Precision, Recall, and F1-Score ($p < 0.05$) but not the LLM-judge measures. The smaller multilingual-e5-base was non-inferior to multilingual-e5-large within a margin of 0.05 on all metrics except Recall and Faithfulness, and k can follow precision or recall priorities. Because the study covers a single regulation and computational cost was not measured, the findings should not be generalized to Indonesian legal corpora.

Keywords: Dense Retrieval; Embedding Model; Indonesian Health Regulation; Multilingual-E5; Retrieval-Augmented Generation

INTRODUCTION

A key component of the health system that aims to improve community well-being through planning, decision-making, and the execution of health programs is accessibility to public health regulations, particularly legal documents pertaining to nutrition guidelines (like Ministry of Health Regulations or Permenkes). Regrettably, these kinds of legal papers are often lengthy, rigidly structured, and rife with formal lexical difficulties (Rina et al., 2025). However, utilizing traditional Large Language Models (LLMs) to create autonomous Q&A systems frequently encounters a crucial problem known as hallucination, in which the model produces material that appears grammatically correct but is not valid in terms of law or medicine (Gao et al., 2024). This inability to comprehend certain domains necessitates changes in AI system architecture to make them more accurate and trustworthy.

*Author Correspondence



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

The Retrieval-Augmented Generation (RAG) architecture is now acknowledged as the gold standard for addressing the shortcomings of generative models. By including an Information Retrieval process that searches and retrieves particular external documents before feeding them as reference context to the LLM, the RAG framework can reduce the degree of hallucinations (Zhao et al., 2026) (Karpukhin et al., 2020). However, this retrieval stage is crucial to the RAG architecture's success. Based on the calculation principle of garbage in, garbage out, the LLM will continue to generate wrong responses if the search engine (retriever) is unable to filter out irrelevant phrases (noise) from the sea of documents (Es et al., 2024). Consequently, the most important factor in establishing the system's overall accuracy is the semantic-based search component (Dense Retriever).

Several studies on AI systems for Indonesian regulatory text have concentrated on the generation layer. (Febrian & Figueredo, 2024) improved an LLM for Indonesian government financial regulations through RAG, prompt engineering, and fine-tuning, and (Palupi & Irawan, 2025) compared IndoGPT with GPT-4 and GPT-4o in a RAG chatbot for university regulations. Other work still relies on sparse lexical retrieval: (Rakhmawati et al., 2023) built a question answering system for the Indonesian Constitution using TF-IDF with cosine and Jaccard similarity, and (Faisal et al., 2024) used BM25 as the first-stage retriever for Indonesian government regulations.

Dense retrieval has received less attention. To the best of our knowledge, (Rozaq et al., 2025) is among the few studies applying a dense retriever (IndoSBERT) to Indonesian legal text; although it informally contrasted IndoSBERT with IndoBERT and FastText on two case examples, it did not systematically evaluate alternative embedding models or the trade-off between model size and retrieval accuracy. (Faisal et al., 2024) also tested dense retrievers on Indonesian government regulations, but these were general-purpose IndoBERT models that underperformed BM25 and were not compared by model size. (Rahmadi & Setyanto, 2026) adopted multilingual-e5-large-instruct as the embedding model without comparing it against alternative models or model sizes. Likewise, (Ongri et al., 2025) benchmarked knowledge-graph-based RAG frameworks on regulations including Indonesian government regulations, where embedding choice was secondary and only one comparison (Contriever vs. mContriever) was reported within a single framework and one document. Consequently, the efficiency-accuracy trade-off among Transformer-based embedding models for Indonesian legal language remains largely unexplored, which may lead to inefficient use of computational resources.

This study examines the embedding layer of the RAG pipeline. It compares two variants of the same multilingual embedding family, multilingual-e5-large and multilingual-e5-base (Wang et al., 2024), which share a training recipe and differ mainly in model size and vector dimensionality (1024 vs. 768). Holding the corpus, queries, chunking, generator, and evaluation metrics constant isolates the effect of model size. Cross-family comparisons such as (Rao et al., 2025), which used different architectures and English corpora, cannot do this. The evaluation uses Indonesian public nutrition regulation, a domain that previous embedding evaluations have not covered. Retrieval quality is measured with Precision, Recall, and F1-Score (Sembiring & Santoso, 2026). against ground truth, and with LLM-judged Context Relevance, Faithfulness, and Answer Relevance (Baur et al., 2025) (Es dkk., 2024), at three retrieval depths ($k = 3, 5, 7$).

To the best of our knowledge, this study makes three contributions. First, it compares dense embeddings for Indonesian regulatory text under controlled conditions that isolate model size, whereas existing Indonesian legal retrieval studies either rely on lexical methods or evaluate embedding models only informally, without controlling for model size. Second, it shows how retrieval depth affects ground-truth-based and LLM-judge-based metrics differently. Third, it tests whether the smaller base variant is non-inferior to the large variant and reports vector dimensionality as an indicator of storage cost; computational cost was not measured. This study aims to (1) empirically compare the retrieval performance of multilingual-e5-large and multilingual-e5-base on Indonesian public nutrition law documents, (2) evaluate the effect of retrieval depth (k) on precision-based and LLM-judge-based metrics, and (3) assess whether the smaller multilingual-e5-base is a non-inferior alternative to multilingual-e5-large for this domain.

LITERATURE REVIEW

Summary of Related Work

This subsection reviews prior work that bears directly on the research problem: how the size of a dense embedding model affects retrieval quality in a RAG system for Indonesian regulatory text. Studies published between 2023 and 2026 were included if they met at least one of two criteria: (i) they introduced or evaluated embedding models, particularly with respect to model size, or (ii) they built retrieval or RAG systems for Indonesian legal or regulatory documents. The first group (Wang et al., 2024) (Rao et al., 2025) shows what is known about embedding models and size trade-offs in general, while the second group shows what has been done on Indonesian legal text. Table 1 summarizes each study by domain, method, and key findings, and the discussion that follows uses it to identify what remains untested.

Table 1 Summary of Related Studies

Study	Dataset/Domain	Method	Key Findings
(Wang et al., 2024)	Multilingual text pairs (100+ languages)	Contrastive pre-training on the XLM-RoBERTa-large architecture	Introducing the multilingual-e5 model family, like the large and base variants, which are trained with a large-scale contrastive approach, but without evaluating the trade-offs of model size in the legal/regulatory domain.
(Rao et al., 2025)	SciFact, NFCorpus	FIQA, Hybrid retrieval (dense + sparse + graph) with LLM re-ranking	Small embedding models (MiniLM-v6) can outperform large models (BGE-Large), but the tests were done on different models and domains (not multilingual-e5 and not Indonesian legal corpus)
(Rakhmawati et al., 2023)	Indonesian Constitution	Question answering with TF-IDF, cosine similarity, and Jaccard similarity	Relies on lexical matching only; dense embeddings were not evaluated.
(Faisal et al., 2024)	Indonesian government regulations	Granularity-aware legal QA with BM25 retrieval and a BERT re-ranker	BM25 outperformed the Siamese-BERT retrievers (exact match 29.73% vs. 4.86–15.68%), and the reranker raised exact match to 35.68%. The authors attribute the weak dense results partly to pretrained models not trained for retrieval; embedding models were not compared by size.
(Rozaq et al., 2025)	Indonesian Civil Code (2,074 articles)	IndoSBERT embeddings (256-dimensional) with Qdrant and cosine similarity	Recommendation accuracy of 76.66% by legal experts. IndoBERT and FastText were contrasted informally on two case examples, without systematic metrics or a model-size comparison.
(Rahmadi & Setyanto, 2026)	Regional regulations (393 regulations)	Fine-tuned Llama 3.2-1B with RAG (FAISS, multilingual-e5-large-instruct)	Focuses on the generator; the embedding model was fixed and not compared with alternatives.
(Pratama et al., 2025)	Indonesian Criminal Code	RAG chatbot with LLama	Tested on 10 questions (80% hit rate); the embedding model was neither compared nor evaluated.
(Ongri et al., 2025)	EU Directives and Indonesian government regulations	KG-based and hybrid RAG frameworks vs. naive RAG; Ragas answer accuracy	Hybrid frameworks were most accurate, and naive RAG was fast and competitive. The only embedding comparison (Contriever vs. mContriever) was limited to one framework and one document.
This study	Indonesian public nutrition regulation (Permenkes No. 41 of 2014; 22 questions)	Controlled comparison of multilingual-e5-large and multilingual-e5-base in a RAG pipeline at k = 3, 5, 7; Precision, Recall, F1-Score, Context Relevance, Faithfulness, Answer Relevance	Addresses the gap left by the studies above: a controlled comparison of embedding model size within one multilingual family, on Indonesian regulatory text, with both ground-truth and LLM-judge metrics.

In most of the Indonesian legal studies in Table 1, the embedding model is fixed in advance rather than studied. Two rely mainly on lexical retrieval (Faisal et al., 2024; Rakhmawati et al., 2023), two build on one main model (Rahmadi & Setyanto, 2026; Rozaq et al., 2025), one does not specify its model (Pratama et al., 2025), and one uses the default of each framework (Ongri et al., 2025). (Faisal et al., 2024) did test dense retrievers built on two general-purpose IndoBERT checkpoints, but these underperformed BM25, and the authors attributed this partly to

pretrained models that were not trained for retrieval. None of these studies can therefore show whether a smaller or larger embedding model of the same family would have changed the outcome. Their evaluation designs also make results hard to attribute to the embedding itself. Expert-judged accuracy on a single code (Rozaq et al., 2025) lacks a systematic comparison, end-to-end answer accuracy scored by an LLM (Ongri et al., 2025) mixes the contribution of the retriever with that of the generator, and a hit rate on ten questions (Pratama et al., 2025) rests on a small test set.

The general embedding literature leaves a different gap. (Wang et al., 2024) introduce the multilingual-e5 family but do not report how its variants differ on regulatory text. (Rao et al., 2025) show that smaller embedding models can match or beat larger ones, but their comparison is confounded by differences in architecture, hybrid retrieval, and LLM re-ranking, and it uses English corpora. It therefore remains unresolved whether model size alone matters for retrieving Indonesian regulatory text, and how this interacts with the number of retrieved chunks. None of the Indonesian studies in Table 1 targets public health or nutrition regulation either. This study addresses these gaps. On Indonesian public nutrition regulation, it compares two sizes of one embedding family under identical conditions, scores retrieval directly with Precision, Recall, and F1-Score alongside LLM-judged metrics, and varies retrieval depth.

METHOD

Fig. 1 summarizes the experimental workflow. The Permenkes No. 41/2014 document is loaded and split into 600-character chunks with a 100-character overlap, and each chunk keeps its page number. The same chunks are encoded separately by multilingual-e5-large and multilingual-e5-base and stored in one Chroma collection per model. For each of the 22 test questions, the system retrieves the top-k chunks with $k = 3, 5$, and 7 , which gives six model-k combinations. A single generative model, identical across all combinations, produces an answer from the retrieved chunks. Each combination is evaluated with ground-truth metrics computed from page matches (Precision, Recall, and F1-Score) and with LLM-judge metrics (Context Relevance, Faithfulness, and Answer Relevance), which yields 132 data points. These data points are then compared with Wilcoxon signed-rank tests, Friedman tests with post-hoc comparisons, and non-inferiority tests.

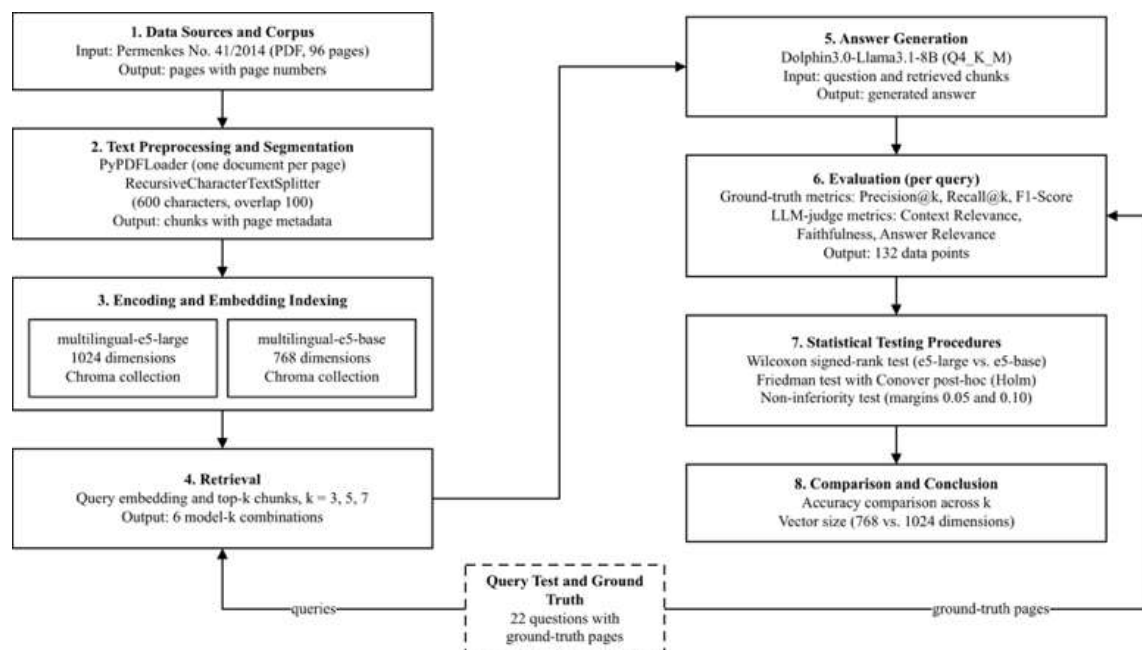


Fig. 1 Research Methodology Flowchart

Research Design

This study uses a quantitative experimental approach to compare the performance of two dense embedding model variants, namely multilingual-e5-large and multilingual-e5-base, in a Retrieval-Augmented Generation (RAG) architecture for Indonesian public nutrition law document question-answering tasks. Both models are tested on identical corpora, queries, generative components, and evaluation metrics, so any differences in results can be attributed to differences in embedding model size (Yafie et al., 2026). As an additional variable, three k values (the number of chunks retrieved) are also compared for each model, resulting in six experimental combinations.

* Author Correspondence



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Data Sources and Corpus

The document corpus is sourced from the Indonesian Ministry of Health Regulation Number 41 of 2014 concerning Guidelines for Balanced Nutrition, an official publication by the Ministry of Health of the Republic of Indonesia, consisting of 96 pages. This document includes the main body of the regulation as well as technical appendices, organized into six chapters: Chapter I Introduction, Chapter II Principles of Balanced Nutrition, Chapter III Balanced Nutrition Messages, Chapter IV Slogans and Visuals, Chapter V Strategies and Implementation of Information and Education Communication, and Chapter VI Closing. This document was chosen because it contains characteristics of Indonesian legal language with dense conditional clauses and specific technical terminology, as explained in the Literature Review section.

Text Preprocessing and Segmentation

PDF documents are loaded using PyPDFLoader from the LangChain library. The loaded text is segmented into chunks using RecursiveCharacterTextSplitter with a chunk size of 600 characters and an overlap of 100 characters between consecutive chunks, so as to preserve context between consecutive chunks of the same page. Each PDF page is loaded as a separate document that carries its page index, and the splitter is applied to each page separately. Every chunk therefore belongs to exactly one page and inherits that page number as metadata, converted to a 1-based number so that it matches the page numbers shown in the PDF viewer. Chunks never cross page boundaries: a passage that continues from one page to the next is split into separate chunks on each page, and the 100-character overlap applies only within a page. The ground truth of each question is the set of pages on which the information needed to answer it appears (Table 3), and a range of pages is counted page by page. A retrieved chunk is judged at the page level, that is, through the page it belongs to, and several retrieved chunks from the same page count as one relevant page. This mapping does not check whether the retrieved chunk itself contains the answer text, so a chunk from a correct page that does not contain the answer is still counted as relevant, and an answer that spans two pages requires a chunk from each page to be retrieved for full Recall.

Encoding and Embedding Indexing

Each document chunk is encoded into a vector representation using two separate variants of embedding models. The embedding models used are multilingual-e5-large (intfloat/multilingual-e5-large, 1024 dimensions) and multilingual-e5-base (intfloat/multilingual-e5-base, 768 dimensions), accessed via the HuggingFaceEmbeddings interface in LangChain. The vector representations of each model are stored in separate Chroma collections and given unique names per model to avoid vector dimension conflicts between the two models during the indexing process. At the retrieval stage, the system is tested with three different k values, namely k = 3, 5, and 7, to determine the impact of the number of retrieved chunks on the performance of each embedding model.

The search for semantic similarity between the query embedding q and the embedding of document chunk d_i is performed internally by the Chroma vector store using cosine similarity:

$$\cos(q, d_i) = \frac{q \cdot d_i}{|q| \cdot |d_i|} \quad (1)$$

which produces a score in the range $[-1, 1]$, with a value of 1 indicating identical vector directions. The set of context results from retrieval for each query is defined as:

$$C = \{d_{(1)}, d_{(2)}, \dots, d_{(k)}\} \quad (2)$$

with $d_{(1)}$ being the chunk with the highest similarity score, $d_{(2)}$ the second highest, and so on up to the k-th chunk.

Answers were generated with Dolphin3.0-Llama3.1-8B (GGUF, Q4_K_M quantization) run through llama.cpp, with a temperature of 0.1, a maximum of 1,024 new tokens, a context window of 4,096 tokens, and a repetition penalty of 1.2. A fixed Indonesian-language prompt instructed the model to answer concisely, to use only the retrieved context, and to cite page numbers. The same generation settings were used for all six combinations. Because generation uses sampling, answers differ between runs; the retrieval metrics are deterministic and were reproduced exactly in a second full run of the pipeline, and all LLM-judge results reported here come from the final run.

Query Test and Ground Truth

Twenty-two test questions were prepared based on the contents of the attachment of Ministry of Health Regulation Number 41 of 2014. The questions cover pages 3 to 84 of the 96-page document and include definitions, lists, numerical recommendations, explanations, comparisons, and a calculation formula; 14 of the 22 questions have a single ground-truth page, and the other eight have between two and twenty. The questions were prepared through discussion within the research team, which included a nurse working in a hospital, and were



drawn from the content of the regulation's appendices so that they reflect information needs in practice. The nurse reviewed the questions for relevance. Because the question set was selected by a small team, it may not represent all information in the regulation. Text was extracted from the PDF with PyPDFLoader only, without optical character recognition or image analysis, so only the text layer of the document was available to the pipeline. The questions were therefore designed to be answerable from the text, and information that appears only in figures was not used.

Evaluation

The performance of each combination is evaluated using two sets of metrics calculated simultaneously for each query. The ground-truth-based metric group—comprising Precision@k, Recall@k, and F1-Score is calculated based on the match between the page numbers of the retrieved chunks and the ground truth, while the LLM-as-judge metric group such as Context Relevance, Answer Faithfulness, and Answer Relevance is adapted from the framework by (Khalila et al., 2025)(Yu et al., 2025) and are calculated by the same generative model (Dolphin3.0-Llama3.1-8B), which acts as an automated judge.

Precision and recall at rank k are calculated as:

$$Precision@k = \frac{| \{d(1), \dots, d(k)\} \cap R |}{k} \quad (3)$$

$$Recall@k = \frac{| \{d(1), \dots, d(k)\} \cap R |}{|R|} \quad (4)$$

where the numerator counts the distinct ground-truth pages that appear among the pages of the k retrieved chunks, so several chunks from the same page count once, and R is the set of ground truth pages. The F1-Score is calculated as the harmonic mean of Precision and Recall. Context Relevance and Faithfulness are calculated as:

$$Context\ Relevance = \frac{\text{Number of Relevant Context Sentences}}{\text{Total context sentences}} \quad (5)$$

$$Faithfulness = \frac{\text{Number of claims supported by context}}{\text{Total Claims on Answers}} \quad (6)$$

while Answer Relevance is obtained directly from the score provided by the judge model (range 0–1), without additional calculations. Since the judge model uses the same generative model as the generative component, the evaluation results are potentially biased because the model evaluates its own output this limitation is discussed further in the Discussion chapter. The judge was run at temperature 0 with a fixed seed and a grammar that forces the five numbers into a fixed format. It received the full retrieved context and the answer (truncated at 1,500 characters), and each output was validated: faithful claims could not exceed total claims, relevant sentences could not exceed total sentences, and the relevance score had to lie between 0 and 1. Invalid outputs were retried up to twice at temperature 0.3, and an output that was still invalid would have been recorded as missing, not as zero; in the final run all 132 outputs were valid at the first attempt. A first version of the judge, which truncated the context to 1,500 characters and extracted the first number found in free-form output, was replaced for the reasons given in the Data Quality Findings.

Statistical Testing Procedures

The comparison between multilingual-e5-large and multilingual-e5-base across each metric was tested using a paired Wilcoxon signed-rank test that is, by pairing them based on the same query and the same value of k—at a significance level of $\alpha = 0.05$. The effect of the k value on each metric for each model was tested using the Friedman test (a nonparametric test for repeated comparisons of more than two conditions), given that the data consisted of repeated measurements on the same query across three k values. As a robustness check that avoids pooling across k, the differences were also averaged over k for each query ($n = 22$), and 95% confidence intervals of the mean difference were obtained by bootstrapping (10,000 resamples). The pooled analysis across k values ($n = 66$ pairs) is reported only as an exploratory comparison, because observations from the same query are not independent; the primary inference treats the 22 queries as the independent units. For the same reason, the models were also compared separately at each k ($n = 22$ per test), with Holm correction across the 18 resulting tests. When the Friedman test was significant, pairwise differences between k values were examined with the Conover post-hoc test with Holm correction. Because non-significance does not establish equivalence, we also tested whether multilingual-e5-base is non-inferior to multilingual-e5-large with a one-sided paired t-test on the per-query differences averaged over k, using margins of 0.05 and 0.10 on the 0 to 1 scale, and we report the one-sided 95% bootstrap lower bound of the difference. These margins were set after seeing the data, so they express a judgment about practical relevance and not pre-specified thresholds.

*Author Correspondence



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

RESULT

The experiment was conducted on 132 data points (22 queries $\times$ 2 embedding models $\times$ 3 values of k). Table 2 summarizes the mean and standard deviation of each metric per combination.

Table 2 Mean ($\pm$ SD) Precision, Recall, F1, and LLM-Judge Metrics by Model and k Value

Model	k	Precision	Recall	F1	Context Relevance	Faithfulness	Answer Relevance
e5-large	3	0.348 $\pm$ 0.162	0.737 $\pm$ 0.379	0.433 $\pm$ 0.185	0.738 $\pm$ 0.142	0.880 $\pm$ 0.160	0.810 $\pm$ 0.156
e5-large	5	0.245 $\pm$ 0.137	0.824 $\pm$ 0.326	0.347 $\pm$ 0.146	0.771 $\pm$ 0.157	0.900 $\pm$ 0.154	0.778 $\pm$ 0.148
e5-large	7	0.182 $\pm$ 0.100	0.847 $\pm$ 0.320	0.278 $\pm$ 0.124	0.744 $\pm$ 0.136	0.912 $\pm$ 0.137	0.808 $\pm$ 0.095
e5-base	3	0.333 $\pm$ 0.145	0.735 $\pm$ 0.351	0.430 $\pm$ 0.149	0.756 $\pm$ 0.144	0.915 $\pm$ 0.120	0.791 $\pm$ 0.138
e5-base	5	0.236 $\pm$ 0.133	0.822 $\pm$ 0.294	0.344 $\pm$ 0.126	0.797 $\pm$ 0.146	0.842 $\pm$ 0.252	0.791 $\pm$ 0.157
e5-base	7	0.175 $\pm$ 0.098	0.845 $\pm$ 0.287	0.275 $\pm$ 0.111	0.746 $\pm$ 0.211	0.891 $\pm$ 0.139	0.791 $\pm$ 0.146

Across all 132 data points, the overall averages are Precision = 0.253, Recall = 0.802, F1 = 0.351, Context Relevance = 0.759, Faithfulness = 0.890, and Answer Relevance = 0.795. For both models, Precision falls and Recall rises as k increases (Table 4), which follows from how the two metrics are built. Precision is the number of distinct relevant pages among the k retrieved chunks divided by k, so each extra chunk lowers it unless it brings a new relevant page. Recall divides the same count by the number of ground-truth pages, so it cannot fall. Because the numerator is capped by the number of ground-truth pages, this ceiling matters: 14 of the 22 questions have a single ground-truth page (Table 3), and for these questions Precision equals exactly 1/k whenever the page is retrieved, which happened for 13 of 14 questions at k = 3 and for all 14 at k = 5 and k = 7 in both models. Recall also levels off. For e5-large, it rises by 0.087 from k = 3 to k = 5 but by only 0.023 from k = 5 to k = 7, while Precision keeps falling (0.348, 0.245, 0.182). Most relevant pages therefore appear within the top five chunks, and the sixth and seventh chunks mostly add non-relevant material. F1-Score declines for the same reason, since Precision falls faster than Recall rises, and it is highest at k = 3 (0.433 for e5-large and 0.430 for e5-base). On this corpus, k = 7 adds little coverage at the cost of a longer context for the generator. The three LLM-judge metrics (Context Relevance, Faithfulness, Answer Relevance) show no clear upward or downward trend with k.

The Effect of k Value (Friedman Test)

Table 3 shows that the k value has a statistically significant effect on Precision, Recall, and F1-Score for both embedding models ($p < 0.05$), but it does not have a significant effect on the three LLM-judge metrics (Context Relevance, Faithfulness, Answer Relevance across all $p > 0.05$). This finding supports the descriptive pattern in Table 2 the k value noticeably affects page-based retrieval metrics, but it does not impact the semantic quality assessment of answers by the LLM-judge within the tested range of k=3, 5, 7.

Table 3 Friedman Test Results for the Effect of k Values (n=22 queries, df=2)

Metric	χ^2 (e5-large)	p (e5-large)	χ^2 (e5-base)	p (e5-base)
Precision	32.667	<0.001 *	32.667	<0.001 *
Recall	8.400	0.015 *	8.400	0.015 *
F1-Score	20.667	<0.001 *	20.667	<0.001 *
Context Relevance	0.571	0.751	0.531	0.767
Faithfulness	0.519	0.772	1.056	0.590
Answer Relevance	2.235	0.327	1.105	0.575

The two metric groups differ in what they measure. Precision, Recall, and F1-Score are direct functions of k, so a significant Friedman result is almost guaranteed. The LLM-judge metrics score the content of the context and the answer, and the revised judge, which received the full retrieved context at every k, still showed no effect of k (all $p > 0.3$), so this null result is not an artifact of truncated context. Three features of our data make a k effect hard to detect. First, the judge is lenient where page matching is strict. The document repeats content across chapters, for example the four pillars on page 7 (Question 2) and page 17 (Question 19), and the definition of balanced nutrition on pages 3 and 8 (Questions 1 and 22), so a chunk from a different page count as an error for Precision but can be fully relevant for the judge. Second, the judge scores are compressed near the top of the scale: 81 of the 132 Faithfulness scores equal 1.0, which leaves little room for any factor to raise them. Third, the judge is only moderately reliable. When the same 44 rows were scored twice at temperature 0.1, the Spearman correlations between the two scorings were 0.21 for Context Relevance, 0.71 for Faithfulness, and 0.52 for Answer Relevance, and the judge scores were unrelated to retrieval F1-Score (Spearman correlations between 0.01 and 0.06). With only 22 queries, the Friedman test also has little power. The result should therefore be read as no detectable effect in this sample, not as proof that k does not matter.

Comparison of Multilingual-E5-Large vs Multilingual-E5-Base (Wilcoxon Test)

Based on Table 4, no statistically significant differences were found between multilingual-e5-large and multilingual-e5-base on any of the six metrics tested (all $p > 0.05$), whether tested by pooling across k -values or separately per k -value (per- k results were consistent, not shown separately because the pattern is similar). The average differences across all metrics were below 0.02, well below the standard deviation of each metric (ranging from 0.10–0.38 in Table 3), indicating that the variation between queries is much greater than the variation between embedding models.

Table 4 Result Wilcoxon Signed-Rank Test, e5-large vs e5-base

Metric	Average Difference (large-base)	p-value	Significant ($\alpha=0.05$)
Precision	+0.0102	0.444	No
Recall	+0.0023	0.786	No
F1-Score	+0.0037	0.790	No
Context Relevance	-0.0151	0.520	No
Faithfulness	+0.0145	0.612	No
Answer Relevance	+0.0080	0.648	No

No significant difference means that, in this setting, the larger model gave no measurable benefit in retrieval or answer quality. The mean differences are at most 0.02 (Table 4), and the 95% confidence intervals of the per-query differences (Table 5) lie within ± 0.07 for all six metrics, so the data rule out differences larger than about 0.07 in either direction. Two features of the test set explain why the models look alike. First, retrieval is nearly saturated: both models found the single relevant page for 13 of the 14 single-page questions at $k = 3$ and for all 14 at $k = 5$, which leaves little room for a stronger embedding to show an advantage. Second, the models fail on different questions that offset each other: e5-large scored a higher mean F1-Score on five questions and e5-base on two, and the per-query F1-Scores of the two models correlate at $r = 0.73$, so equal averages hide different failure patterns. In practical terms, e5-base produces 768-dimensional instead of 1024-dimensional vectors, so each stored vector is 25% smaller. Encoding time, query latency, and memory use were not measured, so we make no claim about computational efficiency beyond vector size, and the non-inferiority tests in Table 9 address accuracy only. The finding comes from one document and a test set on which retrieval was nearly saturated (both models retrieved the relevant page for all 14 single-page questions at $k = 5$), so harder questions, such as paraphrased or multi-page ones, or larger and more heterogeneous corpora, may separate the models.

Table 5 Per-Query Comparison of e5-large vs e5-base (differences averaged over k ; $n = 22$ queries; 95% bootstrap CI, 10,000 resamples)

Metric	Average Difference (large-base)	95% CI (bootstrap)	p-value (Wilcoxon, $n=22$)
Precision	+0.0102	[-0.024, +0.043]	0.497
Recall	+0.0023	[-0.068, +0.065]	0.799
F1-Score	+0.0037	[-0.037, +0.039]	0.799
Context Relevance	-0.0151	[-0.066, +0.032]	0.557
Faithfulness	+0.0145	[-0.037, +0.065]	0.509
Answer Relevance	+0.0080	[-0.026, +0.043]	0.778

Tested separately at each k ($n = 22$ per test), none of the 18 comparisons between the two models was significant (smallest unadjusted $p = 0.350$; all Holm-adjusted $p = 1.000$). Because the Friedman test was significant for Precision, Recall, and F1-Score, pairwise differences between k values were examined with the Conover post-hoc test with Holm correction (Table 6).

Table 6 Post-Hoc Pairwise Comparisons of k Values (Conover Test with Holm Correction; Identical for Both Models)

Metric	Pair of k	Mean ranks (first, second)	Adjusted p (Holm)
Precision	3 vs 5	2.82, 2.05	<0.001
Precision	3 vs 7	2.82, 1.14	<0.001
Precision	5 vs 7	2.05, 1.14	<0.001
Recall	3 vs 5	1.80, 2.07	0.044
Recall	3 vs 7	1.80, 2.14	0.015
Recall	5 vs 7	2.07, 2.14	0.555
F1-Score	3 vs 5	2.59, 2.14	0.045
F1-Score	3 vs 7	2.59, 1.27	<0.001

* Author Correspondence



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Metric	Pair of k	Mean ranks (first, second)	Adjusted p (Holm)
F1-Score	5 vs 7	2.14, 1.27	0.001

The post-hoc comparisons show that Precision and F1-Score decreased significantly between every pair of k values. Recall increased significantly from k = 3 to k = 5 and from k = 3 to k = 7 but not from k = 5 to k = 7, which confirms that the gain in Recall levels off after k = 5. The results were identical for the two models because the pattern of ranks across k was the same: the models returned identical Precision and Recall in 55 of the 66 query-k combinations.

Because a non-significant difference does not establish equivalence, we tested whether multilingual-e5-base is non-inferior to multilingual-e5-large using the per-query differences (base minus large, averaged over k) against margins of 0.05 and 0.10 (Table 7).

Table 7 Non-Inferiority Test of multilingual-e5-base Relative to multilingual-e5-large (Per-Query Differences Averaged over k; n = 22)

Metric	Mean difference (base – large)	One-sided 95% lower bound	p (margin 0.05)	p (margin 0.10)
Precision	-0.0102	-0.039	0.017	<0.001
Recall	-0.0023	-0.055	0.092	0.005
F1-Score	-0.0037	-0.034	0.016	<0.001
Context Relevance	+0.0151	-0.025	0.009	<0.001
Faithfulness	-0.0145	-0.057	0.099	0.002
Answer Relevance	-0.0080	-0.037	0.016	<0.001

Multilingual-e5-base was non-inferior to multilingual-e5-large within a margin of 0.05 for Precision, F1-Score, Context Relevance, and Answer Relevance (one-sided p between 0.009 and 0.017), but not for Recall (p = 0.092) or Faithfulness (p = 0.099), for which the lower bounds of the difference were -0.055 and -0.057. Within the wider margin of 0.10, non-inferiority held for all six metrics (all p ≤ 0.005). Non-inferiority does not mean that the models are interchangeable on every query, and the margins were chosen after seeing the data.

Extreme Cases: Best and Worst Performance

Eleven of the 132 combinations had an F1-Score of 0.1 or lower, and they come from four questions (Table 10). Three mechanisms explain them. First, content that is repeated on several pages produces errors for page-based metrics even when the retrieved passage is correct. The four pillars (Question 2, ground truth on page 7) are also described on page 17 (Question 19), and e5-large ranked page 17 first at k = 3. The appendix definition of balanced nutrition (Question 22, page 8) is also given in the body of the regulation on page 3, which e5-base ranked first at k = 3, probably because the word that separates the two questions (“appendix”) has little weight in a dense embedding. Second, the question on the ten general messages has a ground truth of 20 pages (31 to 50), so even at k = 7 at most seven of those pages can be retrieved and Recall cannot exceed 0.35. e5-large ranked the opening chunk of that section (page 31) first at every k but found no other page of it, whereas e5-base ranked passages from earlier chapters first (pages 6, 17, and 7), probably because they discuss the messages of the guidelines in general terms. Third, for the question on single-food diets (Question 20, pages 17 and 18), e5-large ranked two chunks from page 80 first, a passage that restates the conclusion that eating from all four food groups meets nutritional needs and so overlaps strongly with the wording of the question. At the other end, the best case was the carbohydrate-rich staple foods question (Precision = 0.667, Recall = 1.000, F1 = 0.800 for e5-large at k = 3), which has a short and specific answer on two adjacent pages (8 and 9). The wide spread across queries (Table 4, with standard deviations up to 0.38 for Recall) therefore reflects how differently the questions are tied to the document, not random noise in the models. For the RAG system, this indicates that aggregate questions that need many scattered passages are a weak point of fixed-size top-k retrieval, and that repeated content would call for deduplication or for evaluation at the level of the answer instead of the page. The causes in Table 8 were assigned by reading the retrieved passages and were not tested experimentally.

Table 8 Analysis of the Retrieval Failures (F1-Score ≤ 0.1)

Question	Model, k	Ground-truth pages	Top-ranked pages	Likely cause
Q22: definition of Balanced Nutrition according to the appendix	e5-base, k = 3	8	3, 81, 80	The same definition appears in the body of the regulation (page 3); only the word “appendix” separates the questions
Q4: ten general messages	e5-base, k = 3, 5, 7	31 to 50 (206, 17, 7 pages)	206, 17, 7	Ground truth spread over 20 pages; earlier passages about the guidelines ranked first
Q4: ten general messages	e5-large, k = 3, 5, 7	31 to 50 (2031, 77, 75 pages)	2031, 77, 75	Section opening found at rank 1, but Recall is capped at one page of 20

* Author Correspondence



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Question	Model, k	Ground-truth pages	Top-ranked pages	Likely cause
Q20: why one food cannot meet nutritional needs	e5-large, k = 3, 5, 7	17, 18	80, 80, 5	A summary statement on page 80 outranks the original explanation
Q2: four pillars	e5-large, k = 3	7	17, 82, 17	The same content is described on page 17 (Question 19)

Data Quality Findings in the LLM-as-Judge Scheme

An earlier version of the judge extracted the first number found in free-form output after a prompt that contained a worked example. In a first run it produced degenerate output in 5 of 132 rows (3.8%): the model repeated the example text, and the extracted numbers probably came from that example and not from an evaluation of the answer. The failures occurred in both embedding models (2 rows in e5-base, 3 rows in e5-large) at $k = 3$ and $k = 5$. The same check found none in a second run, so the failure is intermittent. The extraction function also returned 0 whenever a field was missing, so a failed extraction could not be told apart from a true score of zero. We therefore replaced the judge: the output format is forced by a grammar, the example with numbers was removed from the prompt, the full retrieved context is provided, every output is validated and, if necessary, retried, and unresolved cases are recorded as missing and not as zero. With the revised judge, all 132 outputs were valid at the first attempt, none was missing, and none reproduced the example. The scores of the two judge versions were unrelated (Spearman correlations between -0.10 and 0.00), which indicates that the earlier scores carried little information about the answers; all LLM-judge results in this article therefore come from the revised judge.

DISCUSSIONS

The Wilcoxon test results in Table 6 indicate that multilingual-e5-large does not significantly outperform multilingual-e5-base on any of the six metrics tested, despite the model size and embedding dimensions differing considerably (1024 vs. 768 dimensions). With 22 queries and formal statistical testing, this finding is methodologically stronger compared to a mere descriptive comparison. Substantively, the results suggest that for retrieval tasks on a single, small- to medium-sized corpus (one Ministry of Health regulation document, totaling 96 pages) with questions that have relatively explicit references within the document, the additional representational capacity of the large variant does not translate into measurable advantages in retrieval performance or answer quality. Non-inferiority tests (Table 9) show that multilingual-e5-base stays within 0.05 of multilingual-e5-large on all metrics except Recall and Faithfulness, and within 0.10 on all six.

Trade-off Between Precision and Recall at k Values, with No Detectable Effect on Answer Quality

The results of the Friedman test in Table 5 confirm that the k value significantly affects Precision, Recall, and F1-Score in both models, as theoretically expected: increasing k definitely lowers Precision (denominator grows) and tends to raise Recall (higher chance of finding relevant chunks). However, k does not have a significant impact on Context Relevance, Faithfulness, or Answer Relevance. This suggests that adding context chunks in the range of $k=3, 5, 7$ does not automatically make the LLM's answers more (or less) faithful or semantically relevant the final answer quality seems to be more determined by the model's generative ability and the relevance of the top chunks, rather than the number of additional chunks included. This result should be read with two caveats. First, the judge scores are compressed near the top of the scale (81 of 132 Faithfulness scores equal 1.0) and only moderately repeatable (Spearman correlation between two scorings at temperature 0.1: 0.21 for Context Relevance, 0.71 for Faithfulness, and 0.52 for Answer Relevance), so small effects of k would be hard to detect. Second, the judge is the same model as the generator and has not been validated against human ratings. The finding therefore means no detectable effect in this sample, not proof that k does not matter.

Reliability of the Custom LLM-as-Judge Scheme

Findings in the subchapter on Data Quality Findings in the LLM-as-Judge scheme show that the first version of the custom judge (Khalila et al., 2025) was not reliable: it produced degenerate output in 3.8% of the rows in one run, it coded failed extractions as zero, and its scores did not agree with those of the revised judge. The revised judge removed these technical failures, but its reliability is still limited. Scoring the same 44 rows twice at temperature 0.1 gave Spearman correlations of 0.21 for Context Relevance, 0.71 for Faithfulness, and 0.52 for Answer Relevance, with identical scores in 50.0%, 84.1%, and 63.6% of the rows, and the judge scores were unrelated to retrieval F1-Score. This is consistent with the limited capacity of a small quantized model used as a judge. Because the judge is also the generator and no human validation was carried out, the LLM-judge results should be treated as exploratory. Future studies should use a stronger or different judge model and validate it against human ratings.

Comparison with Literature

(Rao et al., 2025) reported that smaller embedding models (MiniLM-v6) can outperform larger models (BGE-Large) when combined with LLM-based re-ranking, across different domains and languages from this study. However, in our results, there was no significant difference between the large and base variants of the same model (multilingual-e5). This aligns with (Rao et al., 2025) general finding that model size is not the main factor in retrieval performance, even though it is not identical because our study compared two variants of the same model family, not two different architectures. This finding also supports (Khalila et al., 2025) argument that the quality of RAG heavily depends on the alignment between query, document, and retrieval strategy, not just on the embedding model size used.

Limitation

This study has several limitations. First, the judge model used in evaluating LLM-as-judge is the same generative model as the generative component (Dolphin3.0-Llama3.1-8B), so the evaluation results could be biased since the model is assessing its own outputs (self-evaluation bias). Cross-validation against human assessment has not been carried out, and a second judge model was not tested. A test-retest of the revised judge on 44 rows at temperature 0.1 gave Spearman correlations of 0.21 (Context Relevance), 0.71 (Faithfulness), and 0.52 (Answer Relevance), so the LLM-judge results are only moderately repeatable and should be treated as exploratory. Second, the first version of the judge showed score extraction failures (5 of 132 rows, 3.8%, in one run) because of repetitive output and an extraction function that took the first numerical match without validation; this version was replaced, but the scores of the revised judge are compressed near the top of the scale (81 of 132 Faithfulness scores equal 1.0), which limits their ability to discriminate. Third, all experiments were conducted on a single corpus (Permenkes Number 41 of 2014) and one domain (nutrition/public health regulations in Indonesian), so generalization to other domains or languages would require further testing. Fourth, the range of k values tested, limited to 3, 5, and 7, has not been shown to indicate if a similar pattern would hold for much larger or smaller k values. Fifth, the Wilcoxon signed-rank test between embedding models (Table 6) was conducted by pooling data across k values ($n = 66$ pairs = 22 queries $\times$ 3 k), so the actual independent units are still 22 queries, not 66; this pooling could lead to pseudo-replication that affects the validity of the p -value. Considering that the average differences between models across all metrics are relatively small (<0.02) compared to the standard deviation (0.10–0.38), the statistical power of a test with 22 independent units might not be sufficient to detect small differences; thus, the finding of no significant difference should be interpreted as simply not detecting a difference with this design and sample size, not as definitive evidence of no performance difference between the two models. As a check that avoids pooling, Table 7 reports per-query differences with bootstrap confidence intervals, which exclude differences larger than about 0.07 for all six metrics. The k effect test (Friedman test) did not run into similar issues because its design is indeed repeated-measures within the same query, and the results ($p < 0.001$ for Precision and F1-Score) indicate that $n=22$ is already sufficient to detect that effect. Sixth, the answers, and therefore the LLM-judge scores, come from the final run of the pipeline. Generation uses sampling (temperature 0.1), and the scores of the first judge version differed substantially between two runs (for example, mean Faithfulness of e5-base at $k = 3$ was 0.781 in the first run and 0.693 in the second), whereas the retrieval metrics were identical in both runs. Seventh, computational cost (encoding time, query latency, and index size) was not measured directly, so the efficiency implications rest on vector dimensionality and model size alone. Eighth, text was extracted with PyPDFLoader without optical character recognition or image analysis, so information that appears only in figures or scanned images was not available to the pipeline. Ninth, relevance was judged at the page level, so a retrieved chunk from a correct page was counted as relevant even if it did not contain the answer text, and the non-inferiority margins (0.05 and 0.10) were chosen after seeing the data.

Practical Implications

Based on the results obtained, no significant differences were found between multilingual-e5-large and multilingual-e5-base across all tested metrics. The main practical implication of this study is that multilingual-e5-base, with its smaller vector dimension (768 vs 1024) and consequently lower storage requirements per vector, was non-inferior to multilingual-e5-large within a margin of 0.05 on all metrics except Recall and Faithfulness (Table 9), for similar use cases (RAG on single-document Indonesian health regulation documents). Computational cost was not measured directly, so no efficiency gain is claimed beyond the smaller vector size. Regarding the value of k , since k had no detectable effect on the quality of the answers but does noticeably impact Precision/Recall, choosing k should be adjusted according to the application's priority. A smaller k , like $k=3$, can help minimize noise and speed up inference if precision is prioritized, or a larger k , like $k=7$, can be used if coverage (recall) of information is more important, with no detectable drop in answer quality in this sample.

CONCLUSION

This study compares the performance of multilingual-e5-large and multilingual-e5-base on the RAG architecture for question-answering documents of Permenkes Number 41 of 2014 on Balanced Nutrition Guidelines, across different k values, such as $k = (3, 5, 7)$ and 22 test queries (132 evaluation data points). The Wilcoxon signed-rank test results showed no statistically significant difference between the two embedding models across all six metrics tested (Precision, Recall, F1-Score, Context Relevance, Faithfulness, Answer Relevance, all $p > 0.05$). The Friedman test results showed that the k value has a significant effect on retrieval metrics based on ground truth (Precision, Recall, F1-Score; $p < 0.05$) but does not significantly affect answer quality metrics based on LLM-judge. These findings suggest that for using RAG on single Indonesian-language health regulation documents, choosing between a large or base embedding model does not result in a detectable performance difference in this sample, so the smaller base version is a reasonable alternative, although the storage saving is small for a corpus of this size and computational efficiency was not measured. The study also has limitations that affect how these findings should be read. The judge in the custom LLM-as-judge setup is the same model as the generator (Dolphin3.0-Llama3.1-8B), so self-evaluation bias cannot be ruled out. A test-retest of the revised judge on 44 rows at temperature 0.1 showed only moderate repeatability (Spearman correlations of 0.21 for Context Relevance, 0.71 for Faithfulness, and 0.52 for Answer Relevance), and the scores were compressed near the top of the scale, so the LLM-judge results are treated as exploratory. Validation through human evaluation and a different judge model was not carried out in this study. In addition, because the corpus is a single regulation with 22 questions, the findings should not be generalized to the wider Indonesian legal corpus.

Building on these limitations, several directions for future research are recommended. First, to address self-evaluation bias, the LLM-as-judge scores should be validated against human evaluators, and a different judge model, preferably stronger than the 8B quantized model used here, should be used instead of the generator. Second, the models should be tested on a larger and more diverse Indonesian legal corpus with more questions, including paraphrased and multi-page questions, since the present test set was nearly saturated for single-page questions. Third, additional embedding models, such as the small variant of multilingual-e5 and other multilingual or Indonesian-specific models, should be evaluated to show whether the absence of a size effect holds beyond this model family. Fourth, other retrieval strategies, such as hybrid sparse-dense retrieval, re-ranking, structure-aware chunking, or query decomposition for questions that span many pages, and a wider range of k values should be compared. Finally, measuring encoding time, query latency, and index size would show whether the smaller vectors of the base model translate into measurable efficiency gains. Addressing these points would show how the limitations identified in this study can be resolved and how the proposed RAG system can be further improved.

REFERENCES

- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). *RAGAS: Automated Evaluation of Retrieval Augmented Generation*.
- Faisal, D. R., Darari, F., & Ryanda, R. A. (2024). Granularity-aware legal question answering: A case study of Indonesian government regulations. *International Journal of Advances in Intelligent Informatics*, 10(3), 359. <https://doi.org/10.26555/ijain.v10i3.1105>
- Febrian, G. F., & Figueredo, G. (2024). *KemenkeuGPT: Leveraging a Large Language Model on Indonesia's Government Financial Data and Regulations to Enhance Decision Making* (arXiv:2407.21459). arXiv. <https://doi.org/10.48550/arXiv.2407.21459>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Khalila, Z., Nasution, A. H., Monika, W., Onan, A., Murakami, Y., Radi, Y. B. I., & Osmani, N. M. (2025). Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *International Journal of Advanced Computer Science and Applications*, 16(2). <https://doi.org/10.14569/IJACSA.2025.01602134>
- Ongri, J. G., Tobing, B. C. L., & Faisal, D. R. (2025). *Benchmarking KG-based RAG Systems: A Case Study of Legal Documents*.
- Palupi, A. C., & Irawan, A. (2025). *Evaluating IndoGPT for Legal Queries: A Benchmark Against GPT-4 Models: ISSN: 2599-1663. 09(02), 22–27*. <https://doi.org/10.25077/9.1.22-27.2025>



- Pratama, A. Z., Marinta, A., Triyudanto, B., Saman, M., & Noor, T. (2025). Retrieval-Augmented Generation for Indonesian Criminal Law Information Using the LLaMA Model. *Innovative Informatics and Artificial Intelligence Research (IIAIR)*, 1(1).
- Rahmadi, N., & Setyanto, A. (2026). LexIndoLLM: Large Language Model untuk Konsultasi Regulasi Daerah di Indonesia. *Jurnal Buana Informatika*, 17(1), 1–10.
- Rakhmawati, N. A., Satriawan, I., Gunawan, Y., & Noviyana, R. F. (2023). DEVELOPMENT OF INDONESIAN CONSTITUTION QUESTION ANSWERING SYSTEM. *UPB Scientific Bulletin, Series C: Electrical Engineering and Computer Science*, 85(4), 59–68.
- Rao, A., Alipour, H., & Pendar, N. (2025). *Rethinking Hybrid Retrieval: When Small Embeddings and LLM Re-ranking Beat Bigger Models* (arXiv:2506.00049). arXiv. <https://doi.org/10.48550/arXiv.2506.00049>
- Rina, W., Km, S., Kl, M., Robert, D., Kes, M., Suryaningsih, C., Kep, S., Kep, M., Kep, S. T., Kes, M., Sya'diah, Y., Wenno, S. Z., Kes, M., Tahulending, A. A., Pd, S., & Kes, M. (2025). ANALISIS KEBIJAKAN KESEHATAN. *PT Media Pustaka Indo*.
- Rozaq, A., Fitri, R., Akbar, M. S., Sumari, M. H., Khafiyya, N., Wardiah, I., & Sholihin, F. (2025). Legal Literacy in Indonesia: Leveraging Semantic-Based AI and NLP for Enhanced Civil Law Access. *E3S Web of Conferences*, 622, 03002. <https://doi.org/10.1051/e3sconf/202562203002>
- Sembiring, A., & Santoso, H. (2026). Comparative Academic Performance Prediction in Primary Schools Using Linear Regression and Random Forest. *Sinkron*, 10(2), 1104–1113. <https://doi.org/10.33395/sinkron.v10i2.15953>
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 Text Embeddings: A Technical Report* (arXiv:2402.05672). arXiv. <https://doi.org/10.48550/arXiv.2402.05672>
- Yafie, H. A., Zailani, A. U., Muttaqin, W., Daffansyah, M. S. A., & Ramzi, M. R. (2026). Academic Chatbot for Campus Information Services Using Retrieval-Augmented Generation. *Sinkron*, 10(3), 1330–1339. <https://doi.org/10.33395/sinkron.v10i3.16088>
- Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2025). *Evaluation of Retrieval-Augmented Generation: A Survey* (pp. 102–120). https://doi.org/10.1007/978-981-96-1024-2_8
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2026). Retrieval-Augmented Generation for AI-Generated Content: A Survey. *Data Science and Engineering*, 11(1), 1–29. <https://doi.org/10.1007/s41019-025-00335-5>

