LOCAL BENCHMARKING OF SENTENCE EMBEDDING MODELS FOR RETRIEVAL-AUGMENTED GENERATION

MURAT KARAKAYA

Department of Software Engineering, TED University, Ankara, Turkey
E-mail: murat.karakaya@tedu.edu.tr

Abstract - Retrieval-Augmented Generation (RAG) systems heavily depend on high-quality embedding models to effectively retrieve semantically relevant text chunks from document corpora. Current benchmarking practices predominantly utilize static, general-purpose datasets, which may inadequately represent the nuances and specificities of domain-focused applications. To address this limitation, we introduce a fully automated benchmarking pipeline enabling practitioners to evaluate sentence-transformer embedding models directly on their customized document collections. Our system leverages synthetic query generation coupled with LLM-based automated relevance judgments, thus simulating realistic retrieval scenarios without manual annotation efforts. In our study, we benchmarked several prominent Sentence Transformers embedding models on a specialized technical corpus, rigorously analyzing their retrieval performance using five distinct metrics. The proposed pipeline further provides detailed statistical comparisons, visual performance diagnostics, and practical throughput assessments, significantly aiding in embedding model selection for real-world RAG system deployments.

Retrieval-Augmented Generation, Embedding Models, Sentence Transformers, Information Retrieval, Benchmarking, Synthetic Queries.

BACKGROUND, MOTIVATION, AND OBJECTIVE

Retrieval-Augmented Generation (RAG) systems are increasingly adopted in applications that require grounded and context-aware responses [1], [2], [3]. These systems typically utilize vector databases that store chunked representations of domain-specific document collections.

Practitioners developing RAG systems must select appropriate embedding models to transform document chunks into effective vector representations [4]. However, while several prominent benchmarks exist to evaluate embedding models, these generic datasets often fail to represent the specific semantic nuances and characteristics of a practitioner's local corpus. Consequently, an embedding model performing well on widely-used benchmarks may not necessarily deliver optimal performance within a specialized domain.

To mitigate this challenge, we propose a fully automated benchmarking pipeline that practitioners can execute locally on their own document corpus. Our approach synthesizes realistic queries using a large language model (LLM), employs the same LLM to automatically judge query-chunk relevance, and evaluates retrieval efficacy across multiple sentence-transformer models.

The central motivation behind our work is to empower researchers and practitioners to benchmark embedding models directly against their own corpus rather than relying solely on generalized datasets [5], [6]. The proposed pipeline emphasizes reproducibility and interpretability by producing comprehensive evaluation metrics and intuitive visualizations, effectively capturing the retrieval behavior and performance characteristics of each evaluated model.

STATEMENT OF CONTRIBUTION / METHODS

We present a modular pipeline explicitly designed for benchmarking sentence-transformer embedding models on arbitrary document corpora. The core components of this pipeline, illustrated in Figure 1, include:

- Document Processing: Source PDF documents are ingested and segmented into semantically coherent chunks using LangChain's RecursiveCharacterTextSplitter.
- Synthetic Query Generation: Queries that semantically align with randomly selected document chunks are automatically generated using Gemini 2.0 Flash.
- Automated Relevance Judgments: The same large language model (LLM) generates automated relevance assessments for retrieved chunks, establishing ground-truth relevance mappings (qrels) without the need for manual annotations.
- Embedding Model Evaluation: Several popular sentence-transformer models embed the processed corpus, and their retrieval performance is rigorously evaluated using standard metrics such as MAP@10, Recall@10, Precision@10,

NDCG@10, and MRR@10 via the SentenceTransformers evaluation framework integrated with the ChromaDB vector store.

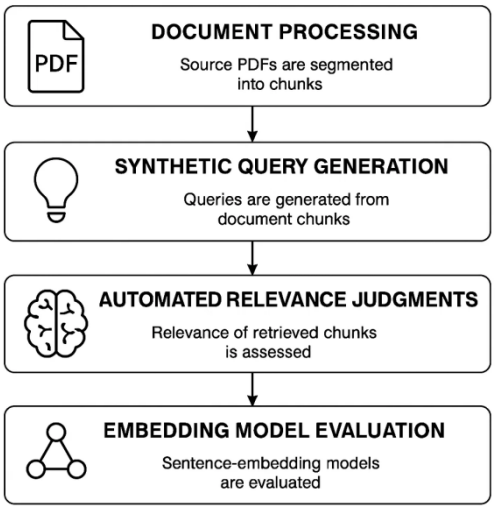

Figure 1: An automated pipeline for benchmarking sentence-transformer embedding models.

The embedding models selected for benchmarking are among the most downloaded and widely utilized on the Hugging Face platform [7]. Their prevalent use in both research and production environments underscores their suitability and relevance for this benchmarking task.

The pipeline is designed to closely simulate real-world RAG retrieval scenarios. By employing synthetic yet targeted queries, it robustly assesses each model's capability to accurately capture semantic similarities specific to the user-provided corpus.

DATASET DESCRIPTION

The evaluation corpus used in this study was constructed from six influential research papers on large language models and transformer architectures: Attention Is All You Need [8], DeepSeek-V3 Technical Report [9], LLaMA: Open and Efficient Foundation Language Models [10], Qwen2.5 Technical Report [11], Mistral 7B [12], and Gemma 3 Technical Report [13]. This carefully curated collection simulates realistic scenarios where practitioners seek to evaluate embedding models directly on specialized corpora that significantly differ from standard benchmarks.

The corpus comprised 588 document chunks extracted from technical PDFs using a chunk size of 1000 characters with an overlap of 200 characters. These chunks averaged 151.9 words per chunk, with lengths ranging from 33 to 381 words (standard deviation of 32.5). To emulate realistic user queries, 50 synthetic queries were generated, each averaging 19.7 words (range: 10 to 33 words). Relevance judgments were automated via LLM-based evaluation, producing 220 relevant query-document pairs and an average of 4.9 relevant documents per query. The resulting coverage ratio of 37.41% posed a meaningful retrieval challenge.

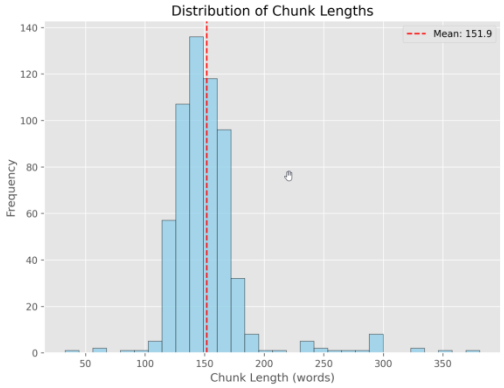

Figure 2: Corpus statistics illustrating the distribution of chunk lengths.

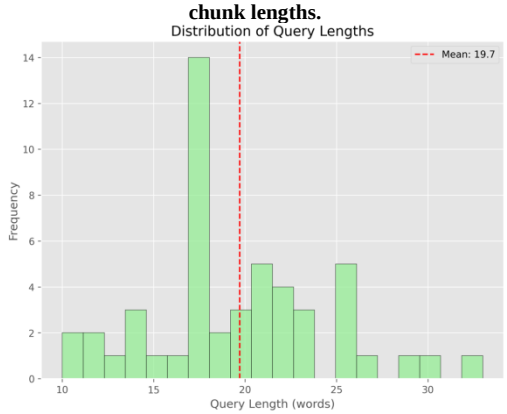

Figure 3: Corpus statistics illustrating the distribution of query lengths.

Figure 4: Corpus statistics illustrating the distribution of relevance density.

These visualizations provide valuable insights into the dataset's diversity and structural characteristics. Specifically, Figure 2 shows that while chunk lengths cluster around the mean, significant variation exists, potentially influencing embedding consistency. Figure 3 indicates a balanced distribution of query lengths, effectively representing a range of user information needs from concise to elaborate. Figure 4 highlights variability in the number of relevant documents per query, underscoring the complexity

inherent to the retrieval task. Such heterogeneity ensures robust evaluation, thoroughly assessing embedding models across varied retrieval scenarios.

SELECTED EMBEDDING MODELS

We evaluated four widely recognized sentence-transformer models that rank among the most frequently downloaded from the Hugging Face "sentence-transformers" library as of June 2025 [7]: all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-mpnet-base-v2, and paraphrase-multilingual-MiniLM-L12-v2. These models represent a strategic balance between performance, computational size, and multilingual capabilities, making them highly suitable for inclusion in our locally deployable benchmarking pipeline as seen at Table I.

The evaluated models are characterized as follows:

- all-MiniLM-L6-v2: A compact 6-layer transformer model optimized for efficient computation, particularly effective for sentence similarity tasks, clustering, and retrieval.
- all-MiniLM-L12-v2: An enhanced, deeper model variant that achieves improved retrieval accuracy without significantly increasing computational latency.
- all-mpnet-base-v2: Provides larger, 768-dimensional embeddings and was trained with contrastive learning objectives over one billion sentence pairs, thus ensuring superior semantic alignment and embedding quality.
- paraphrase-multilingual-MiniLM-L12-v2: Extends the capabilities of the MiniLM architecture to multilingual scenarios, ideal for tasks involving cross-lingual sentence embedding and retrieval.

While these models were selected for their prevalent use and established effectiveness in both research and production contexts, our benchmarking pipeline is highly configurable. Users can readily adapt the evaluation framework to include alternative or domain-specific sentence-transformer models by modifying the pipeline configuration file. This flexible design ensures the benchmarking process remains both adaptable and relevant to diverse user requirements and experimental contexts.
This approach facilitates repeatable and nuanced experimentation, enabling detailed analysis of embedding model performance across varying tasks and specific document collections.

Model	Embedding Dim	Size (params)	HF Downloads	Notes
all-MiniLM-L6-v2	384	~22M	90.4M	Fast, compact, high English accuracy
all-MiniLM-L12-v2	384	~33M	4.97M	Deeper version, optimized for speed
all-mpnet-base-v2	768	~110M	20.3M	Highest semantic quality, slower inference
paraphrase-multilingual-MiniLM-L12-v2	384	~118M	11M	Multilingual capability (supports 50+ languages)

Table I Overview of Evaluated Embedding Models

RESULTS, DISCUSSION, AND INSIGHTS

As explained above, all evaluations presented in this section are based on a sample domain-specific corpus constructed from research papers. These documents simulate a realistic scenario in which practitioners want to assess embedding models directly on their own corpus, which may differ significantly from standardized benchmarks.

Therefore, it is important to emphasize that the results and rankings reported here do not imply any global superiority of one embedding model over another. Rather, they serve as evidence that our proposed pipeline enables reliable and interpretable evaluation of sentence transformers tailored to a specific corpus. The utility lies in the ability to replicate this analysis across arbitrary corpora to determine the best-fitting model for a particular RAG deployment.

MODEL PERFORMANCE SUMMARY

The performance results in Table IIillustrate the retrieval effectiveness of four popular sentence-transformer models when applied to our corpus. The all-MiniLM-L6-v2 model achieves the highest performance across all five retrieval metrics, including MAP@10, Recall@10, and MRR@10.

However, this does not imply its global superiority; the takeaway is that this model aligns particularly well with the characteristics of our chosen dataset. The second-tier model, all-mpnet-base-v2, shows competitive performance in MRR and precision but lags behind in recall. The remaining models, especially the multilingual variant, underperform—likely due to a mismatch between their general-purpose multilingual training objectives and the specific structure of the technical English documents in our corpus.

Figure 5 provides a visual comparison, showing the strength of MiniLM-L6-v2across all axes. Importantly, the visualization highlights the nuance of trade-offs: some models perform better on precision or ranking, others on recall.

Model	MAP@10	Recall@10	Precision@10	NDCG@10	MRR@10
all-MiniLM-L6-v2	0.8264	1.0000	0.4889	0.9059	0.9259
all-mpnet-base-v2	0.4829	0.6522	0.2933	0.6137	0.7867
all-MiniLM-L12-v2	0.4353	0.6033	0.2800	0.5744	0.7699
paraphrase-multilingual-MiniLM-L12-v2	0.3823	0.5500	0.2400	0.5275	0.7788

Table II Retrieval metrics for each model

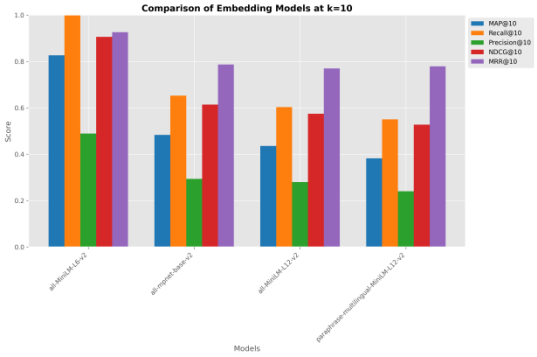

Figure 5: Comparison of retrieval performance across evaluated models.

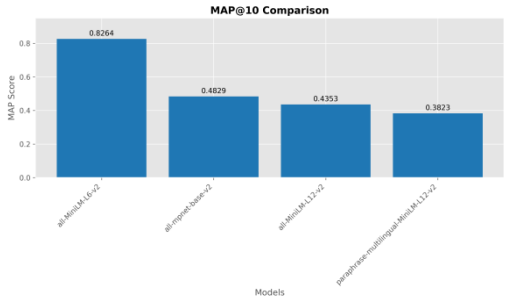

Figure 6: MAP@10 comparison across models. Differences exceeding 0.05 (dashed threshold line) are considered statistically meaningful.

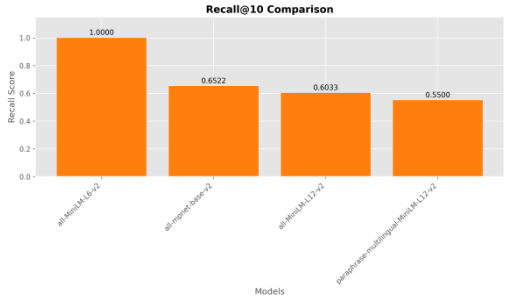

Figure 7: Recall@10 performance across models. The sharp contrast in recall between top and bottom models shows wide coverage variance.

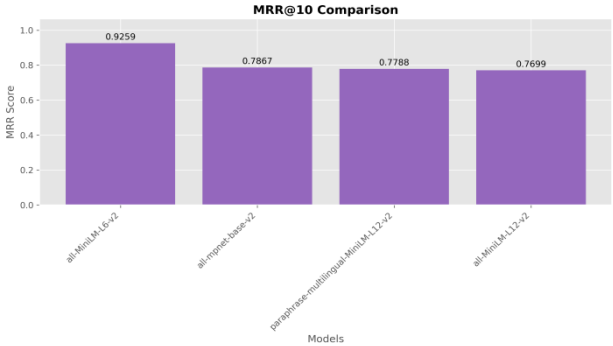

Figure 8: MRR@10 comparison: the high first-relevant-rank consistency of MiniLM-L6-v2 is evident.

STATISTICAL SIGNIFICANCE ANALYSIS

The deltas in Table III quantify the pairwise performance differences and reinforce the observations from the previous subsection. These differences, especially those exceeding 0.05, are considered

Comparison	MAP@10 Δ	Recall@10 Δ	MRR@10 Δ
MiniLM-L6-v2 vs mpnet-base-v2	0.3435	0.3478	0.1393
MiniLM-L6-v2 vs MiniLM-L12-v2	0.3911	0.3967	0.1560
MiniLM-L6-v2 vs multilingual	0.4441	0.4500	0.1471
mpnet-base-v2 vs MiniLM-L12-v2	0.0476	0.0489	0.0167
mpnet-base-v2 vs multilingual	0.1006	0.1021	0.0079
MiniLM-L12-v2 vs multilingual	0.0530	0.0532	0.0089

Table III pairwise performance differences (Δ) in MAP@10, RECALL@10, and MRR@10

practically and statistically meaningful in our evaluation. The most pronounced improvements are observed between MiniLM-L6-v2 and the multilingual model, especially in recall and MAP.

Figures 6–8 provide complementary visualizations, confirming the robustness of MiniLM-L6-v2 across multiple retrieval quality dimensions within this specific evaluation setup. These plots help identify model strengths and limitations visually and are integral to model selection using our framework.

EMBEDDING THROUGHPUT

Table 1 highlights the embedding throughput of each model. The MiniLM variants are all similarly fast, processing over 40 documents per second. In contrast, mpnet-base-v2 is significantly slower, taking over 140 seconds for the same task. This discrepancy may render it impractical for large-scale or latency-sensitive pipelines.
This analysis demonstrates that model choice is multi-dimensional. MiniLM-L6-v2 not only excels in retrieval metrics but also provides near-optimal processing efficiency, making it a strong candidate for deployment in environments where both quality and speed are critical.

Again, the takeaway is not that MiniLM-L6-v2 is universally optimal—it is that it best fits the structure, content, and goals of this specific corpus evaluation. This illustrates the pipeline's power to guide such decisions based on empirical local evidence.

DISCUSSION AND PRACTICAL IMPLICATIONS

This study presents a pragmatic, corpus-specific methodology for evaluating embedding models in Retrieval-Augmented Generation (RAG) systems. By leveraging synthetic query generation and LLM-based relevance assessments, our pipeline eliminates the need for manual annotation and enables fully localized benchmarking. The use of a technical, domain-specific corpus—comprising foundational model reports—demonstrates the pipeline's real-world applicability.

The principal contribution of this work lies in empowering practitioners with a plug-and-play evaluation tool. By simply supplying a document corpus, users can invoke an automated pipeline that performs document chunking, generates synthetic queries, constructs relevance judgments, and conducts comparative retrieval evaluations across multiple sentence-transformer models. This streamlined workflow allows users to identify the model that best suits their data characteristics and retrieval needs, avoiding reliance on generic benchmark leaderboards.

Our findings emphasize that embedding model performance is highly corpus-dependent. Models that rank highly on public leaderboards may underperform on domain-specific datasets, particularly in metrics such as recall, semantic relevance, or runtime efficiency. This variability reinforces the importance of conducting task-aligned evaluations to ensure downstream effectiveness.

This framework is especially valuable in settings where annotated data is scarce or user information needs evolve dynamically. Practical applications span enterprise knowledge retrieval, academic search systems, biomedical literature mining, legal document review, and more—any environment where retrieval quality must be adapted to specific content structures.

By promoting reproducible, efficient, and transparent model comparisons, our approach bridges the gap between abstract benchmarking and real-world deployment. It supports evidence-driven decision-making in embedding model selection, advancing robust and context-aware RAG system development.

CONCLUSION

This paper introduced a fully automated benchmarking pipeline for evaluating embedding

models in Retrieval-Augmented Generation (RAG) settings. Unlike static public benchmarks, our pipeline is tailored to user-supplied corpora and is capable of end-to-end processing—from document chunking to retrieval evaluation—using synthetic queries and automated relevance judgments.

The experimental results demonstrated that all-MiniLM-L6-v2 delivered the best overall performance in terms of MAP@10, Recall@10, and MRR@10, while also achieving efficient embedding throughput. Comparative analysis showed statistically significant differences between models, confirming that model selection has a measurable impact on retrieval outcomes.

A key insight from our work is that embedding models must be validated on the actual content they will serve. Performance varied widely across models when evaluated on our domain-specific corpus, underscoring the limitations of one-size-fits-all benchmark results for applied RAG workflows.

By supporting reproducible, domain-sensitive evaluation, our pipeline enables practitioners to make informed choices about embedding models suited to their specific documents and use cases. Future work will explore extending the pipeline's flexibility across query types and relevance standards to deepen interpretability and further support local benchmarking practices.

Future enhancements will focus on: expanding performance and efficiency metrics (e.g., memory usage, latency, and scalability); improving query analysis (including complexity and language-based performance); implementing robust statistical validation techniques; and introducing advanced visualizations for diagnostic insights. Additional improvements to infrastructure—such as multi-threading, checkpoint recovery, and detailed progress monitoring—will further improve usability and scalability. Finally, we plan to extend automated query generation workflows and support multilingual, multi-hop, and compositional information needs, increasing the pipeline's relevance to diverse RAG applications.

[1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks." 2021. Available: https://arxiv.org/abs/2005.11401

[2] S. Packowski, I. Halilovic, J. Schlotfeldt, and T. Smith, "Optimizing and evaluating enterprise retrieval-augmented generation (RAG): A content design perspective," in Proc. 8th int'l conf. On advances in artificial intelligence (ICAAI 2024), New York, NY, USA: Association for Computing Machinery, 2024, pp. 162–167. doi: 10.1145/3704137.3704181.

[3] D. Rau, S. Wang, H. Déjean, S. Clinchant, and J. Kamps, "Context embeddings for efficient answer generation in retrieval-augmented generation," in Proc. 18th ACM int'l conf. On web search and data mining (WSDM '25), New York, NY, USA: Association for Computing Machinery, 2025, pp. 493–502. doi: 10.1145/3701551.3703527.

[4] A. Afzal, J. Vladika, G. Fazlija, A. Staradubets, and F. Matthes, "Towards optimizing a retrieval augmented generation using large language models on academic data," in Proc. 8th int'l conf. On natural language processing and information retrieval (NLPIR 2024), New York, NY, USA: Association for Computing Machinery, 2025, pp. 250–257. doi: 10.1145/3711542.3711575.

[5] H. V. Tran, T. Chen, Q. V. H. Nguyen, Z. Huang, L. Cui, and H. Yin, "A thorough performance benchmarking on lightweight embedding-based recommender systems," ACM Transactions on Information Systems, vol. 43, no. 3, pp. 63:1–63:32, 2025, doi: 10.1145/3712589.

[6] B. Chen, J. Tackman, M. Setälä, T. Poranen, and Z. Zhang, "Integrating access control with retrieval-augmented generation: A proof of concept for managing sensitive patient profiles," in Proc. 40th ACM/SIGAPP symposium on applied computing (SAC '25), New York, NY, USA: Association for Computing Machinery, 2025, pp. 915–919. doi: 10.1145/3672608.3707848.

[7] H. Face, "Hugging face sentence-transformers model hub." https://huggingface.co/sentence-transformers, 2025.

[8] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[9] D. Team, "DeepSeek-V3 technical report," arXiv preprint arXiv:2412.19437, 2024.

[10] H. Touvron, T. Lavril, G. Izacard, et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

[11] Q. Team, "Qwen2.5 technical report," arXiv preprint arXiv:2412.15115, 2024.

[12] M. AI, "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.

[13] G. Team, "Gemma 3 technical report," arXiv preprint a Xiv:2503.19786, 2025.

★★★