JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer
ISSN: 2615-577X (Online)

Evaluating Multilingual Sentence Embeddings for Indonesian Open-Domain Passage Retrieval

Shinvi Nur Najmil Jannah[1], Alam Rahmatulloh[2]

Program Studi Informatika, Fakultas Teknik, Universitas Siliwangi
Jl. Mugarsari, Tasikmalaya, Jawa Barat, Indonesia
[1]237006008@student.unsil.ac.id
[2]alam@unsil.ac.id

Abstract— **Indonesian search systems rarely have labeled data for fine-tuning dense retrievers, yet it is unclear which ready-to-use multilingual encoders can beat a properly configured lexical baseline. This study compares LaBSE, paraphrase-multilingual-mpnet-base-v2, and multilingual-e5-large, used without fine-tuning by the authors, against BM25 on the Indonesian subset of MIRACL: 960 dev queries over a 50,000-passage Wikipedia corpus that contains all 8,286 judged passages. Passages were indexed with FAISS and systems were scored with nDCG@10, MRR, and Recall@100. Multilingual-e5-large was the only encoder to surpass BM25 with regex tokenization, reaching nDCG@10 of 0.713 versus 0.579 (+23.1%), whereas paraphrase-multilingual-mpnet (0.502) and LaBSE (0.227) fell below it. A naive whitespace-tokenized BM25 scored only 0.354, which would have made two of the three encoders look superior. The results suggest that retrieval-oriented training matters more than language coverage, although multilingual-e5-large was also fine-tuned on MIRACL training data and the evaluation covers only Wikipedia passages. A search prototype built on the best model shows practical use.**

Keywords*— *BM25, dense retrieval, FAISS, Indonesian passage retrieval, information retrieval, multilingual sentence embedding

Abstrak— **Sistem pencarian berbahasa Indonesia jarang memiliki data berlabel untuk fine-tuning model dense retrieval, padahal belum jelas encoder multibahasa siap pakai mana yang mampu mengungguli baseline leksikal yang dikonfigurasi dengan benar. Penelitian ini membandingkan LaBSE, paraphrase-multilingual-mpnet-base-v2, dan multilingual-e5-large, yang dipakai tanpa fine-tuning oleh penulis, terhadap BM25 pada subset bahasa Indonesia MIRACL: 960 query dev pada korpus Wikipedia 50.000 passage yang memuat seluruh 8.286 passage berlabel. Passage diindeks dengan FAISS dan sistem dinilai dengan nDCG@10, MRR, dan Recall@100. Multilingual-e5-large menjadi satu-satunya encoder yang melampaui BM25 dengan tokenisasi regex, dengan nDCG@10 0,713 berbanding 0,579 (+23,1%), sedangkan paraphrase-multilingual-mpnet (0,502) dan LaBSE (0,227) berada di bawahnya. BM25 dengan tokenisasi spasi sederhana hanya mencapai 0,354 sehingga dua dari tiga encoder tampak unggul. Hasil ini menunjukkan bahwa pelatihan yang berorientasi retrieval lebih menentukan daripada cakupan bahasa, meskipun multilingual-e5-large juga telah di-fine-tune pada data latih MIRACL dan evaluasi hanya mencakup passage Wikipedia. Prototipe pencarian berbasis model terbaik menunjukkan penerapan praktisnya.**

Kata kunci*— *BM25, dense retrieval, FAISS, passage retrieval bahasa Indonesia, sentence embedding multibahasa, temu balik informasi

I. INTRODUCTION

Search over Indonesian text, from digital libraries to institutional document portals, depends on an information retrieval (IR) component that surfaces relevant passages for a user query. Retrieval methods range from lexical approaches such as BM25 [1] to dense retrieval, which encodes queries and passages into a shared vector space and ranks passages by similarity [2], [3], [4]. BM25 nevertheless remains a strong reference point: the BEIR benchmark [5], which evaluates retrieval models across 18 heterogeneous datasets, found that dense models often fail to generalize beyond their training domain.

Dense retrieval became practical with siamese sentence encoders [2] and with dual encoders trained on question–passage pairs, which can outperform BM25 on open-domain question answering [4]. A recent survey [3] concludes that dense retrieval based on pretrained language models has become the dominant paradigm, but most of the models it reviews are developed and evaluated on English benchmarks. For languages with little labeled retrieval data, such as Indonesian, using or fine-tuning such models is therefore uncertain [6].

Multilingual sentence encoders offer a way around this data gap because a single model covers many languages, including Indonesian, and can be applied without task-specific fine-tuning. This study examines three widely used encoders built with different training objectives: LaBSE [7], trained with a translation ranking objective for bitext alignment; paraphrase-multilingual-mpnet-base-v2, obtained by distilling a monolingual paraphrase model into a multilingual student [8]; and multilingual-e5-large [9], pretrained contrastively on text pairs and then fine-tuned on retrieval data. All three list Indonesian among their supported languages and are distributed through the sentence-transformers ecosystem, which makes them realistic first choices for practitioners. Their different objectives also make them informative contrasts: if

language coverage were sufficient, the three should perform similarly, whereas if the training objective matters, retrieval-oriented training should show an advantage.

Benchmarks now make such comparisons possible: MTEB [10] standardizes the evaluation of text embeddings across tasks, and MIRACL [11] provides human-annotated relevance judgments for 18 languages, including Indonesian, extending earlier resources such as Mr. TyDi [12]. Prior work with these resources has mostly studied models trained or fine-tuned for retrieval [11], [13], has shown that language support does not guarantee uniform quality across languages [14], and has reported that lexical baselines remain competitive [5]. Indonesian-specific work compares BM25 with IndoSBERT variants [15], and IndoNLU [16] concentrates on classification-style tasks rather than passage retrieval. Earlier work published in JATIKOM also shows how strongly Indonesian lexical matching depends on preprocessing: Rabin-Karp hashing for thesis plagiarism detection cannot match text in which words are replaced by synonyms [17], and an Enhanced Confix Stripping stemmer used for readability measurement still makes over- and understemming errors [18]. The multilingual E5 report [9] compares its models with BM25 and mDPR baselines on MIRACL but does not vary the tokenization of the BM25 baseline. To our knowledge, no study has examined, for Indonesian passage retrieval, how the comparison between off-the-shelf multilingual encoders and BM25 depends on that tokenization, so it remains unclear which encoder, if any, justifies replacing a lexical system when no labeled Indonesian data are available.

Table I summarizes the most closely related studies and states how the present study relates to each of them.

TABLE I
SUMMARY OF CLOSELY RELATED WORK AND POSITION OF THIS STUDY

Study	Focus and method	Relation to this study
Thakur et al. [5]	Zero-shot benchmark; BM25, dense retrievers, and re-rankers on 18 datasets	Shows BM25 is a strong reference; motivates a carefully tokenized BM25 baseline here
Muennighoff et al. [10]	Text embedding benchmark; 33 models, 8 task types	Evaluates embeddings in general, not Indonesian passage retrieval
Zhang et al. [11]	MIRACL: human-annotated retrieval benchmark in 18 languages; BM25 and mDPR baselines	Supplies the benchmark used here; off-the-shelf multilingual encoders are not compared
Zhang et al. [13]	Recipes for training multilingual dense retrievers in several data scenarios	Addresses fine-tuning; this study uses encoders without fine-tuning
Yang et al. [14]	Language bias of multilingual retrievers	Motivates testing encoders that claim Indonesian support
Lasminingrum et al. [15]	Lexical vs. semantic retrieval for Indonesian documents; BM25 and IndoSBERT variants	Indonesian-specific encoders; this study evaluates multilingual ones
This study, 2026	**LaBSE, mpnet, and e5-large vs. BM25 (regex and whitespace) on MIRACL Indonesian**	**Only e5-large beats BM25 (regex); baseline tokenization changes the conclusion**

Building on this gap, the research problem is that, without labeled Indonesian data, practitioners cannot tell a priori which off-the-shelf multilingual encoder, if any, outperforms a properly configured lexical baseline for passage retrieval. This leads to the research question: how much do LaBSE, paraphrase-multilingual-mpnet-base-v2, and multilingual-e5-large, used without fine-tuning by the authors, improve Indonesian passage retrieval over BM25? Accordingly, the objective is to compare their retrieval effectiveness with BM25 on the Indonesian subset of MIRACL [11] using nDCG@10, MRR, and Recall@100. The main contributions are fourfold: (i) a controlled comparison of three encoders with different training objectives against two BM25 configurations; (ii) evidence that baseline tokenization alone changes which encoders appear superior; (iii) an analysis showing why accuracy-based views of top-10 retrieval are misleading; and (iv) a deployed search prototype, SPIRIT, built on the best encoder. The results show that only multilingual-e5-large outperforms BM25 with regex tokenization, by about 23% in nDCG@10. The remainder of this paper is organized as follows: Section II describes the dataset, models, and evaluation procedure; Section III presents the results; Section IV discusses the findings and their limits; and Section V concludes the paper.

II. MATERIAL AND METHODS

Figure 1 summarizes the overall pipeline, from corpus preparation through retrieval with off-the-shelf encoders, evaluation, and deployment of the best-performing model into a web application.

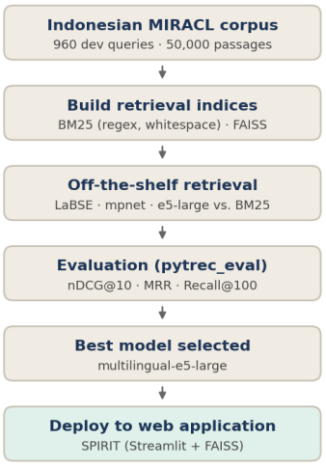

Figure 1. Overall research pipeline, from corpus preparation to web application deployment

A. Dataset

This study uses the Indonesian subset of MIRACL [11], a retrieval benchmark whose relevance judgments were produced by native-speaker annotators. The 960 queries of the dev partition serve as evaluation queries. Their qrels contain 9,668 judged query–passage pairs, of which 3,088 are positive (relevance 1) and 6,580 are judged negative (relevance 0), covering 8,286 unique passages. The evaluation corpus was built in two steps: all 8,286 judged passages were taken from the full Indonesian MIRACL corpus of Wikipedia passages (1,446,315 passages), and 41,714 distractor passages that were not judged for any dev query were then sampled uniformly at random without replacement (random seed 42), which brings the corpus to 50,000 passages. Keeping every judged passage preserves all relevant passages and all judged negatives, while the random distractors keep indexing cost manageable. Because a 50,000-passage corpus is easier to search than the full corpus, absolute scores are not comparable with results on the full corpus; the comparison between systems is the quantity of interest.

B. Models Evaluated

Five retrieval systems are compared. BM25 serves as the lexical baseline and is implemented with the rank_bm25 library (BM25Okapi with its default parameters, k1 = 1.5 and b = 0.75). Two tokenizations are evaluated: lowercase text split on word characters (regex \w+), which is the main baseline, and lowercase whitespace splitting, which is retained to show the effect of tokenization. Three dense encoders are used as released, without fine-tuning by the authors (Table II): LaBSE [7], paraphrase-multilingual-mpnet-base-v2 [8], and multilingual-e5-large [9]. They were chosen because they cover three training objectives, all support Indonesian, and are widely used through the sentence-transformers library [2]. Note that multilingual-e5-large was supervised-fine-tuned on retrieval data that include the MIRACL training split [9]. The dev queries evaluated here are held out, but the model is therefore not zero-shot with respect to MIRACL. Each encoder was run with its default maximum input length, which differs across models.

TABLE II
ENCODERS EVALUATED

Model	Training objective	Max. tokens	Dim.
LaBSE [7]	Translation ranking for bitext alignment	256	768
paraphrase-multilingual-mpnet-base-v2 [8]	Multilingual distillation of a paraphrase model	128	768
multilingual-e5-large [9]	Contrastive pre-training, then supervised retrieval fine-tuning (includes MIRACL training split)	512	1024

C. Indexing and Retrieval

Every passage was encoded with each dense model using a batch size of 64 and L2-normalized, so that inner product equals cosine similarity. The vectors were indexed with a FAISS IndexFlatIP [19] for exhaustive search, and the 100 highest-scoring passages were retrieved for each query encoded with the same model. For multilingual-e5-large, the main experiments encode queries and passages without the “query: ” and “passage: ” prefixes recommended in its model card; a sensitivity run with the prefixes is reported in Section III. BM25 returns its 100 highest-scoring passages for the tokenized query. Experiments ran on Google Colab with a Tesla T4 GPU (Python 3.13, sentence-transformers 5.7.0, PyTorch 2.11.0, FAISS-CPU 1.15.1, rank_bm25 0.2.2, and pytrec_eval 0.5). The model checkpoints were sentence-transformers/LaBSE (revision 836121a053), sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (4328cf2639), and intfloat/multilingual-e5-large (3d7cfbdacd), as available in October 2026. Effectiveness was evaluated once per system, because the corpus, queries, and encodings are fixed and FAISS flat search is exact; latency was measured in three repeated runs.

D. Evaluation Metrics

Evaluation uses pytrec_eval [20]. nDCG@10 and Recall@100 are the official metrics of MIRACL [11]: nDCG@10 rewards placing relevant passages near the top, and Recall@100 measures how many relevant passages enter the top 100, which matters when a retriever feeds a re-ranker. Mean Reciprocal Rank (MRR), computed over the 100 retrieved passages, complements them by reporting how early the first relevant passage appears. The MIRACL qrels were used without modification: passages labeled positive receive relevance 1 and judged-negative passages receive 0, and a retrieved passage without a judgment is counted as non-relevant by pytrec_eval. Scores are averaged over all 960 queries. As a supplementary view, retrieval was also framed as binary classification over the 9,668 judged pairs: a pair is predicted relevant if its passage appears in the system’s top 10 and not relevant otherwise, which yields true positive (TP), false positive (FP), false negative (FN), and true negative (TN) counts; retrieved passages without judgments are ignored.

E. Web Application Implementation

To demonstrate practical use, the best-performing encoder (multilingual-e5-large) was deployed in a Streamlit search application named SPIRIT (Semantic Passage Information Retrieval for Indonesian Text), which loads the same embeddings and FAISS index used in the evaluation. Users enter a free-text query in Indonesian and receive the top-k passages ranked by cosine similarity, and a statistics page summarizes the offline evaluation results. Figure 2 shows the search interface and an example for the query “sejarah indonesia” (Indonesian history). The application was checked with a

manual black-box test of four input scenarios. A scenario passes when (C1) the application remains responsive without an unhandled error, (C2) the user receives either a ranked result list or an explicit message, and (C3) the top results are topically related to the query on manual inspection, which applies to meaningful queries only. Retrieval latency was measured offline, as described in Section III.

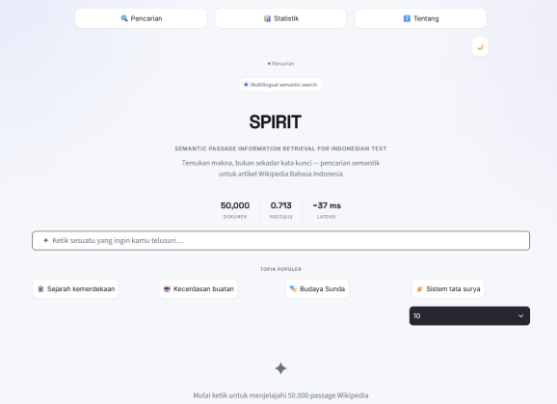

Figure 2. SPIRIT web application : search interface

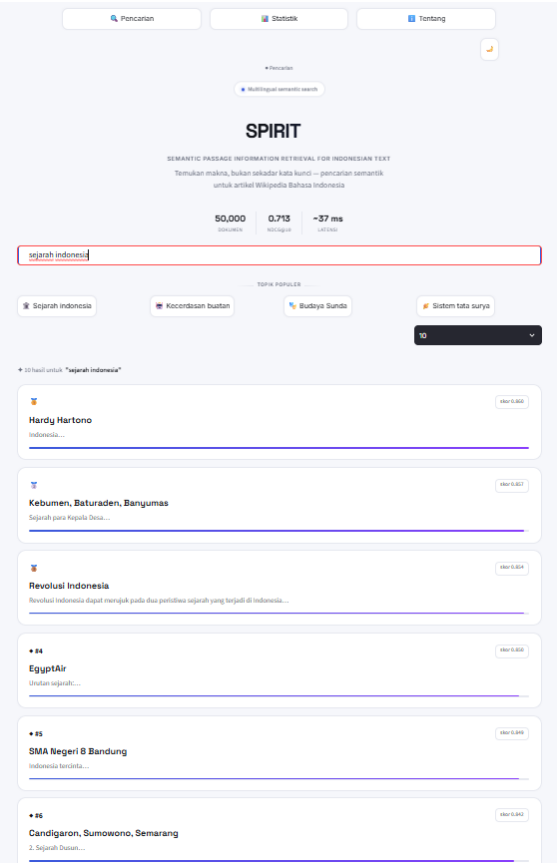

Figure 3. SPIRIT web application : top results returned for the query "sejarah indonesia"

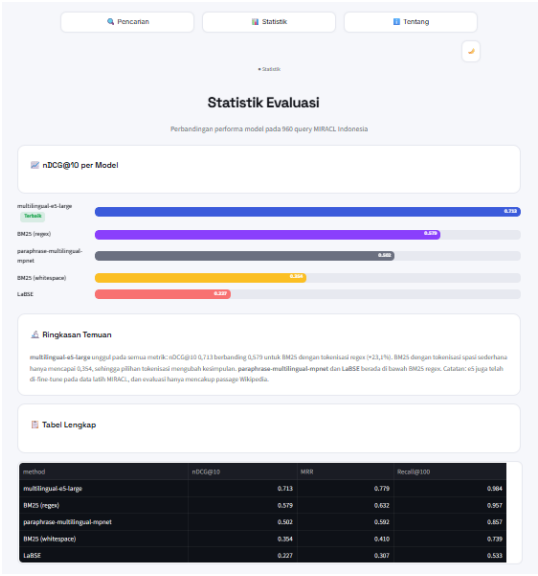

Figure 4. SPIRIT web application : Statistic Evaluating

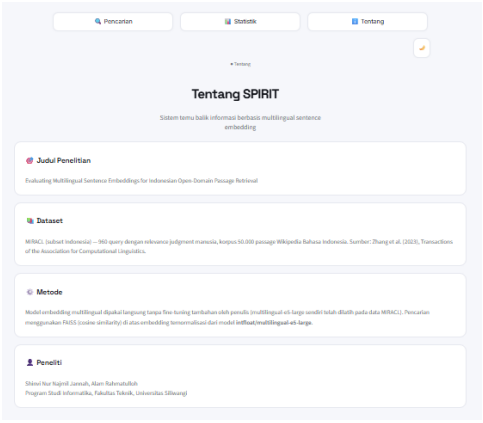

Figure 5. SPIRIT web application : About the SPIRIT

III. RESULTS

Table III reports the results of the five retrieval systems on the 960 dev queries, sorted by nDCG@10. The percentage in parentheses is the relative change with respect to BM25 with regex tokenization, the main baseline, and Figure 6 plots the same scores.

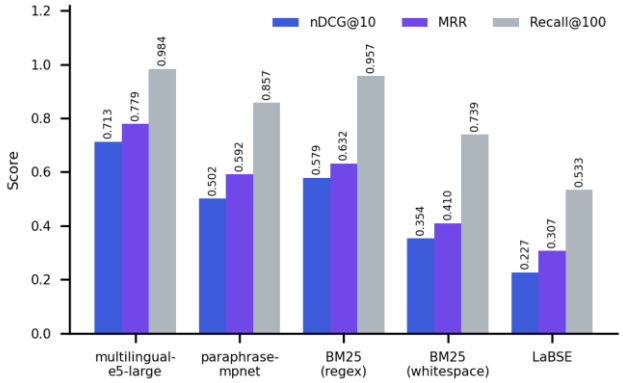

Figure 6. Comparison of nDCG@10, MRR, and Recall@100 across five retrieval systems on 960 MIRACL Indonesian dev queries

TABLE III
RETRIEVAL RESULTS ON 960 MIRACL INDONESIAN DEV QUERIES
(RELATIVE CHANGE VS. BM25 WITH REGEX TOKENIZATION)

Method	nDCG@10	MRR	Recall@100
multilingual-e5-large	0.713 (+23.1%)	0.779 (+23.2%)	0.984 (+2.8%)
BM25 (regex), baseline	**0.579**	**0.632**	**0.957**
paraphrase-multilingual-mpnet	0.502 (−13.4%)	0.592 (−6.4%)	0.857 (−10.5%)
BM25 (whitespace)	0.354 (−38.9%)	0.410 (−35.1%)	0.739 (−22.8%)
LaBSE	0.227 (−60.9%)	0.307 (−51.4%)	0.533 (−44.4%)
Sensitivity: e5-large with prefixes	*0.726 (+25.4%)*	*0.789 (+24.8%)*	*0.986 (+3.0%)*

Note: Relative changes are computed from the unrounded scores, so they can differ in the last digit from changes computed from the rounded values shown (for paraphrase-multilingual-mpnet, nDCG@10: 0.502/0.579 − 1 gives −13.3% from the rounded values, whereas the unrounded scores give −13.4%).

Against the regex-tokenized BM25 baseline (nDCG@10 0.579, MRR 0.632, Recall@100 0.957), only multilingual-e5-large is better on every metric: +0.134 in nDCG@10 (+23.1%), +0.147 in MRR (+23.2%), and +0.027 in Recall@100 (+2.8%). The gain is therefore concentrated in the ordering of the top results rather than in candidate recall, where BM25 already finds 95.7% of the relevant passages within the top 100. Paraphrase-multilingual-mpnet falls below BM25 on all three metrics (−0.077, −0.040, and −0.100, or −13.4%, −6.4%, and −10.5%), and LaBSE is far below it (nDCG@10 0.227, −60.9%; Recall@100 0.533, −44.4%), even though both models list Indonesian as a supported language.

The baseline configuration changes the picture. With whitespace tokenization BM25 reaches an nDCG@10 of only 0.354, so multilingual-e5-large would appear to double it (+101.5%) and paraphrase-multilingual-mpnet to beat it (+41.8%). Regex tokenization alone raises BM25 by about 64% in nDCG@10 and removes the advantage of paraphrase-multilingual-mpnet. Both BM25 variants use the same corpus and the same library, so the difference is due to tokenization only.

Adding the "query: " and "passage: " prefixes recommended for E5 raises nDCG@10 from 0.713 to 0.726 (+0.013), MRR from 0.779 to 0.789, and Recall@100 from 0.984 to 0.986, so the comparison is not sensitive to this choice. The main results keep the configuration without prefixes, which is also the one used by the SPIRIT index.

As a supplementary analysis, retrieval was framed as binary classification over the 9,668 judged query–passage pairs (3,088 relevant and 6,580 not relevant): a pair is predicted relevant if its passage appears in the system's top 10. Figure 7 shows the resulting confusion matrices.

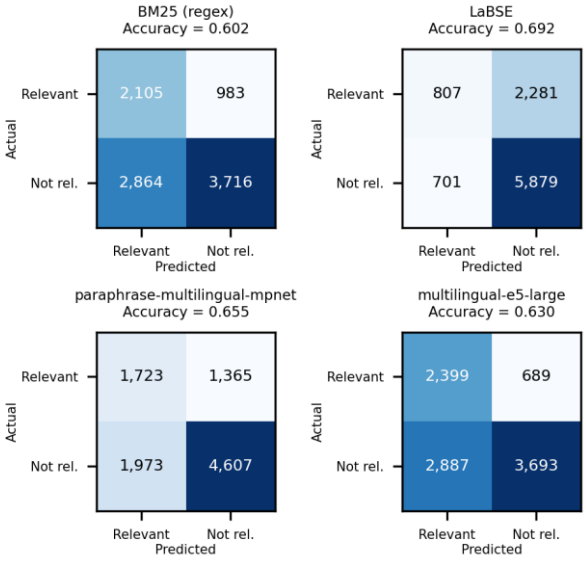

Figure 7. Confusion matrices at the top-10 cutoff over 9,668 judged query–passage pairs (MIRACL Indonesian dev queries)

The accuracy ranking in Figure 7 (LaBSE 0.692, paraphrase-multilingual-mpnet 0.655, multilingual-e5-large 0.630, BM25 with regex tokenization 0.602; BM25 with whitespace tokenization, not shown, 0.616) almost reverses the nDCG@10 ranking in Table III. The reason is class imbalance: 68.1% of the judged pairs are not relevant, so a system that returned nothing would reach an accuracy of 0.681, higher than every system except LaBSE, which exceeds it by only 0.011. LaBSE attains the highest accuracy and precision (0.535) by retrieving few relevant passages, and its recall is only 0.261, whereas multilingual-e5-large has the highest recall (0.777) and BM25 with regex tokenization reaches 0.682. Accuracy therefore rewards what a system leaves out, which is why rank-aware metrics are used as the main criteria.

Table IV reports per-query latency on the Tesla T4 GPU: encoding the query and searching the index for the top 10 passages, averaged over the 960 dev queries after ten warm-up queries and then averaged over three repeated runs (the standard deviation of the mean across runs was at most 0.9 ms). Multilingual-e5-large needs 36.6 ms on average (95th percentile 42.6 ms), compared with about 23 ms for LaBSE and paraphrase-multilingual-mpnet, which have the same depth and hidden size and therefore nearly identical latency. BM25 took 78.2 ms per query for the whitespace-tokenized variant, but this reflects the pure-Python rank_bm25 implementation rather than an optimized engine, so it should not be read as a general efficiency comparison. Encoding the 50,000 passages with multilingual-e5-large took 18 min 33 s at a batch size of 64; this time was measured in the run with the prefixes, whereas the main results and the SPIRIT index use the encodings without prefixes.

TABLE IV
PER-QUERY RETRIEVAL LATENCY ON A TESLA T4 GPU (960 DEV QUERIES)

Method	Mean (ms)	Median (ms)	p95 (ms)
BM25 (whitespace)	78.2	75.6	119.6
LaBSE	23.3	22.7	27.5
paraphrase-multilingual-mpnet	23.2	22.6	27.0
multilingual-e5-large	36.6	35.7	42.6

In addition to the offline evaluation, the SPIRIT web application was checked with four manual input scenarios judged against criteria C1–C3 defined in Section II. Table V summarizes the scenarios and the results.

TABLE V
FUNCTIONAL CHECKS ON THE SPIRIT WEB APPLICATION

Scenario	Input	Observed behavior	Result
Empty query	No text entered	Shows a prompt to start typing; no results and no error	Pass (C1, C2)
Long query	Multi-paragraph text (about 2–3 paragraphs)	Ranked results returned normally	Pass (C1, C2)
Special characters	Symbols (e.g., "@#$%^&*")	Responds normally and displays results	Pass (C1, C2)
Cross-lingual query	English query (e.g., "history of Indonesia")	Topically relevant results returned	Pass (C1–C3)

All four scenarios pass. For an empty query the application shows a prompt inviting the user to start typing (Figure 2) instead of an empty page or an error. These checks cover basic robustness only and are not a usability study.

IV. DISCUSSION

The measurements support three findings. First, only multilingual-e5-large outperformed a properly tokenized BM25, by about 23% in nDCG@10 and MRR but by only 2.8% in Recall@100, so its advantage lies in ordering the top results. Second, paraphrase-multilingual-mpnet and LaBSE did not outperform BM25 although both support Indonesian, which shows that language coverage alone does not predict retrieval quality, in line with reports of uneven multilingual retrieval quality [14]. Third, the conclusion depends on the baseline: with whitespace tokenization BM25 looks weak enough for two encoders to appear superior, which is consistent with the role of BM25 as a strong reference point in BEIR [5].

The following explanations are hypotheses, because the experiments do not isolate them. The training objective may explain the ordering: LaBSE is optimized to align translations [7], which rewards cross-lingual equivalence rather than topical relevance between a short query and a passage; paraphrase-multilingual-mpnet learns from a paraphrase-oriented teacher [8], so its similarity reflects sentence meaning more than answer relevance; and multilingual-e5-large was trained contrastively on large collections of text pairs and then fine-tuned for retrieval [9]. Two confounds, however, prevent attributing the advantage of multilingual-e5-large to the objective alone. Its fine-tuning data include the MIRACL training split [9], so it benefited from in-domain supervision, and its input limit of 512 tokens is far above the 128 tokens of paraphrase-multilingual-mpnet, which may truncate longer passages and lower its scores. Model size was not controlled either. Isolating these factors would require truncation-matched runs and encoders trained on comparable data.

These results agree with BEIR [5] in showing that a lexical baseline can be hard to beat for encoders that are not trained for the task, and with studies of language bias [14] in showing that nominal language support is a poor predictor of quality. They extend MIRACL [11], whose reference dense baseline is fine-tuned on retrieval data, by comparing off-the-shelf encoders with a carefully tokenized BM25 on a single language. Indonesian work comparing BM25 with IndoSBERT variants [15] addresses the same lexical-versus-semantic question with Indonesian-specific encoders, whereas this study evaluates multilingual encoders that can be reused without Indonesian training data.

For practitioners building Indonesian passage search without labeled data, multilingual-e5-large is the only encoder in this study that justifies replacing BM25 for ranking the top results, while a tokenized BM25 remains a competitive and inexpensive source of candidates (Recall@100 of 0.957 versus 0.984). A hybrid design is a natural next step but was not tested. The SPIRIT prototype shows that the model can serve queries at about 37 ms of retrieval latency on a T4 GPU; caching the results of repeated queries, which cut the retrieval time of a single record from about 2.1 s to 0.02 s with Redis in a small single-user JATIKOM experiment [21], is a further serving option that was not tested here. Several limits restrict generalization. The corpus is a 50,000-passage subsample that retains all judged passages, so absolute scores are likely higher than on the full corpus and may not transfer; all passages come from Wikipedia, and the models may behave differently on news, legal, or conversational text; each system was evaluated once on one corpus sample, without confidence intervals or significance tests; the BM25 baseline uses no stemming or stopword removal, so a stronger Indonesian lexical pipeline could narrow the gap further, although by how much is uncertain, because Indonesian stemmers such as Enhanced Confix Stripping still make over- and understemming errors [18] and earlier JATIKOM work on Rabin-Karp text matching found only a small difference between runs with and without stopword removal and stemming [17]; and the encoders were used with default input lengths and without prefixes except in the sensitivity run.

V. CONCLUSION

This study set out to compare off-the-shelf multilingual sentence encoders with BM25 for Indonesian passage retrieval, and the evidence answers that objective directly: among LaBSE, paraphrase-multilingual-mpnet-base-v2,

and multilingual-e5-large, only multilingual-e5-large improves on a properly tokenized BM25 (nDCG@10 of 0.713 versus 0.579, +23.1%), whereas the other two encoders fall below it despite supporting Indonesian. Retrieval-oriented training appears to matter more than language coverage, although the exposure of multilingual-e5-large to MIRACL training data and its longer input limit prevent a clean attribution.

The main contributions are a controlled comparison of encoders with different training objectives, the finding that baseline tokenization alone can change which encoders appear superior, evidence that accuracy-based views of top-10 retrieval mislead, and the SPIRIT prototype. These conclusions hold for Indonesian Wikipedia passages in a 50,000-passage subsample of MIRACL and should not be generalized to other domains or to the full corpus without further evidence.

Future work should (i) evaluate on the full MIRACL corpus and on more diverse Indonesian collections such as news, legal, and conversational text, (ii) test hybrid BM25 and dense retrieval, for example with reciprocal rank fusion, (iii) measure efficiency at larger scale with approximate indexes and query-result caching, (iv) run truncation-matched and fine-tuning ablations to separate the training objective from input length and training data, and (v) add significance tests and stronger Indonesian lexical baselines with stemming and stopword removal.

ACKNOWLEDGMENT

The authors would like to thank the Informatics Study Program, Faculty of Engineering, Universitas Siliwangi, for the support provided throughout the research and writing of this article.

REFERENCES

[1] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Language Process. 9th Int. Joint Conf. Natural Language Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[3] W. X. Zhao, J. Liu, R. Ren, and J. R. Wen, "Dense text retrieval based on pretrained language models: A survey," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1–60, 2024, doi: 10.1145/3637870.

[4] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. 2020 Conf. Empirical Methods Natural Language Process. (EMNLP)*, Online, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[5] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2104.08663.

[6] J. Wu, Z. Ren, and S. Verberne, "What are the limits of cross-lingual dense passage retrieval for low-resource languages?," 2024, doi: 10.48550/arXiv.2408.11942.

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguist. (Volume 1: Long Papers)*, Dublin, Ireland, 2022, pp. 878–891. doi: 10.18653/v1/2022.acl-long.62.

[8] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. 2020 Conf. Empirical Methods Natural Language Process. (EMNLP)*, Online, 2020, pp. 4512–4525. doi: 10.18653/v1/2020.emnlp-main.365.

[9] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," 2024. doi: 10.48550/arXiv.2402.05672.

[10] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL)*, Dubrovnik, Croatia, 2023, pp. 2014–2037. doi: 10.18653/v1/2023.eacl-main.148.

[11] X. Zhang *et al.*, "MIRACL: A multilingual retrieval dataset covering 18 diverse languages," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1114–1131, 2023, doi: 10.1162/tacl_a_00595.

[12] X. Zhang, X. Ma, P. Shi, and J. Lin, "Mr. TyDi: A multi-lingual benchmark for dense retrieval," in *Proc. 1st Workshop Multiling. Represent. Learn. (MRL)*, Punta Cana, Dominican Republic, 2021, pp. 127–137. doi: 10.18653/v1/2021.mrl-1.12.

[13] X. Zhang, K. Ogueji, X. Ma, and J. Lin, "Toward best practices for training multilingual dense retrieval models," *ACM Transactions on Information Systems*, vol. 42, no. 2, 2023, doi: 10.1145/3613447.

[14] J. Yang, F. Jiang, and T. Baldwin, "Language bias in multilingual information retrieval: The nature of the beast and mitigation methods," in *Proc. 4th Workshop Multiling. Represent. Learn. (MRL)*, Miami, FL, USA, 2024, pp. 280–292. doi: 10.18653/v1/2024.mrl-1.23.

[15] D. P. Lasminingrum, E. Y. Puspaningrum, and B. M. Mulyo, "Perbandingan metode lexical dan semantic retrieval: Bm25, Indosbert baseline, dan Indosbert fine-tuned pada pencarian dokumen berbahasa Indonesia," in *Prosiding Seminar Nasional Penelitian dan Pengabdian Kepada Masyarakat LPPM Universitas 'Aisyiyah Yogyakarta*, 2026.

[16] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," *in Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics and 10th Int. Joint Conf. Nat. Lang. Process.*, pp. 843–857, 2020, doi: 10.18653/v1/2020.aacl-main.85.

[17] T. Suryati, Y. Wibisono, and Y. Wihardi, "Aplikasi deteksi plagiarisme dokumen skripsi dengan algoritma Rabin-Karp," *JATIKOM: Jurnal Teori dan Aplikasi Ilmu Komputer*, vol. 1, no. 2, pp. 91–95, 2018, doi: 10.17509/jatikom.v1i2.25190.

[18] Y. N. Fadziah, Rasim, and E. R. Fitrajaya, "Penerapan algoritma enhanced confix stripping dalam pengukuran keterbacaan teks menggunakan gunning fog index," *JATIKOM : Jurnal Teori dan Aplikasi Ilmu Komputer*, vol. 1, no. 1, pp. 14–22, 2018, doi: 10.17509/jatikom.v1i1.25143.

[19] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.

[20] C. Van Gysel and M. de Rijke, "Pytrec_eval: An extremely fast python interface to trec_eval," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2018, pp. 873–876. doi: 10.1145/3209978.3210065.

[21] F. A. Sinaga, M. Hidayat, and R. Djati, "Penggunaan redis untuk meningkatkan performance retrieve data dari database," *JATIKOM: Jurnal Teori dan Aplikasi Ilmu Komputer*, vol. 9, no. 1, pp. 15–21, 2026.