Evaluation of Embedding Models for Hungarian Question-Answer Retrieval on Domain-Specific and Public Benchmarks

Margit Antal

***Abstract*—Embedding models have become a fundamental component of modern natural language processing, yet their performance in morphologically rich, low-resource languages such as Hungarian remains underexplored. In this paper, we present a systematic evaluation of state-of-the-art embedding models for Hungarian question–answer retrieval. We construct two complementary evaluation datasets: (i) a domain-specific corpus collected from company documentation, preprocessed into topical chunks with human-verified question–answer pairs and (ii) the publicly available HuRTE benchmark. Using Chroma as the vector database, we compare eight multilingual and cross-lingual embedding models alongside keyword-based search baseline.**

Performance is measured using Mean Reciprocal Rank (MRR) and Recall@k. Results show substantial variation across models and datasets, with notable differences between domain-specific and general-purpose retrieval tasks. BGE-M3 and XLM-ROBERTA achieved the highest accuracy (MRR: 0.90) on the Clearservice dataset, while GEMINI demonstrated superior performance on HuRTE (MRR: 0.99). We complement the evaluation with comprehensive error analysis, highlighting challenges posed by Hungarian domain-specific terminology, synonyms, and overlapping topics, and discuss trade-offs in efficiency through index build time and query latency measurements. Our findings provide a comparative study of embedding-based retrieval in Hungarian, offering practical guidance for downstream applications and setting a foundation for future research in Hungarian representation learning. The dataset and the corresponding evaluation code are publicly accessible at https://github.com/margitantal68/hungarian-embeddings.

***Index Terms*—Hungarian language, embedding models, question-answer retrieval, vector similarity search.**

I. Introduction

Embedding models have become fundamental to modern natural language processing (NLP), providing dense vector representations that encode semantic relationships between words, phrases, and documents. The evolution of embedding techniques has progressed from early feedforward neural networks to static models like Word2Vec [1] and FastText [2], and subsequently to dynamic, contextualized embeddings derived from Transformer-based architectures including BERT [3], GPT [4], and T5 [5]. These advances have significantly improved performance across numerous NLP tasks, particularly in high-resource languages such as English.

However, the effectiveness of modern embedding models in morphologically rich, low-resource languages remains insufficiently explored. Hungarian exemplifies these challenges due to its agglutinative morphology, where words can contain multiple morphemes that substantially alter meaning and grammatical function. This morphological complexity often results in performance degradation compared to English [6].

The importance of robust embeddings extends beyond traditional NLP tasks to modern applications such as Retrieval-Augmented Generation (RAG) systems, where embeddings enable efficient knowledge retrieval from large databases by capturing semantic similarity beyond surface-level keyword matching. In these systems, embedding quality directly impacts retrieval accuracy and, consequently, the overall system performance. Despite this critical role, systematic evaluation of state-of-the-art embedding models for Hungarian remains limited.

Previous work on Hungarian embeddings has been sparse and focused primarily on static representations. Gedeon [7] presented the most comprehensive evaluation to date, but concentrated exclusively on static word embeddings, leaving modern contextualized models largely unexplored. To the best of our knowledge, only a single study [8] to date has systematically evaluated embedding models for Hungarian texts, focusing exclusively on the legal domain. However, no comprehensive assessment has yet been conducted for other types of Hungarian texts.

This paper addresses these research gaps by presenting the first comprehensive evaluation of state-of-the-art embedding models for Hungarian question-answer retrieval. Our primary contributions are threefold: (1) we provide a systematic comparison of modern embedding models on Hungarian retrieval tasks, (2) we establish evaluation benchmarks using both domain-specific and general-purpose datasets, and (3) we offer practical guidance for selecting appropriate models for Hungarian NLP applications.

To achieve these objectives, we construct two complementary evaluation datasets. The first comprises domain-specific data extracted from technical documentation, preprocessed into semantically coherent chunks with human-annotated question-answer pairs. The second utilizes the publicly available HuRTE benchmark [9], providing standardized evaluation conditions. We employ Chorma [10] for efficient vector storage and similarity search.

Our evaluation examines eight diverse embedding models, including multilingual transformers (BGE-M3, E5-BASE, XLMROBERTA, NOMIC), language-specific models (HU-BERT), and commercial API solutions (OpenAI, Google). Performance is assessed using established information retrieval metrics: Mean Reciprocal Rank (MRR) and Recall@k. Beyond quantitative analysis, we conduct detailed error analysis to identify failure patterns related to Hungarian morphology, compound word processing, and domain-specific terminology. Additionally, we analyze practical considerations including inference latency, computational requirements, and cost-effectiveness to provide comprehensive guidance for practi-

tioners.

The remainder of this paper is organized as follows: Section 2 reviews related work in multilingual embeddings and Hungarian NLP; Section 3 details our experimental methodology and datasets; Section 4 presents quantitative results and comparative analysis; Section 5 discusses error patterns and morphological challenges; Section 6 discusses the results, highlighting practical trade-offs and deployment considerations, while Section 7 concludes by outlining the implications for future Hungarian NLP research.

II. Related Work

Teaching machines to comprehend human language is a fundamental step in developing intelligent systems, a task often facilitated by word embeddings. These dense vector representations map similar words to similar vectors and are capable of capturing complex semantic relationships. The field has evolved from early feedforward neural networks for language modeling to highly effective static models like Word2Vec and FastText. However, a key limitation of these static embeddings is their inability to capture context-dependent meanings, leading to the development of dynamic, contextualized word embeddings from Transformer-based models such as BERT, GPT, and T5.

For the Hungarian language, which is considered an underrepresented language due to its complex morphology and agglutinative nature, high-quality embedding models are insufficiently evaluated. Few empirical measurements exist to assess embedding model performance specifically for Hungarian, making it difficult for developers of Hungarian Q&A systems to determine which models are best suited for their applications.

Despite the success of dynamic models, static word embeddings remain relevant for various applications due to their lower computational requirements. Research has shown a significant performance drop in Hungarian word analogy tasks compared to English, attributed to the language's high morphological variation and less stable semantic representations.

Several studies have focused on evaluating word embeddings in Hungarian. Gedeon [7] provides a comprehensive analysis of various static word embeddings, including traditional models like Word2Vec and FastText, as well as static embeddings derived from BERT-based models using different extraction methods. For intrinsic evaluation using a word analogy task (measuring embedding quality without using them in a real application), FastText demonstrated superior performance, achieving high accuracy and Mean Reciprocal Rank (MRR) scores. Among the BERT-based models, the X2Static method for extracting static embeddings showed superior performance compared to decontextualized and aggregate methods, approaching the effectiveness of traditional static embeddings. This method leverages contextual information from a teacher model to generate static embeddings, and a Turkish study similarly found X2Static to be the most effective for extracting static embeddings from BERT-based models. For extrinsic evaluation (test embeddings in a real application, such as NER or POS tagging), Gedeon utilized a bidirectional LSTM model for Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tasks. The results indicated that embeddings derived from dynamic models, particularly those extracted using the X2Static method, outperformed purely static embeddings. ELMo embeddings achieved the highest accuracy in both NER and POS tagging, highlighting the benefits of contextualized representations even when used in a static form. ELMo generates contextualized word embeddings using a bidirectional LSTM language model, capturing polysemy and context-dependent meanings.

BERT (Bidirectional Encoder Representations from Transformers) [3] and its derivatives have become central to modern NLP. For Hungarian, huBERT is a state-of-the-art Hungarian cased BERT-base model trained on the Webcorpus 2.0. It has been shown to outperform multilingual BERT models in tasks such as morphological probing, POS tagging, and NER. Nemeskey [11] introduced the huBERT family, which achieved state-of-the-art performance in NER and NP chunking for Hungarian. Another significant Hungarian BERT model is PULI BERT-Large [12], a BERT large model with 345 million parameters. XLM-RoBERTa (XLM-R) is a transformer-based multilingual masked language model that also includes Hungarian data.

The sentence transformers method has gained popularity for creating semantically meaningful sentence embeddings that enable comparison using cosine similarity. This approach, sometimes extended to multilingual models using knowledge distillation, involves a teacher model generating desired sentence embeddings in one language, which a student model then replicates across multiple languages using parallel sentences. Hatvani and Yang [13] addressed the lack of high-quality embedding models for Hungarian in RAG systems. They developed three encoder-only language models: xml_roberta_sentence_hu, hubert_sentence_hu, and minilm_sentence_hu. These models, trained using a distillation method with paraphrase-distilroberta-base-v2 as the teacher model and FLORES-200 and OpenSubtitles corpora, demonstrated substantial improvements in semantic similarity tasks. The hubert_sentence_hu model achieved the highest accuracy and F1-Score on a custom news article test corpus.

Beyond general NLP tasks, Hungarian embedding models have been applied and evaluated in specific domains. Osváth et al. [14] used BERT topic modeling with huBERT and HIL-SBERT embeddings to analyze patient narratives from a Hungarian online forum, identifying major topics and using a fine-tuned BERT model for sentiment analysis. Their findings highlighted dominantly negative sentiments in patient experiences and comments.

Yang and Váradi [15] explored developing deep neural network language models for Hungarian with low computational and data resources. They pre-trained and fine-tuned five transformer models: ELECTRA, ELECTRIC, RoBERTa (small), BART (base), and GPT-2 on various NLP tasks, including sentence-level sentiment analysis, NER, noun phrase chunking, and text summarization. While these experimental models generally did not surpass the state-of-the-art huBERT model in classification tasks, they achieved competitive results with fewer parameters and resources. Notably, their BART

model achieved a significantly higher F-score in abstractive summarization compared to huBERT-based tools, and the models offered advantages in terms of smaller carbon footprint and mobile application suitability.

Tóth et al. [16] developed LMEZZ, a learning application to help students with Hungarian sentence analysis based on school grammar rules, utilizing transformer-based BERT models (huBERT and PULI BERT-Large) for improved reliability over convolutional neural network-based SpaCy models.

A recent study [8] presents a semantic search system developed to efficiently identify Hungarian court decisions with similar factual backgrounds. Its primary objective is to retrieve relevant legal precedents by matching court rulings based on semantic similarity, using factual case summaries as queries. The research evaluated twelve embedding models on a corpus of 1,172 Hungarian court decisions. Given that legal documents are typically lengthy—often exceeding the context window of most transformer-based architectures—the authors examined seven different strategies for handling long texts, including simple chunking, striding (overlapping chunks), and Last Chunk Scaling (LCS), which mitigates the overrepresentation of small final segments in the averaged embedding vector. Model performance was assessed using the Mean Reciprocal Rank (MRR) metric. The study found that the Cohere embed-multilingual-v3.0 model achieved the best results, reaching an MRR of 0.95. Notably, this demonstrates that a well-optimized 512-token model can outperform several models with substantially larger context windows (up to 8192 tokens). The authors also evaluated models pretrained specifically for the Hungarian language, including the base huBERT model [11] without fine-tuning, as well as two adapted variants: the sbert_hubert model [14], fine-tuned for sentence-level semantic similarity, and the danieleff model, fine-tuned for question–answering (Q&A) tasks. The danieleff model was trained on 170 question–answer pairs derived from sections (1,000–5,000 characters) of university academic regulations. Among the Hungarian-language models, danieleff achieved the highest performance.

Modern multilingual embedding models have made significant progress in understanding multiple languages simultaneously. Current state-of-the-art models can work with more than 100 languages and perform well on standard evaluation benchmarks [17]. Transformer-based models, especially BERT, XLM-R, and XLM-RoBERTa, have become the most widely used approaches for this task. Several key innovations have improved these models. First, researchers developed methods to adapt models trained on one language to work with others by creating specialized word representations [18]. Second, they created systems that can understand sentences across languages by sharing vocabulary encoding methods, enabling models to work on new languages without additional training [19]. Third, they combined different types of embeddings with improved alignment techniques to better match meanings across languages [20]. These advances have led to efficient multilingual systems that can process long texts (up to 8192 tokens) while maintaining good performance across different language tasks [17]. However, these models have not been thoroughly tested on morphologically complex languages like Hungarian. Most evaluations focus on widely-used languages, which may not reveal the challenges that arise with Hungarian's complex word structure and limited available training data.

III. Methods

A. Embeddings

Text embeddings are numerical representations of words, sentences, or documents in a continuous vector space. They capture semantic meaning, so texts with similar meanings end up close together in that space, even if they use different wording. In RAG systems, embeddings are crucial because they enable efficient retrieval of relevant knowledge from large databases. Instead of relying only on keyword matching, embeddings allow the system to understand context and intent, leading to more accurate and meaningful results.

Multilingual embeddings extend this capability across languages, mapping semantically similar texts in different languages to nearby positions in the same vector space. This makes it possible for a RAG system to retrieve knowledge in one language and use it to answer questions in another, breaking down language barriers and improving accessibility. In practice, high-quality multilingual embeddings are essential for building global, cross-lingual RAG applications that can serve diverse users and knowledge sources.

Embedder models were utilized in three ways: commercial models accessed via their APIs, and open-source models run either through a local Ollama server or via SentenceTransformers [21], a Python library for generating dense vector representations (embeddings) of sentences, paragraphs, and documents.

In this paper we employed the following embedder models:

- BGE-M3 – bge-m3 [22]: This model offers robust embeddings for multilingual and general-purpose semantic tasks, with emphasis on large-scale retrieval and clustering.
- E5-BASE – intfloat/multilingual-e5-base [23]: The Multilingual E5-Base model is a transformer-based text embedder that generates semantically rich, language-agnostic sentence embeddings across over 100 languages, enabling effective multilingual retrieval, clustering, and semantic similarity tasks.
- GEMINI – gemini-embedding-001 [24]: Text embeddings were generated using the *Gemini-embedding-001* model, which produces 768-dimensional vectors. To ensure consistency with other embedding models in our evaluation, the input-type parameter was not specified. The embeddings were obtained through the Gemini API's v1beta endpoint.
- HUBERT – danieleff/hubert-base-cc-sentence-transformer [8]: This model model was fine-tuned on 170 Hungarian question–answer pairs derived from sections of university academic regulations ranging from 1,000 to 5,000 characters in length.
- NOMIC – nomic-embed-text-v1 [25]: Designed for general-purpose embeddings, NOMIC excels in large-scale retrieval, clustering, and semantic search tasks with high efficiency.

TABLE I
COMPARISON OF POPULAR EMBEDDING MODELS BY USAGE MODE, DIMENSION, SEQUENCE LENGTH, DOMAIN, AND SIZE.

Model Name	Usage	Dimension	Sequence Length (#tokens)	Domain	Model Size
BGE-M3	SentenceTransformer - local	1024	8192	Multilingual	≈ 560M
E5-BASE	SentenceTransformer - local	768	512	Multilingual	≈ 278M
GEMINI	Google API	768	2048	General	Undisclosed
HUBERT	SentenceTransformer - local	768	512	Hungarian	≈ 110M
NOMIC	Ollama - local	768	2048	General	≈ 137M
OPENAI-3SMALL	OpenAI API	1536	8192	General	Undisclosed
OPENAI-ADA	OpenAI API	1536	8192	General	Undisclosed
XLMROBERTA	SentenceTransformer - local	768	128	Multilingual	≈ 270M

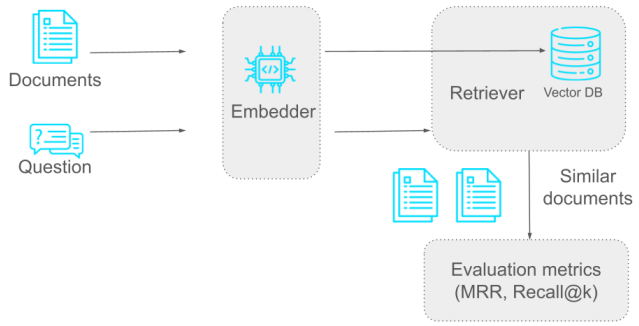

Fig. 1. Evaluation pipeline

- OPENAI-3SMALL – text-embedding-3-small: A lightweight and cost-effective embedding model from OpenAI's API v1, optimized for production-scale semantic tasks where efficiency and performance must be balanced.
- OPENAI-ADA – text-embedding-ada-002: A versatile embedding model available through OpenAI's API v1, widely adopted for applications such as semantic search, clustering, classification, and recommendation systems, supporting a broad range of text inputs.
- XLMROBERTA – paraphrase-xlm-r-multilingual-v1 [26], [27]: Tis model is a multilingual Sentence-Transformer based on XLM-RoBERTa, designed to produce high-quality, language-agnostic sentence embeddings for over 50 languages, optimized for tasks like semantic similarity and multilingual retrieval.

Table I presents the key characteristics of the embedding models.

B. Evaluation pipeline

The evaluation pipeline is illustrated in Fig. 1. During the ingestion stage, documents are transformed into vector representations using an embedding model and subsequently stored in a vector database, specifically Chroma in our implementation. Following ingestion, the evaluation of a question dataset proceeds through three steps: (1) vectorization of the input questions, (2) retrieval of the top-k most similar documents for each question, and (3) computation of retriever performance metrics.

As a baseline for comparison with semantic search, we incorporated a keyword-based retrieval method. Specifically, we employed the BM25 [28] algorithm to retrieve the top-k most relevant documents.

C. Datasets

1) Clearservice: The Clearservice dataset[1] is a custom-made dataset created from the data of the company of the same name. The dataset consists of two parts: (1) A file called topics.txt, which groups the data into topics and serves as the search space. (2) A set of questions in cs_qa.csv, containing 50 questions. Each question is associated with a specific topic and can be answered based on it. For each question, the corresponding topic and a reference answer are provided. The reference answer, however, is not used in this study.

2) HuRTE: The HuRTE dataset[2] is the Hungarian adaptation of the Recognizing Textual Entailment (RTE) corpora originally included in the GLUE benchmark. It forms part of the Hungarian Language Understanding Evaluation Benchmark Kit (HuLU) [29] [9] and was created through translation and re-annotation of the English RTE instances. The dataset consists of 4,504 examples, each comprising a premise — sometimes a multi-sentence passage — and a single-sentence hypothesis, with the task being to determine whether the premise entails the hypothesis. This is framed as a binary classification problem, where labels indicate entailment ("1") or non-entailment ("0"). The corpus is divided into training (2,132 instances), validation (243 instances), and test splits; however, test labels are not provided. The data is distributed in JSON format, with each entry containing an identifier, a premise, a hypothesis, and the corresponding label.

We measure the quality of retrieval using two types of evaluations. The *HuRTE-Positive* evaluation is performed using only the positive examples (label 1) both in the index and in the question evaluation. In this setting, the training set contains 1,092 positive examples, and the validation set contains 135. The hypothesis sentences are searched for within the premise texts, and retrieval quality is assessed accordingly.

The *HuRTE-All* evaluation is performed using all examples in the index, while still evaluating the questions using only the positive examples. This allows us to assess retrieval performance in a more realistic setting, where irrelevant data

[1]https://github.com/margitantal68/hungarian-embeddings/tree/master/data/clearservice
[2]https://github.com/nytud/HuRTE

is present in the index, but only the positives matter for evaluation.

D. Metrics

We evaluated retrieval performance using Mean Reciprocal Rank (MRR) and Recall. MRR measures the rank position of the first relevant document, averaged across all queries. For each (query, document) pair, documents were retrieved using semantic search in Chroma, and the rank of the corresponding ground-truth context was recorded. Recall@1 and Recall@3 capture the proportion of queries for which the correct context appears within the top one or top three retrieved results, respectively.

IV. Results

All measurements were conducted on a MacBook Pro equipped with an Apple M1 Pro processor and 32 GB of unified memory, running macOS Sequoia version 15.7.1. The experiments involving the Nomic embedder utilized the Ollama runtime (version 0.12.6).

TABLE II
EMBEDDING MODELS EVALUATION ON CLEARSERVICE DATASET.

Embedder	MRR	Recall@1	Recall@3
BGE-M3	0.90	0.86	0.96
E5-BASE	0.79	0.70	0.92
GEMINI	0.87	0.78	0.98
HUBERT	0.78	0.74	0.84
NOMIC	0.71	0.64	0.80
OPENAI-3SMALL	0.80	0.70	0.94
OPENAI-ADA	0.80	0.72	0.90
XLMROBERTA	0.90	0.86	0.96
BM25	0.77	0.68	0.80

TABLE III
COMPARISON OF MODELS ON HURTE DATASET *HuRTE-Positive* EVALUATION.

Model	MRR		Recall@1		Recall@3	
	Val	Train	Val	Train	Val	Train
BGE-M3	0.98	0.89	0.96	0.82	1.00	0.97
E5-BASE	0.93	0.84	0.90	0.77	0.97	0.92
GEMINI	0.99	0.91	0.97	0.85	1.00	0.98
HUBERT	0.82	0.63	0.77	0.53	0.88	0.74
NOMIC	0.90	0.72	0.85	0.65	0.95	0.80
OPENAI-3SMALL	0.94	0.85	0.92	0.78	0.97	0.92
OPENAI-ADA	0.94	0.84	0.91	0.78	0.98	0.92
XLMROBERTA	0.94	0.82	0.91	0.75	0.98	0.91
BM25	0.82	0.72	0.78	0.64	0.84	0.79

We applied our evaluation pipeline to both the Clearservice and HuRTE datasets. For HuRTE, we conducted two types of evaluations: *HuRTE-Positive*, which uses only the positive examples in both the index and the evaluation, and *HuRTE-All*, which uses all examples in the index while evaluating only the positive questions. Each type of evaluation was performed separately on the validation set (243 samples, 135 positives) and the training set (2132 samples, 1092 positives), allowing us to analyze how performance generalizes from a smaller dataset to a larger one of the same type.

The evaluation protocol is described in III-B. The results are summarized in the following tables: Table II presents the outcomes for the Clearservice dataset, while Table III reports the results for the HuRTE-Positive dataset.

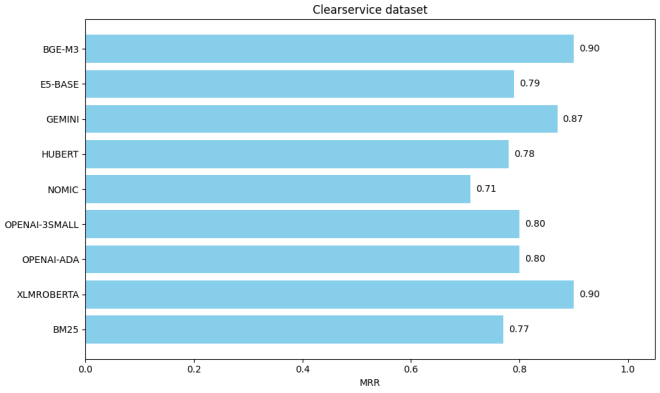

Fig. 2. Models' performance on the Clearservice dataset using the MRR metric.

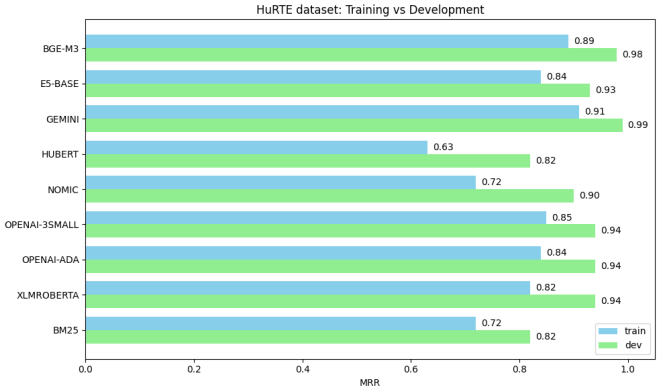

Fig. 3. Model performance on the HuRTE dataset (*HuRTE-Positive* evaluation) using the MRR metric.

Among the available metrics, MRR (Mean Reciprocal Rank) was chosen for visual representation, as it reflects both the position and relevance of the first correct result, providing a more informative measure of retrieval effectiveness than Recall@1 or Recall@3. Figs. 2, 3, and 4 show visual representations of the results.

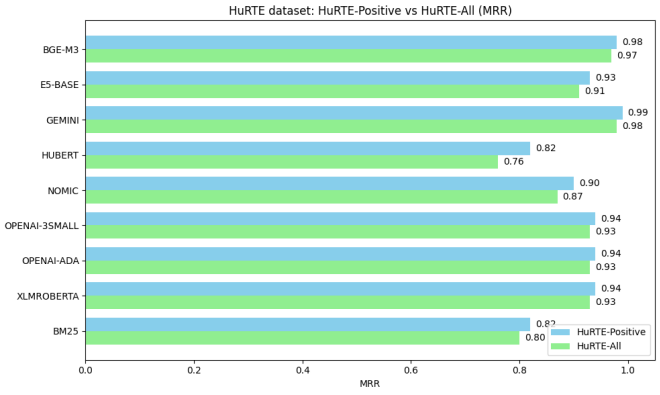

Fig. 4. Model performance on the HuRTE dataset (validation subset), comparing *HuRTE-Positive* and *HuRTE-All* evaluations using the MRR metric.

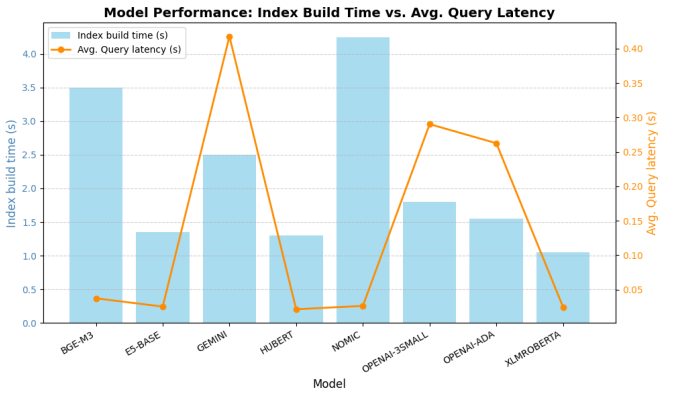

Fig. 5. Index build time vs. average query latency evaluated on the HuRTE dataset, validation subset, (*HuRTE-Positive* evaluation).

A. Analysis of Model Performance

Two types of time measurements were performed: index build time, representing the time required to create the model index, and average query latency, indicating the response time per query. The results are shown in Fig. 5.

HUBERT and XLMROBERTA achieved the best overall time performance, with both low build times and minimal latency. E5-BASE also performed efficiently across both metrics. GEMINI, OPENAI-3SMALL, and OPENAI-ADA exhibited notably higher query latencies despite moderate build times, likely due to API communication overhead. NOMIC, run locally via Ollama, and BGE-M3 had longer index build times but maintained low query latency. Overall, HUBERT and XLMROBERTA demonstrated the best balance between index setup efficiency and query responsiveness.

We evaluated both the retrieval quality and efficiency of embedding models, which are critical for performance in RAG systems. By comparing MRR (ranking quality) against query latency and index build time, this reveals how effectively each model balances accuracy with speed. Models positioned toward the top-left of the diagram achieve the best trade-off, offering high-quality retrieval with minimal response time. The results are shown in Fig. 6.

V. Error Analysis

We conducted a comprehensive error analysis on the Clearservice dataset. The questions and their corresponding error rates are presented in Appendix A.

Our analysis revealed that some errors are systematic—where seven out of eight embedding models failed—while others are occasional. The failures can be broadly categorized into three groups: (i) synonyms and paraphrases - The embeddings failed to recognize equivalence between different phrasings of the same concept. (ii) Overlapping topics - Relevant information appears across multiple sections, leading to confusion between semantically related topics. (iii) Domain-specific terminology - Specialized vocabulary (e.g., HR or legal terms) was not consistently captured by the embeddings.

In the following, we analyze the questions with the highest error rates: Q23 and Q9.

Q23: *Milyen elvárás van a munkavégzéssel kapcsolatban?*(What are the expectations regarding the work?) - error_rate = 0.875. This question includes the terms elvárás (expectation) and munkavégzés (work). The correct match is Topic 6 – *Munkavégzés* (Work), but Topic 9 – *Elvárások és Dokumentáció* (Expectations and Documentation) is also semantically close, even though it refers to application requirements rather than work performance. This semantic proximity likely caused confusion among the embedding models.

Q9: *Mi jár vasárnapi és ünnepnapi munkára?* (What is provided for Sunday and holiday work?) — error_rate = 0.625. The ground-truth topic is Topic 2 – *Fizetés* (Salary). The question is framed in terms of benefits (mi jár), while the relevant text specifies compensation percentages. Embeddings may incorrectly associate it with Topic 10 – *Szabadság és Hazautazás* (Vacation and Travel Home) due to the lexical overlap with ünnep (holiday).

To further investigate these confusions, we computed the Recall@3 confusion matrix, which measures how often embedding models retrieved the correct topic among their top three results. For each question, the ground-truth topic was compared against the top three retrieved topics, and the results were aggregated into a matrix with true topics as rows and retrieved topics as columns. Off-diagonal entries highlight frequent mismatches between semantically related topics.

By visualizing this matrix as a heatmap, we identified which topics are most frequently confused, revealing systematic weaknesses such as overlapping categories, synonym mismatches, and domain-specific ambiguities. Confusion matrices for all embedding models are shown in Appendix A.

VI. Discussion

Our evaluation of eight embedding models and BM25 across two Hungarian-language datasets reveals important insights into retrieval performance for domain-specific applications.

BGE-M3 and XLMROBERTA emerged as top performers on the Clearservice dataset, both achieving an MRR of 0.90 and Recall@1 of 0.86. GEMINI demonstrated the strongest performance on HuRTE-Positive (MRR: 0.99 validation, 0.91 training), followed closely by BGE-M3 (MRR: 0.98 validation, 0.89 training). The consistent performance gap between validation and training sets suggests that model behavior generalizes well from smaller to larger datasets of similar characteristics.

The traditional BM25 baseline achieved competitive results (MRR: 0.77 on Clearservice), outperforming NOMIC and matching HUBERT on certain metrics, demonstrating that lexical matching remains valuable for Hungarian text retrieval. However, neural embedding models consistently surpassed BM25, particularly on Recall@3 metrics.

Efficiency analysis revealed critical trade-offs between accuracy and speed. HUBERT and XLMROBERTA offered the best balance, with low index build times and minimal query latency. While GEMINI achieved superior retrieval quality, it exhibited significantly higher query latency due to API communication overhead, making it less suitable for real-time applications. BGE-M3, despite longer index build times, maintained competitive query latency while delivering top-tier accuracy.

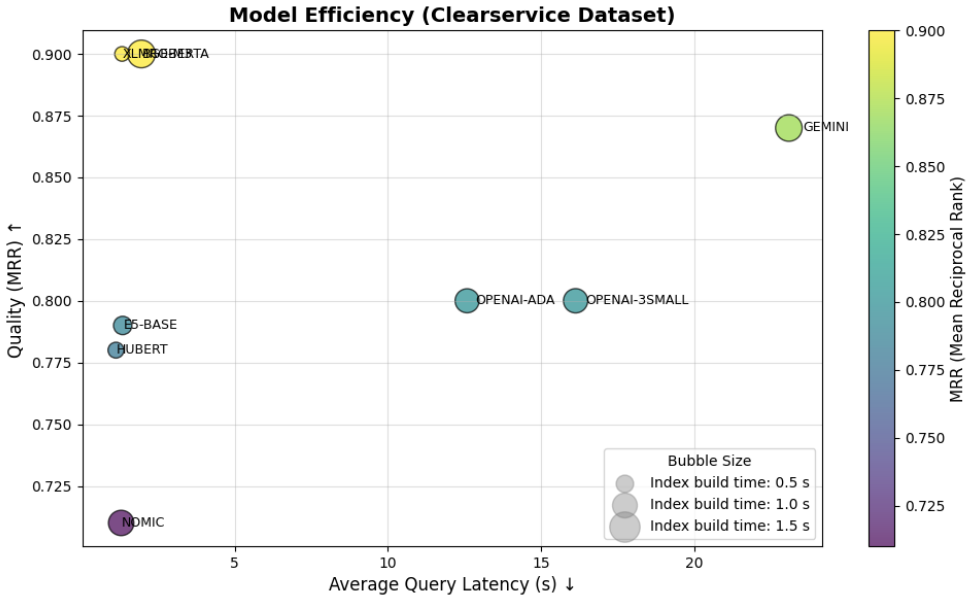

Fig. 6. Model efficiency on the Clearservice dataset using the MRR metric.

The error analysis on Clearservice exposed three primary failure modes: (i) synonyms and paraphrases, where embeddings failed to recognize semantic equivalence; (ii) overlapping topics, particularly when relevant information spans multiple sections; and (iii) domain-specific terminology, especially HR and legal vocabulary.

Systematic errors, where seven of eight models failed, highlight fundamental limitations in capturing Hungarian domain-specific semantics. The confusion between Topics 6 and 9 (Q23) and between Topics 2 and 10 (Q9) demonstrates that lexical overlap and semantic proximity can mislead even state-of-the-art embeddings. The Recall@3 confusion matrices further confirm these patterns, revealing consistent misclassifications between semantically related topics.

VII. Conclusions

In this paper, we conducted a comprehensive analysis of embedding models for Hungarian texts. Eight embedding models were evaluated against a baseline lexical search in the context of an information retrieval task for a Q&A system.

Our findings demonstrate that BGE-M3 and XLM-ROBERTA offer the best overall performance for Hungarian text retrieval, balancing high accuracy (MRR: 0.90) with operational efficiency. While GEMINI achieves superior accuracy, it comes at the cost of increased latency, making the choice between these models dependent on specific application requirements.

For production RAG systems, the trade-off between accuracy and speed is critical. HUBERT and XLMROBERTA provide optimal latency profiles, while BGE-M3 offers a strong middle ground for applications that can tolerate longer index build times in exchange for improved retrieval quality. This efficiency analysis is particularly valuable for practitioners deploying real-time information retrieval systems.

The systematic errors observed across models indicate that domain-specific Hungarian terminology and subtle semantic distinctions remain challenging for current embedding approaches. These persistent challenges suggest that future work should focus on fine-tuning strategies that explicitly incorporate domain knowledge and synonym relationships to better capture the nuances of specialized vocabulary.

While the evaluation provides valuable insights into Hungarian embedding-based retrieval, the limited size of the Clearservice dataset and the entailment-only focus of the HuRTE subset constrain the generalizability of the results. Future work will address these limitations by expanding the domain-specific dataset and incorporating more diverse and balanced retrieval benchmarks to ensure broader applicability.

Declarations

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the author used ChatGPT in order to refine language. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceedings of Workshop at ICLR*, vol. 2013, 01 2013.

[2] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431. [Online]. Available: https://aclanthology.org/E17-2068/

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., 2018, openAI Technical Report. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019. [Online]. Available: https://arxiv.org/abs/1910.10683

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2017. [Online]. Available: https://arxiv.org/abs/1607.04606

[7] M. Gedeon, "A comparative analysis of static word embeddings for hungarian," *Infocommunications Journal*, vol. XVII, pp. 28–34, 06 2025. [Online]. Available: https://doi.org/10.36244/ICJ.2025.2.4

[8] G. M. Csányi, D. Lakatos, J. P. Vadász, D. Nagy, and I. Üveges, "A kontextusablakon kihajolni nem veszélyes: jogi szövegek hatékony szemantikus keresése," in *Proceedings of the XXI. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary, February 6–7 2025, pp. 3–17, mONTANA Knowledge Management Ltd.; National University of Public Service.

[9] N. Ligeti-Nagy, G. Ferenczi, E. Héja, L. J. Laki, N. Vadász, Z. G. Yang, and T. Váradi, "HuLU: Hungarian language understanding benchmark kit," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 8360–8371. [Online]. Available: https://aclanthology.org/2024.lrec-main.733

[10] ChromaDB Team, "Chroma: Open-source embedding and vector database," https://github.com/chroma-core/chroma, 2023, version 1.3.0. Available at https://www.trychroma.com. Licensed under Apache 2.0.

[11] D. M. Nemeskey, "Introducing hubert," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, 2021, pp. 3–14.

[12] Z. G. Yang, R. Dodé, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, Kőrös, L. J. Laki, N. Ligeti-Nagy, N. Vadász, and T. Váradi, "Jönnek a nagyok! bert-large, gpt-2 és gpt-3 nyelvmodellek magyar nyelvre," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*. Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 247–262.

[13] P. Hatvani and Z. Yang, "Training embedding models for hungarian," in *Proceedings of the 3rd Conference on Information Technology and Data Science (CITDS)*, 08 2024, pp. 1–6.

[14] M. Osváth, Z. Yang, and K. Kósa, "Analyzing narratives of patient experiences: A bert topic modeling approach," *Acta Polytechnica Hungarica*, vol. 20, pp. 153–171, 01 2023.

[15] Z. Yang and T. Váradi, "Training experimental language models with low resources, for the hungarian language," *Acta Polytechnica Hungarica*, vol. 20, pp. 169–188, 01 2023.

[16] N. Tóth, B. Oszkó, and Z. Yang, "Hungarian sentence analysis learning application with transformer models," *Acta Cybernetica*, vol. 27, pp. 83–91, 03 2025.

[17] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024. [Online]. Available: https://arxiv.org/abs/2402.03216

[18] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4623–4637. [Online]. Available: https://aclanthology.org/2020.acl-main.421/

[19] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 09 2019. [Online]. Available: https://doi.org/10.1162/tacl_a_00288

[20] T. Schuster, O. Ram, R. Barzilay, and A. Globerson, "Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1599–1613. [Online]. Available: https://aclanthology.org/N19-1162/

[21] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[22] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2318–2335. [Online]. Available: https://aclanthology.org/2024.findings-acl.137/

[23] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," arXiv preprint arXiv:2402.05672, 2024, model "multilingual-e5-base" released on Hugging Face, intfloat/multilingual-e5-base.

[24] J. Lee, F. Chen, S. Dua, D. Cer, M. Shanbhogue, I. Naim, G. H. Ábrego, Z. Li, K. Chen, H. S. Vera, X. Ren, S. Zhang, D. Salz, M. Boratko, J. Han, B. Chen, S. Huang, V. Rao, P. Suganthan, F. Han, A. Doumanoglou, N. Gupta, F. Moiseev, C. Yip, A. Jain, S. Baumgartner, S. Shahi, F. P. Gomez, S. Mariserla, M. Choi, P. Shah, S. Goenka, K. Chen, Y. Xia, K. Chen, S. M. K. Duddu, Y. Chen, T. Walker, W. Zhou, R. Ghiya, Z. Gleicher, K. Gill, Z. Dong, M. Seyedhosseini, Y. Sung, R. Hoffmann, and T. Duerig, "Gemini embedding: Generalizable embeddings from gemini," 2025. [Online]. Available: https://arxiv.org/abs/2503.07891

[25] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, "Nomic embed: Training a reproducible long context text embedder," 2024.

[26] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[27] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. [Online]. Available: https://arxiv.org/abs/1911.02116

[28] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, p. 333–389, Apr. 2009. [Online]. Available: https://doi.org/10.1561/1500000019

[29] N. Ligeti-Nagy, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, L. J. Laki, N. Vadász, Z. G. Yang, and T. Váradi, "Hulu: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából," in *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2022, pp. 431–446.

TABLE IV
RETRIEVAL ERROR RATES FOR THE CLEARSERVICE DATASET

Index	Question	Error Rate
23	*Milyen elvárás van a munkavégzéssel kapcsolatban?*	0.875
9	*Mi jár vasárnapi és ünnepnapi munkára?*	0.625
5	*Milyen szállodákban biztosít munkát a cég?*	0.250
15	*Kiutazás előtt mit kell teljesíteni?*	0.250
26	*Hogyan biztosítják a munkaegyenlőséget?*	0.250
39	*Milyen személyazonosító okmány szükséges?*	0.250
48	*Hol van a munkavállaló hivatalosan bejelentve?*	0.250
2	*Hány magyar munkavállaló dolgozik jelenleg a cégben?*	0.125
10	*A magyar alapbér előleg?*	0.125
12	*Milyen típusú lakásokban szállásolják el a dolgozókat?*	0.125
17	*Ki biztosítja a nyelvoktatást?*	0.125
19	*Milyen pozíciók érhetők el?*	0.125
24	*Milyen egészségügyi állapot kizáró ok?*	0.125
25	*Ki állapítja meg az egészségügyi alkalmasságot?*	0.125
28	*Mennyi ideig tart a tréning?*	0.125
30	*Ki fizeti a tréninget?*	0.125
35	*Van lehetőség hétvégén hazautazni?*	0.125
38	*Milyen munkaviszony szükséges az elmúlt egy évben?*	0.125
40	*Milyen erkölcsi feltétel van?*	0.125
41	*Hogyan kell felmondani a meglévő munkahelyen?*	0.125

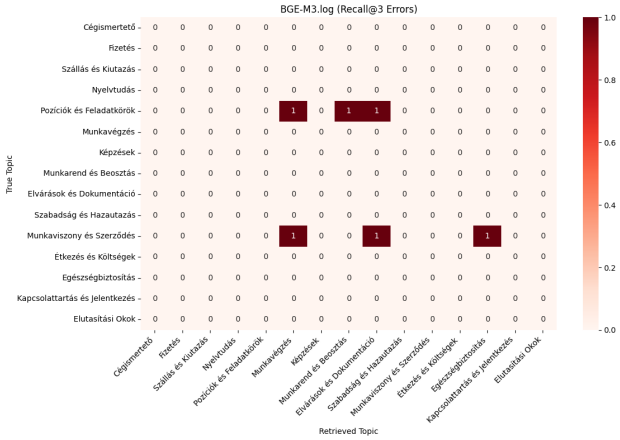

Fig. 7. Confusion matrix for the BGE-M3 model

Fig. 8. Confusion matrix for the E5-BASE model

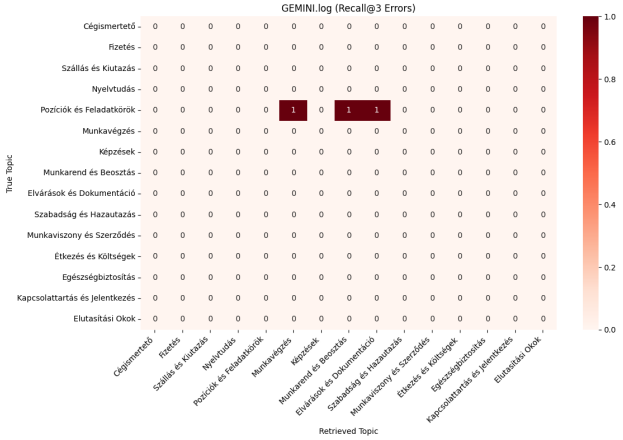

Fig. 9. Confusion matrix for the GEMINI model

APPENDIX

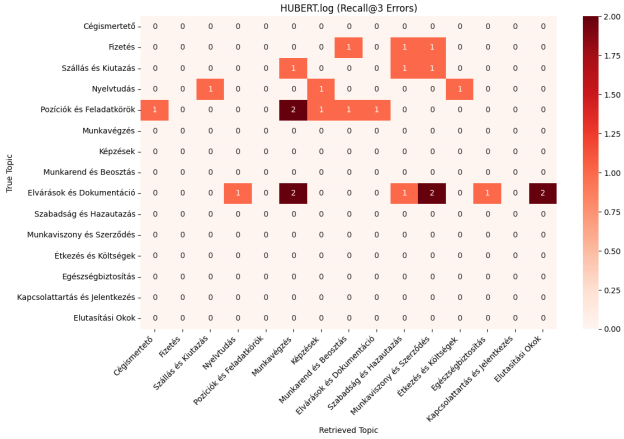

Fig. 10. Confusion matrix for the HUBERT model

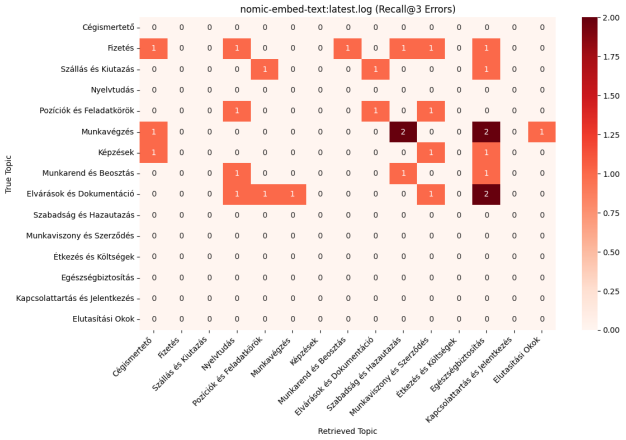

Fig. 11. Confusion matrix for the NOMIC model

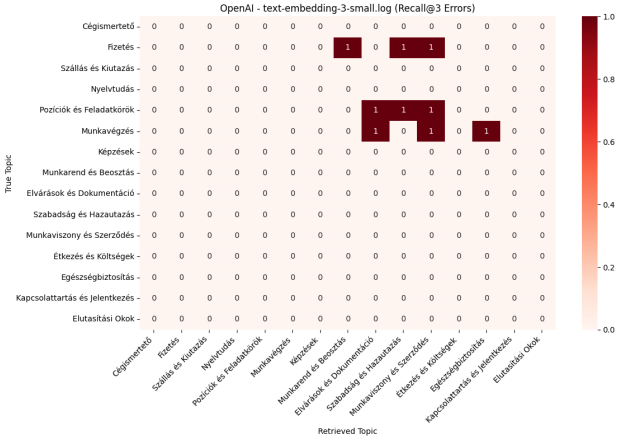

Fig. 12. Confusion matrix for the OPENAI-3SMALL model

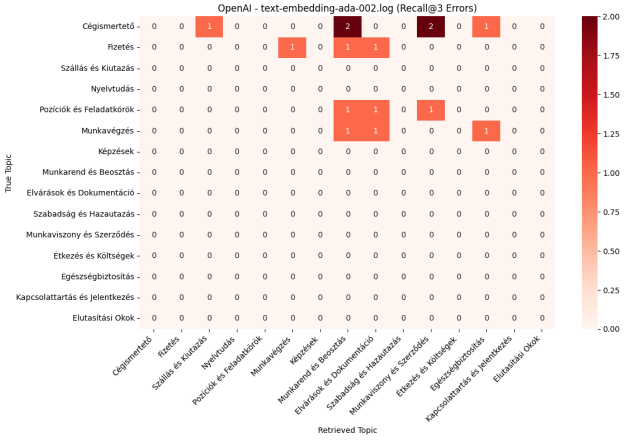

Fig. 13. Confusion matrix for the OPENAI-ADA model

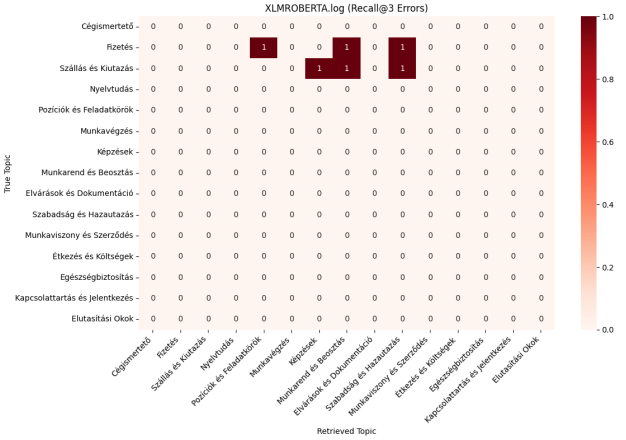

Fig. 14. Confusion matrix for the XLMROBERTA model