Embedding Models for Hungarian Chatbots: Retrieval Performance and Efficiency Analysis

Csiszér Rolland-Zsombor, Margit Antal

Sapientia Hungarian University of Transylvania, Târgu Mureș, Romania

csiszer.rolland@student.ms.sapientia.ro manyi@ms.sapientia.ro

This study investigates the effectiveness and efficiency of different embedding models for retrieval-augmented generation (RAG) in the Hungarian language context. The research was motivated by the need to build a chatbot for ClearService, a company whose internal documentation is entirely in Hungarian. Since embedding model performance can vary significantly across languages, we systematically evaluated eight widely used multilingual and cross-lingual embedders. Using ChromaDB as the vector database, retrieval performance was tested on a set of 50 domain-specific Hungarian questions with evaluation metrics including Mean Reciprocal Rank (MRR), Recall@1, and Recall@3.

Results revealed a clear performance gap among models: BGE-M3 achieved the highest retrieval accuracy, followed by paraphrase-multilingual-MiniLM and OPENAI-3 Small. Beyond accuracy, we also performed a detailed timing analysis, including build time, search time, and per-query throughput. The paraphrase-multilingual-MiniLM model proved to be the most efficient, sustaining up to 40 queries per second (QPS) with minimal latency, making it attractive for real-time applications. In contrast, BGE-M3 and OPENAI-3 Small required higher computational resources due to their larger embedding dimensions but compensated with superior retrieval quality. These findings emphasize the trade-off between accuracy and efficiency: while BGE-M3 remains the most effective embedder for Hungarian-language retrieval, paraphrase-multilingual-MiniLM represents the fastest option for latency-sensitive, high-throughput deployments.

Keywords: Embeddings, Hungarian, RAG, chatbot, evaluation

References

[1] Antal, M., Buza, K. (2025) Evaluating Open-Source LLMs in RAG Systems: A Benchmark on Diploma Theses Abstracts Using Ragas. *Acta Univ. Sapientiae Inform. 17, 5 (2025).* https://doi.org/10.1007/s44427-025-00006-3

[2] Reimers, Nils and Gurevych, Iryna (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.*

[3] Jianlv Chen and Shitao Xiao and Peitian Zhang and Kun Luo and Defu Lian and Zheng Liu (2024) BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. https://arxiv.org/abs/2402.03216