SkMTEB: Slovak Massive Text Embedding Benchmark and Model Adaptation

Marek Šuppa[α, β*] **Andrej Ridzik**[δ] **Daniel Hládek**[γ]
Natália Kňažeková[α, δ] **Viktória Ondrejová**[β]
[α]Comenius University in Bratislava, Slovakia, [β]Cisco Systems,
[γ]Technical University of Košice, Slovakia,
[δ]Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia

github.com/slovak-nlp/skmteb
huggingface.co/collections/slovak-nlp/skmteb

Abstract

We introduce SkMTEB, the first comprehensive MTEB-style text embedding benchmark for Slovak, a low-resource West Slavic language, comprising 31 datasets across 7 task types—nearly 4× the depth of existing multilingual benchmark coverage for Slovak. Our evaluation of 31 embedding models reveals that large instruction-tuned multilingual models achieve the strongest performance, while existing Slovak-specific models trained for NLU tasks transfer poorly to embedding tasks. To address the need for efficient, locally-deployable Slovak embeddings, we develop `e5-sk-small` (45M parameters) and `e5-sk-large` (365M) by applying vocabulary trimming and fine-tuning to Multilingual E5 models. Despite size reductions of up to 62%, our open-source models achieve competitive performance with proprietary APIs while remaining locally deployable for semantic search and retrieval-augmented generation (RAG). We release the benchmark, models, datasets, and code openly, hoping our approach offers a replicable path for other under-resourced languages.

1 Introduction

Text embeddings are now core infrastructure for semantic search, retrieval-augmented generation (RAG), clustering, and classification. The field has pursued scale—with state-of-the-art models reaching billions of parameters—but benchmark evidence is concentrated in high-resource languages, and the most capable models remain impractical to deploy at low latency or on constrained hardware.

This tension is especially acute for under-resourced languages. Although large multilingual models technically support hundreds of languages, their capacity is predominantly allocated to high-resource languages like English and Chinese. For a language like Slovak – a West Slavic language with approximately 5 million speakers—this means suboptimal representation in model vocabularies, limited training data coverage, and ultimately degraded performance compared to well-resourced languages.

Two complementary developments are needed to address this gap. First, robust evaluation benchmarks are essential to measure progress and identify weaknesses in the Slovak embedding models. Although benchmarks such as MTEB (Muennighoff et al., 2023) have catalyzed embedding research for English, and language-specific benchmarks have emerged for Chinese (Xiao et al., 2023), Polish (Poświata et al., 2024) and other languages, Slovak lacks such evaluation infrastructure. The existing skLEP benchmark (Suppa et al., 2025) addresses natural language understanding but not the embedding tasks critical to retrieval and semantic similarity applications.

Second, efficient adaptation techniques are needed to create compact and performant models that can be used practically. Approaches such as vocabulary trimming (Ushio et al., 2023) and targeted fine-tuning on high-quality data can yield strong results without massive compute budgets. For under-resourced languages, the goal is not to match the largest models on general benchmarks, but to create practical, efficient models that serve the specific language well.

In this work, we address both needs for Slovak. We introduce **SkMTEB**, the first comprehensive Slovak text embedding benchmark, comprising 31 datasets across 7 task types. Beyond evaluation, we demonstrate that effective Slovak embedding models can be trained with relatively modest resources by fine-tuning existing models on curated Slovak data and applying vocabulary trimming to create compact, language-specific variants.

Our contributions can hence be summarized as follows:

* Correspondence: marek@suppa.sk

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45597–45628

- We introduce SkMTEB—the first Slovak Massive Text Embedding Benchmark, for which we collate 31 datasets across 7 diverse task types.
- To ensure the benchmark has sufficient breadth, despite the severe lack of datasets in Slovak, we adapt multiple existing datasets for new tasks while also introducing seven brand-new datasets.
- We use SkMTEB to evaluate 31 open-weight and proprietary embedding models spanning compact, mid-sized, and large multilingual systems.
- We adapt Multilingual E5 into compact Slovak embedding models with vocabulary trimming and targeted fine-tuning, then ablate trimming, fine-tuning, and prompt usage.
- We openly release all models, datasets, and code: models and datasets are available at `https://huggingface.co/collections/slovak-nlp/skmteb` and the code at `https://github.com/slovak-nlp/skmteb`.

2 Related Work

Text embeddings have become fundamental components in modern NLP, powering semantic search, retrieval-augmented generation (Lewis et al., 2020), clustering, and classification systems. The field has witnessed a consistent trend toward larger models: from early approaches averaging Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) word vectors, through BERT-based sentence encoders (Devlin et al., 2019; Reimers and Gurevych, 2019), to today's billion-parameter embedding models. Recent state-of-the-art models exemplify this scaling trend—Multilingual E5 (Wang et al., 2024) ranges from 118M to 560M parameters, BGE-M3 (Chen et al., 2024) contains 568M parameters, Jina Embeddings v3 (Sturua et al., 2024) reaches 570M parameters, and the latest Qwen3-Embedding models (Zhang et al., 2025) scale up to 8B parameters. Alongside this scaling trend, recent releases increasingly emphasise instruction tuning, multi-task training, and practical efficiency: Jina Embeddings v4 (Günther et al., 2025) extends the line toward unified multimodal multilingual retrieval, Nomic Embed v2-MoE (Nussbaum and Duderstadt, 2025) applies sparse Mixture-of-Experts to text embeddings, and compact open models such as Granite Embedding (Awasthy et al., 2025) and EmbeddingGemma (Vera et al., 2025) target deployment-constrained settings. While these large models achieve impressive benchmark performance, their computational requirements pose significant challenges for practical deployment, particularly for applications requiring low latency, edge deployment, or cost-effective scaling.

For under-resourced languages, this scale-efficiency tension is compounded by a structural inefficiency: multilingual models allocate substantial capacity to high-resource languages (Wu and Dredze, 2020), and their embedding matrices—often comprising 30%–40% of total parameters (Ushio et al., 2023)—are dominated by tokens that are irrelevant to any single target language. Several approaches have emerged to address this efficiency gap. Vocabulary Trimming (Ushio et al., 2023) removes tokens irrelevant to the target language, reducing model size by up to 66% while preserving performance, with recent work on Dutch (Banar et al., 2025) demonstrating its applicability to embedding models. Static embedding approximations like Static-Similarity (Sentence-Transformers, 2024) achieve 100–400$\times$ faster inference while retaining 86%–95% of performance, and distillation approaches (Reimers and Gurevych, 2020) enable creating smaller models from larger teachers—illustrating that practical deployment often requires trading some accuracy for substantial efficiency gains.

Evaluating these trade-offs requires robust benchmarks. The Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) established a standardized evaluation framework spanning classification, clustering, semantic textual similarity, retrieval, and pair classification tasks with 56 English datasets. While MTEB primarily targets English, it catalyzed the development of language-specific benchmarks providing comparable depth: C-MTEB (Xiao et al., 2023) for Chinese (35 datasets), PL-MTEB (Poświata et al., 2024) for Polish (28 datasets), the Scandinavian Embedding Benchmark (Enevoldsen et al., 2024) for Danish, Norwegian, and Swedish, FR-MTEB (Ciancone et al., 2024) for French, ruMTEB (Snegirev et al., 2025) for Russian (23 datasets), ArabicMTEB (Bhatia et al., 2025) for Arabic (94 datasets), and FaMTEB (Zinvandi et al., 2025) for Persian.

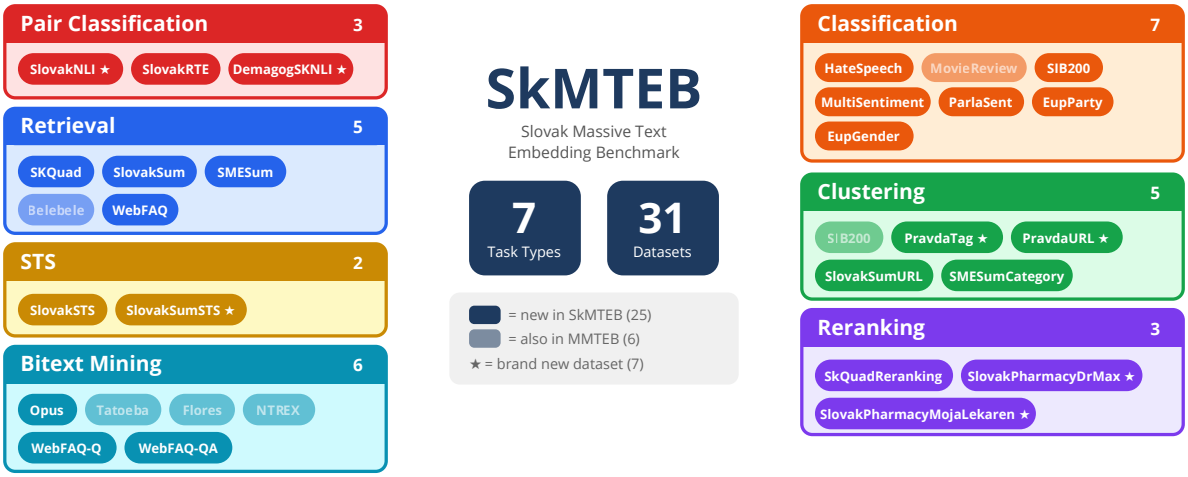

Figure 1: Overview of the SkMTEB benchmark comprising 31 datasets across 7 task types. Lighter shading indicates datasets also present in MMTEB (6), while solid shading marks datasets not in MMTEB (25). Datasets marked with ★ are 7 brand-new datasets created specifically for this work.

The recent MMTEB (Enevoldsen et al., 2025) takes a complementary approach, prioritizing breadth across 250+ languages over depth in any single language. This design necessarily yields shallow per-language coverage: Slovak is represented by only 8 tasks in MMTEB—just 14% of English MTEB's depth and 29% of PL-MTEB's coverage. These 8 tasks consist primarily of subsets from multilingual datasets (SIB-200, FLORES, Tatoeba) that, while enabling cross-lingual comparison, lack Slovak-specific evaluation scenarios such as native retrieval benchmarks, domain-specific tasks, or temporally grounded evaluation. SkMTEB addresses this gap with 31 datasets—nearly 4× MMTEB's Slovak coverage—spanning domains (medical, fact-checking, parliamentary), task formulations (summarization-as-retrieval, URL-based clustering), and temporal ranges (2000–2025). Only 6 datasets overlap, ensuring the benchmarks are complementary: MMTEB enables cross-lingual comparison, while SkMTEB provides the depth needed to diagnose Slovak-specific model behavior.

3 The SkMTEB Benchmark

SkMTEB comprises 31 datasets across 7 task types following the MTEB framework (Muennighoff et al., 2023), providing comprehensive coverage of Slovak text embedding evaluation. The benchmark spans diverse domains including news, government, social media, reviews, and encyclopedic content, with temporal coverage from 2000 to 2025. For each task type below, we first define the task and evaluation metrics, then describe the Slovak datasets we include. Full task and dataset metadata are provided in Appendix E, and curation details for newly created datasets are summarized in Appendices B and C. Figure 1 provides a compact overview of the benchmark composition.

3.1 Retrieval (5 datasets)

From a large corpus of documents and a set of query texts, the model must retrieve (rank) the most relevant documents per query by embedding similarity. The standard metrics include nDCG@k (primarily nDCG@10) (Wang et al., 2013), MRR@k, MAP@k, precision@k, and recall@k. One of the first to evaluate the retrieval of Slovak information was (Hládek et al., 2016).

SKQuadRetrieval is a question-answering retrieval task based on the SK-QuAD dataset (Hládek et al., 2023), evaluating search performance with relevance-scored answers from encyclopedic content. **SlovakSumRetrieval** (Ondrejova and Suppa, 2024) and **SMESumRetrieval** (Suppa and Adamec, 2020) reformulate news summarization datasets as retrieval tasks, using article abstracts as queries to retrieve full documents from collections of 200k+ and 80k articles, respectively. **BelebeleRetrieval** (Bandarkar et al., 2024) provides machine reading comprehension across 122 language variants, while **WebFAQRetrieval** (Dinzinger et al., 2025) contains question-answer pairs from web FAQ pages across 75 languages.

3.2 Reranking (3 datasets)

Given a query and a list of candidate reference texts (some relevant, some irrelevant), the model must produce a ranking of the candidates according to relevance to the query (via embedding similarity). The ranked lists are evaluated using ranking metrics: Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), with MAP as the principal metric.

SkQuadReranking derives from SK-QuAD (Hládek et al., 2023) for article retrieval reranking with manually annotated hard-negatives. **SlovakPharmacyDrMaxReranking** and **SlovakPharmacyMojaLekarenReranking** are reranking datasets built from pharmacist Q&A content; curation details are provided in Appendix B.

3.3 Classification (7 datasets)

Texts with labels are split into train and test sets; embeddings are computed, and then a simple classifier (e.g., logistic regression) is trained on the training embeddings and evaluated on the test embeddings. The main evaluation metric is classification accuracy (with optional F1 and average precision).

SlovakHateSpeechClassification.v2 (Sokolová et al., 2025) annotates social media posts for hateful or offensive language. **SlovakMovieReviewSentimentClassification.v2** (Štefánik et al., 2023) provides binary sentiment classification from 30k+ Slovak movie reviews (2002–2020). **SIB200Classification** (Adelani et al., 2023) is the largest public topic classification dataset, covering 205 languages with 7 topics. **MultilingualSentimentClassification** (Mollanorozy et al., 2023) spans 30 languages with binary sentiment labels. **SlovakParlaSentClassification** (Mochtak et al., 2024) contains 3-level sentiment annotations from parliamentary debates. **MultiEupSlovakPartyClassification** and **MultiEupSlovakGenderClassification** (Yang et al., 2024) predict political groups and gender from European Parliament speeches (2020–2024).

3.4 Clustering (5 datasets)

A collection of texts (sentences, paragraphs) is embedded and then clustered into groups. Clustering quality is assessed with label-agnostic metrics such as V-measure, which is insensitive to label permutations.

SIB200ClusteringS2S (Adelani et al., 2023) clusters up to 1,004 documents into 7 thematic topics. **PravdaSKTagClustering** and **PravdaSKURLClustering** cluster Pravda.sk news articles by tags and URL structure into 50 categories. **SlovakSumURLClustering** (Ondrejova and Suppa, 2024) and **SMESumCategoryClustering** (Suppa and Adamec, 2020) organize news articles into 12 and 11 editorial categories, respectively.

3.5 Bitext Mining (6 datasets)

Given two sets of sentences in different languages, the model must find for each sentence in the first set its best matching translation in the second set using embedding similarity. The performance is primarily measured by F1 (and also precision and recall).

OpusSlovakEnglishBitextMining (Zhang et al., 2020) provides Slovak-English parallel sentences from OPUS-100 (2000–2020). **TatoebaBitextMining** (community, 2021) and **FloresBitextMining** (Goyal et al., 2022) offer multilingual parallel corpora with Slovak support. **NTREXBitextMining** (Federmann et al., 2022) provides Slovak translations of a multilingual news corpus. **WebFAQBitextMiningQuestions** and **WebFAQBitextMiningQAs** (Dinzinger et al., 2025) enable cross-lingual question and Q&A pair retrieval across 75 languages. To keep evaluation focused, we restrict each task to Slovak-English and, where available, Slovak-Czech pairs.

3.6 Pair Classification (3 datasets)

The task is to take a pair of texts and decide whether they are equivalent (e.g., paraphrase, duplicate) or not. The model embeds both texts and computes a distance or similarity; the thresholding then yields binary predictions, and metrics such as average precision (cosine) are used.

SlovakNLI contains handwritten premise-hypothesis pairs for natural language inference (entailment vs. contradiction). **SlovakRTE** (Suppa et al., 2025) is a professionally translated and human-verified recognizing textual entailment dataset from the skLEP benchmark. **DemagogSKNLI** creates NLI pairs from Demagog.sk fact-checking data (2010–2025), pairing evidence with political statements for claim verification.

3.7 Semantic Textual Similarity (2 datasets)

For a pair of sentences, the objective is to predict a continuous similarity score. The embedding sim-

ilarity (e.g., cosine) is correlated with the human-annotated similarity scores, with Spearman correlation being the main evaluation metric.

SlovakSTS (Suppa et al., 2025) is a Slovak translation of the GLUE STSb dataset, providing human-verified semantic similarity scores (0–5) for sentence pairs from blogs and news. **SlovakSumSTS** is a synthetic dataset derived from the SlovakSum news corpus (Ondrejova and Suppa, 2024), with LLM-generated similarity scores that have been human-verified; curation details are provided in Appendix C.

3.8 Baseline Models

We evaluate a diverse set of embedding models to establish baselines on SkMTEB. Our selection spans several axes: (i) historical anchors that established multilingual embedding benchmarks, (ii) models with explicit Slovak-language support, (iii) coverage of the size–quality trade-off from lightweight to large-scale, (iv) architectural diversity (dense, sparse, MoE, instruction-tuned), and (v) leading proprietary APIs as upper-bound references. We organize the baselines by model family and briefly motivate each group in historical order.

Sentence-Transformers and early multilingual baselines. `LaBSE` arrived as one of the first truly strong language-agnostic sentence encoders and quickly became a default reference for multilingual retrieval and bitext mining, making it an essential historical anchor for SkMTEB. (Feng et al., 2022) We also include the widely used `paraphrase-multilingual-mpnet-base-v2` and the compact `paraphrase-multilingual-MiniLM-L12-v2`, which set practical baselines for multilingual similarity and remain common production choices due to their balance of quality and efficiency. (Reimers and Gurevych, 2019) The more recent `static-similarity-mrl-multilingual-v1` represents a different trade-off: an explicitly speed-oriented model that prioritized ultra-fast CPU inference at the time of release and thus captures the efficiency end of the multilingual spectrum. (Sentence-Transformers, 2024)

Slovak-specific baselines. `slovakbert-sts-stsb` adapts SlovakBERT for sentence similarity, providing a localized baseline that reflects Slovak linguistic idiosyncrasies rather than cross-lingual transfer alone. (Pikuliak et al., 2022) `slovakbert-skquad-mnlr` extends this idea to retrieval-style supervision, grounding a Slovak-specific model in QA-derived ranking signals and offering a native reference point for retrieval tasks at the time Slovak resources were scarce. (Pikuliak et al., 2022; Hládek et al., 2023)

Multilingual E5 family. The `multilingual-e5-small`, `multilingual-e5-base`, and `multilingual-e5-large` models established a strong, consistently competitive multilingual baseline that was easy to use and broadly effective across task types, making E5 the default yardstick for many benchmarks when it appeared. (Wang et al., 2024) The later `multilingual-e5-large-instruct` variant brought instruction tuning into this family, reflecting the field's shift toward promptable embeddings and improved transfer across heterogeneous tasks. (Wang et al., 2024)

Modern multilingual retrieval families. `bge-m3` was notable for unifying dense, sparse, and multi-vector retrieval in a single model, signaling a move toward multi-functionality rather than single-purpose encoders. (Chen et al., 2024) `gte-multilingual-base` emphasized long-context retrieval and elastic embeddings, capturing the growing demand for longer documents and reranking-style pipelines at the time of release. (Zhang et al., 2024) `snowflake-arctic-embed-l-v2.0` represents Snowflake's second-generation large embedding model, bringing improved multilingual retrieval performance and long-context support through a refined training pipeline that targets practical enterprise retrieval workloads. (Yu et al., 2024)

Nomic and Jina model lines. `nomic-embed-text-v1.5` popularized long-context embedding with Matryoshka representations, letting practitioners trade embedding size for speed without retraining and making it a frequent baseline in applied settings. (Nussbaum et al., 2025) `nomic-embed-text-v2-moe` advanced this line with sparse Mixture-of-Experts, offering a higher quality-efficiency frontier that reflected broader trends in scalable embedding training. (Nussbaum and Duderstadt, 2025) On the Jina side, `jina-embeddings-v3` introduced task-specific LoRA adapters and Matryoshka learning, while `jina-embeddings-v4` expanded to multimodal multilingual retrieval; together they

capture a progression from flexible text embeddings to unified text-image representations that became increasingly important in the field. (Sturua et al., 2024; Günther et al., 2025)

Recent large-scale embedding families. The `granite-embedding-107m-multilingual` and `granite-embedding-278m-multilingual` models represent IBM's modern embedding line, providing compact and mid-sized baselines with contemporary training recipes that were competitive at release time. (Awasthy et al., 2025) `Qwen3-Embedding-0.6B` marks the entry of a strong foundation-model lineage into embeddings, offering an efficient Qwen-family baseline. (Zhang et al., 2025) Finally, `embeddinggemma-300m` reflects the push toward lightweight, high-quality open embeddings, making it a timely reference point for practical deployments with limited compute. (Vera et al., 2025)

Proprietary API models. To complement open-weight baselines, we include leading proprietary embedding services. OpenAI's `text-embedding-3-small` and `text-embedding-3-large`[1] represent the current generation of their embedding API, with the large variant offering higher capacity at increased cost. Cohere's `embed-v4.0`[2] provides another commercial reference point with strong multilingual capabilities. Amazon's `titan-embed-text-v2`[3] completes the proprietary baselines as AWS's current-generation embedding API with strong multilingual capabilities. These API models establish upper bounds for what commercial solutions offer on Slovak text, though their closed nature limits reproducibility and architectural analysis.

All models are evaluated using their default configurations and recommended preprocessing steps to ensure fair comparison across different architectures and training approaches.

4 Adapting Embedding Models for Slovak

Beyond benchmarking, we explore training Slovak-specific embedding models by fine-tuning Slovak-BERT (Pikuliak et al., 2022) and models from the Multilingual E5 family (Wang et al., 2024).

Training Data We use datasets from the skLEP benchmark (Suppa et al., 2025): SK-SQuAD (Hládek et al., 2023) (72K query-context pairs), NLI translated from XNLI (Conneau et al., 2018) (393K pairs), STS from GLUE STS-B (Cer et al., 2017) (6K pairs), and RTE from GLUE (Wang et al., 2018) (2.5K pairs). We also experiment with Slovak Web QA (967K pairs from WebFAQ (Dinzinger et al., 2025) and MFAQ (De Bruyn et al., 2021)), where hard negatives are randomly sampled answers from the same domain; we later exclude this dataset as this construction does not consistently provide meaningful contrastive signal.

Training Configuration We use mean pooling with a max sequence length of 256, multi-task learning with Cosine Similarity Loss for STS and Multiple Negatives Ranking Loss (Henderson et al., 2017) for other tasks. Training uses a batch size of 32, a learning rate 2×10^{-5} with a linear warmup (10% of training steps), and 3 epochs. All experiments use a single NVIDIA H100 GPU with random seed 42. Training completes in under 1 hour per model variant. Full hyperparameters are provided in Appendix F.

Vocabulary Trimming We apply Vocabulary Trimming (VT) (Ushio et al., 2023) to create compact Slovak-specific models from multilingual E5 encoders, following recent work on Dutch embeddings (Banar et al., 2025). VT removes vocabulary tokens irrelevant to the target language; we retain 60K tokens (from 250K) following the recommendation of Ushio et al. (2023), who found this threshold balances vocabulary coverage with model efficiency across multiple target languages. Token retention is determined by frequency in FineWeb2-Slovak,[4] a quality-filtered Slovak web corpus (Penedo et al., 2024). We apply VT before fine-tuning (Pre-FT VT), reducing both model size and training time. This yields size reductions of 62% for E5-small (118M → 45M) and 35% for E5-large (560M → 365M).

4.1 Fine-tuned Models

We explore two approaches to training Slovak embedding models.

[1] https://platform.openai.com/docs/guides/embeddings

[2] https://docs.cohere.com/docs/embed

[3] https://docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html

[4] https://huggingface.co/datasets/ivykopal/fineweb2-slovak

Model (↓)	Average Across		Average per Task Type						
	All	Type	Btxt	Clf	Clust	PrClf	Rrnk	Rtrvl	STS
Number of datasets (→)	(31)	(7)	(6)	(7)	(5)	(3)	(3)	(5)	(2)
Small models (<130M)									
`e5-sk-small` (45M)	70.56	72.01	91.34	60.84	40.95	66.05	84.94	78.64	81.32
`granite-embedding-107m-multilingual` (107M)	65.81	67.31	85.85	56.04	40.80	60.88	78.10	70.93	78.54
`static-similarity-mrl-multilingual-v1` (108M)	58.51	60.90	87.98	48.87	17.47	63.66	70.76	59.35	78.22
`multilingual-e5-small` (118M)	70.32	71.78	91.29	59.81	41.09	64.66	85.13	79.21	81.25
`paraphrase-multilingual-MiniLM-L12-v2` (118M)	67.86	69.16	94.91	60.83	37.44	67.82	73.46	66.03	83.62
`slovakbert-skquad-mnlr` (125M)	67.49	68.91	85.07	62.49	38.07	65.13	76.91	72.82	81.90
`slovakbert-sts-stsb` (125M)	63.44	65.37	81.75	63.05	31.82	68.90	67.15	59.20	85.69
`sturovec-base` (125M)	68.99	70.13	86.97	62.19	41.55	63.27	80.66	76.54	79.73
Base models (>=130M, <350M)									
`nomic-embed-text-v1.5` (137M)	51.52	55.03	46.64	51.40	28.77	61.98	68.60	55.39	72.43
`granite-embedding-278m-multilingual` (278M)	67.69	69.09	87.56	57.81	42.41	61.40	81.25	73.92	79.30
`multilingual-e5-base` (278M)	72.39	73.57	94.76	62.35	40.91	64.10	85.71	83.57	83.59
`paraphrase-multilingual-mpnet-base-v2` (278M)	70.34	71.67	96.23	64.05	38.95	68.21	76.79	70.03	87.44
`gte-multilingual-base` (305M)	71.76	73.26	94.36	61.70	39.28	67.24	83.04	81.63	85.57
`embeddinggemma-300m` (308M)	69.25	70.56	83.89	62.08	43.76	65.54	81.00	78.40	79.28
`nomic-embed-text-v2-moe` (330M)	72.58	73.84	90.03	63.98	43.53	64.19	85.15	85.37	84.60
Large models (>=350M)									
`e5-sk-large` (365M)	74.70	75.88	96.39	66.34	41.43	67.32	87.81	85.60	86.25
`LaBSE` (471M)	66.44	67.52	**97.48**	58.73	34.28	64.39	73.52	63.39	80.85
`multilingual-e5-large` (560M)	74.25	75.49	96.29	65.34	40.35	66.78	87.96	85.80	85.90
`multilingual-e5-large-instruct` (560M)	**77.49**	**78.44**	97.09	70.28	**49.69**	**70.55**	86.49	86.08	88.86
`bge-m3` (568M)	74.43	75.55	96.29	66.81	39.97	67.03	86.72	85.63	86.39
`snowflake-arctic-embed-l-v2.0` (568M)	72.54	73.63	93.46	63.40	40.41	63.11	87.11	85.17	82.76
`jina-embeddings-v3` (572M)	75.10	76.20	96.43	66.89	44.63	66.44	83.44	85.77	**89.82**
`Qwen3-Embedding-0.6B` (596M)	70.53	71.80	90.12	62.84	44.28	64.46	82.02	75.73	83.14
`jina-embeddings-v4` (3.8B)	72.44	73.87	90.49	62.99	44.65	64.43	85.25	83.64	85.65
`Qwen3-Embedding-4B` (4B)	73.70	74.96	94.21	64.26	44.22	65.82	86.29	84.11	85.82
`Qwen3-Embedding-8B` (8B)	74.53	75.75	94.43	65.94	43.36	66.65	87.04	86.27	86.54
API access models									
`embed-v4.0`	71.26	72.82	90.23	60.12	40.13	63.12	87.45	85.71	82.95
`text-embedding-3-small`	70.48	71.39	91.82	62.13	43.64	63.24	82.37	76.96	79.56
`text-embedding-3-large`	75.07	75.89	96.79	66.91	44.22	66.58	86.96	85.55	84.21
`gemini-embedding-001`	77.23	78.07	96.76	**71.01**	46.26	67.08	**88.27**	**88.34**	88.74
`amazon-titan-embed-text-v2`	67.24	69.21	84.98	59.26	33.39	62.78	83.99	77.43	82.67

Table 1: SkMTEB results summary (percent). The table reports the average performance across all tasks (**All**) and the unweighted average across task types (**Type**), followed by per-task-type averages for Bitext Mining (Btxt), Classification (Clf), Clustering (Clust), Pair Classification (PrClf), Reranking (Rrnk), Retrieval (Rtrvl), and STS. The best result per task is **bolded** with the runner-up underlined.

SlovakBERT with Full Training Data Our initial approach fine-tunes SlovakBERT (Pikuliak et al., 2022) on the complete training dataset, including the Slovak Web QA triplets. The resulting model, `sturovec-base`, achieves reasonable performance but falls short of the unfine-tuned `multilingual-e5-small` baseline (68.99 vs. 70.32 average score), despite using 1.4M training examples. Upon analyzing the Slovak Web QA triplets, we identified quality issues stemming from the automated hard negative sampling process, where random answers from the same domain do not consistently provide meaningful contrastive signal. This motivated a refined approach for subsequent experiments.

Vocabulary-Trimmed E5 Models Based on these observations, we fine-tune models from the Multilingual E5 family (Wang et al., 2024) *without* the Slovak Web QA triplets, using only the higher-quality skLEP datasets (SK-SQuAD, NLI, STS, RTE). Prior to fine-tuning, we apply vocabulary trimming to reduce the models to 60K tokens based on FineWeb2-Slovak frequency statistics. This substantially reduces model size: `e5-sk-small` contains 45M parameters compared to 118M in the original `multilingual-e5-small`—a reduction of over 60%. For the large variant, `e5-sk-large` is reduced from 560M to 365M parameters.

Both models are fine-tuned for 3 epochs using the same hyperparameters as the SlovakBERT experiments. Following standard E5 practice, we prepend `query:` and `passage:` prefixes to queries

and documents, respectively, during both training and inference.

5 SkMTEB Results

Our main results are summarized in Table 1, which reports average scores by task type, and in Table 9, which details per-task classification performance. Figure 3 and Figure 2 visualize the relationship between model size and average performance.

Among evaluated models, `multilingual-e5-large-instruct` achieves the highest overall score (77.49), followed closely by `gemini-embedding-001` (77.23). These instruction-tuned and large-scale models excel particularly in classification and clustering tasks, where they outperform smaller alternatives by significant margins. The Multilingual E5 family demonstrates strong performance across model sizes, with even the small variant (118M) achieving competitive results. Very large models do not consistently outperform 500M–600M alternatives on SkMTEB: `jina-embeddings-v4` (3.8B, 72.44) trails `snowflake-arctic-embed-l-v2.0` (568M, 72.54) and `nomic-embed-text-v2-moe` (330M, 72.58), and only narrowly edges `multilingual-e5-base` (278M, 72.39). This suggests diminishing returns from scale alone for Slovak embedding tasks.

Task difficulty varies substantially across the benchmark. Bitext mining proves largely solved, with most models reaching F1 above 90—reflecting the relative ease of cross-lingual alignment for Slovak-English and Slovak-Czech pairs. In contrast, clustering remains challenging, with V-measure ranging from 17 to 50, indicating significant room for improvement. Reranking and retrieval tasks show strong performance from E5-family models, while STS benefits from models with explicit similarity training objectives like `jina-embeddings-v3` (89.82). Existing Slovak-specific models trained primarily for NLU tasks (`slovakbert-skquad-mnlr`, `slovakbert-sts-stsb`) underperform compared to multilingual alternatives, highlighting the need for dedicated embedding model development.

Our vocabulary-trimmed E5 models demonstrate competitive performance, matching proprietary API models: `e5-sk-small` (70.56) performs on par with `text-embedding-3-small` (70.48), while `e5-sk-large` (74.70) achieves comparable results to `text-embedding-3-large` (75.07).

Model Variant	VT	FT	Size	Avg	Δ
`mE5-small` (baseline)			118M	70.32	—
+ VT	✓		45M	70.45	+0.13
+ FT		✓	118M	70.58	+0.26
+ VT + FT	✓	✓	45M	70.56	+0.24
+ VT + FT + prompt	✓	✓	45M	**71.07**	+0.75
`mE5-large` (baseline)			560M	74.25	—
+ VT	✓		365M	74.56	+0.31
+ FT		✓	560M	74.46	+0.21
+ VT + FT	✓	✓	365M	74.70	+0.45
+ VT + FT + prompt	✓	✓	365M	**74.72**	+0.47

Table 2: Ablation study on vocabulary trimming (VT), fine-tuning (FT), and prompt usage. The Δ column shows change relative to the baseline. Size reduction from VT: 62% for small, 35% for large.

TOST equivalence testing (Lakens, 2017) confirms practical equivalence: 90% CIs for both comparisons fall within ±2 points. Crucially, our models offer practical advantages: they are open-weight, can run locally without API costs, and their reduced size enables higher throughput—making them accessible for practitioners working with Slovak text.

6 Ablation Study

We conduct ablation experiments to isolate the contributions of vocabulary trimming (VT), fine-tuning (FT), and inference prompts to our final models. Table 2 summarizes results for the E5-small and E5-large model families.

Vocabulary Trimming VT reduces model size dramatically—from 118M to 45M parameters for E5-small (62% reduction) and from 560M to 365M for E5-large (35% reduction)—while preserving or slightly improving performance (+0.13 for small, +0.31 for large).

Fine-tuning Fine-tuning on skLEP data (excluding the noisy Slovak Web QA triplets) provides modest but consistent improvements: +0.26 for E5-small and +0.21 for E5-large.

Combined Effects Combining VT and FT yields additive benefits for the large model (+0.45 total improvement over baseline while reducing size by 35%). For the small model, the combined approach achieves nearly the same performance as FT alone but with 62% fewer parameters.

Inference Prompts Adding `query:`/`passage:` prefixes during both training and inference provides additional improvements: +0.51 for E5-small (70.56 → 71.07) and +0.02 for E5-large (74.70 → 74.72). The larger effect on the small model

	E5-small		E5-large	
Task (pair)	**Orig.**	**Trim.**	**Orig.**	**Trim.**
Flores (ces-slk)	99.87	99.74	99.87	99.87
Flores (eng-slk)	95.58	95.72	100.00	100.00
NTREX (ces-slk)	98.08	98.11	99.22	99.27
NTREX (eng-slk)	94.64	95.56	98.80	98.66
Tatoeba (slk-eng)	82.69	82.56	93.22	93.22
Opus (slk-eng)	77.08	76.95	86.21	86.27

Table 3: Cross-lingual bitext mining F1 for original vs. vocabulary-trimmed Multilingual E5 models. Differences stay within 1 F1 point on all pairs, averaging 0.25 F1 for the small model and 0.04 F1 for the large model.

suggests that explicit query-passage distinction particularly benefits models with limited capacity.

Cross-lingual Transfer Preservation A natural concern with aggressive vocabulary trimming is that removing non-Slovak tokens could degrade cross-lingual transfer. Table 3 reports per-pair F1 on the six cross-lingual bitext mining tasks in SkMTEB for the original and trimmed E5 models. Differences are small in both directions (maximum absolute change 0.92 F1 for the small model and 0.14 F1 for the large model), and several pairs marginally improve after trimming. Targeted vocabulary reduction thus preserves Slovak-English and Slovak-Czech transfer, addressing a practical concern for deployments that process code-mixed or adjacent-language content.

7 Conclusion

We presented SkMTEB, the first comprehensive text embedding benchmark for Slovak, comprising 31 datasets across 7 task types. Our evaluation reveals that large instruction-tuned multilingual models achieve strong cross-lingual transfer to Slovak, while compact alternatives like multilingual-e5-small offer competitive performance suitable for practical deployment. Existing Slovak-specific NLU models do not transfer well to embedding tasks, highlighting the need for dedicated development.

We also show that combining vocabulary trimming with fine-tuning on curated Slovak data yields compact, competitive embedding models with modest computational resources. Our 45M parameter model achieves 91% of the performance of the best 560M parameter multilingual model, demonstrating that vocabulary trimming combined with targeted fine-tuning offers a practical path to efficient, language-specific embeddings. Ablations show that trimming reduces E5-small by 62% and E5-large by 35% while preserving Slovak-English and Slovak-Czech bitext mining within 1 F1 point. We release SkMTEB, our vocabulary-trimmed models, and all associated code under open-source licenses.

Beyond Slovak, four findings should transfer to other under-resourced languages: (1) NLU-tuned monolingual encoders underperform on embedding tasks, making language-specific embedding evaluation necessary even where NLU benchmarks exist; (2) vocabulary trimming generalizes from NLU to embedding models, yielding 35–62% parameter reductions with negligible in-language or cross-lingual degradation; (3) 4B–8B-parameter embedding models rarely outperform their 500M–600M counterparts on a single target language, suggesting embedding-specific scaling trends distinct from those observed for generative LLMs; and (4) the full SkMTEB pipeline—benchmark construction, broad evaluation, trimming, and fine-tuning—can be replicated with modest compute (under one GPU-hour per adapted model), providing a practical template for teams working on similar languages.

Limitations

Benchmark coverage. While SkMTEB substantially expands Slovak embedding evaluation from 8 to 31 tasks, gaps remain. Several datasets rely on machine translation from English sources with native speaker post-editing, which may not fully capture Slovak-specific linguistic phenomena or cultural contexts. Domain coverage skews toward news, Wikipedia, and web content; specialized domains such as legal, medical, or technical Slovak are underrepresented.

Translated vs. native data. A portion of our benchmark derives from translated datasets (NLI, STS, RTE). While we employed machine translation with native speaker post-editing, translated data may exhibit translationese artifacts and fail to reflect authentic Slovak language use. Future work should prioritize natively authored Slovak datasets.

Model selection. Our baseline evaluation, while extensive, cannot cover all available embedding models. We focused on models with documented multilingual support, including both open-weight models and select proprietary APIs; however, very recent releases may not be represented. Addition-

ally, computational constraints limited evaluation of the largest models (8B+ parameters) across all tasks.

Temporal snapshot. Both the benchmark and model evaluations represent a snapshot in time. The embedding model landscape evolves rapidly, and newer models may substantially outperform those evaluated here. We encourage ongoing community contributions to keep evaluations current.

Vocabulary trimming trade-offs. While vocabulary trimming reduces model size, it may affect performance on code-mixed or multilingual Slovak text. Our per-pair analysis (Table 3) shows that Slovak-English and Slovak-Czech transfer is preserved within 1 F1 point on all tested pairs, but we do not evaluate performance on non-Slavic, non-English pairs or on sentences that interleave multiple languages.

Ethics Statement

Data licensing. All datasets included in SkMTEB are released under permissive licenses or with explicit permission for research use. We document licensing information for each dataset in Appendix E and release our benchmark under a license compatible with downstream research and development.

Privacy and offensive content. The datasets in SkMTEB are derived from publicly available sources (Wikipedia, news articles, web content, parliamentary proceedings) or existing research datasets. We did not collect new data from human subjects. For newly curated datasets (pharmacy Q&A, news clustering), we verified that source data does not contain personally identifying information beyond public figures mentioned in news contexts. The hate speech classification dataset contains offensive language necessary for the task; we include content warnings in dataset documentation and recommend appropriate handling during model development. Parliamentary and news datasets may contain politically sensitive content reflecting public discourse.

Potential biases. Embedding models and the datasets used to train and evaluate them may encode societal biases present in their source data. SkMTEB inherits biases from its constituent datasets, and strong benchmark performance does not guarantee fair or unbiased model behavior. We encourage users to conduct bias audits appropriate to their deployment contexts.

Intended use. SkMTEB is intended for research evaluation of text embedding models for Slovak. While we hope it enables the development of better Slovak NLP systems—for example in public-sector search, accessibility, and moderation over Slovak text—users should validate model performance on their specific use cases rather than relying solely on benchmark scores. Retrieval and clustering over Slovak political discourse could also enable surveillance or political profiling; we recommend context-appropriate governance when deploying models trained on or evaluated with politically salient datasets such as `DemagogSKNLI` and `SlovakParlaSentClassification`.

Environmental impact. Running the full SkMTEB evaluation suite requires substantial computation. We report approximate compute requirements in the appendix to enable carbon footprint estimation. Our work on vocabulary trimming and efficient adaptation aims to reduce the computational burden of deploying embedding models.

Use of AI assistants. In addition to GPT-5 for dataset generation (Appendix C), we used AI assistants (Claude, ChatGPT) to aid with figure visualization, code refactoring, and English grammar and vocabulary checking, as the authors are non-native English speakers. All AI-generated content was reviewed and verified by the authors.

Acknowledgments

This study was funded by the Ministry of Education, Research, Development and Youth of the Slovak Republic under the project KEGA 049TUKE-4/2024, VEGA 1/0685/26 and by the Slovak Research and Development Agency under the project APVV-22-0414.

This work was partially funded by European Union, under the project lorAI - Low Resource Artificial Intelligence, GA No. 101136646, https://doi.org/10.3030/101136646. It was also partially funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I02-03-V01-00029.

Part of the research results was obtained using the computational resources procured in the national project National competence centre for high performance computing (project code:

311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of society, Operational Program Integrated Infrastructure.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. Granite embedding models. *ArXiv*, abs/2502.20204.

Nikolay Banar, Ehsan Lotfi, Jens Van Nooten, Cristina Arhiliuc, Marija Kliocaite, and Walter Daelemans. 2025. MTEB-NL and E5-NL: Embedding benchmark and models for dutch. *arXiv preprint arXiv:2509.12340*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775. Association for Computational Linguistics.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Extending the massive text embedding benchmark to French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy. ELRA and ICCL.

Tatoeba community. 2021. Tatoeba: Collection of sentences and translations.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. Association for Computational Linguistics.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186. Association for Computational Linguistics.

Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. 2025. Webfaq: A multilingual collection of natural q&a datasets for dense retrieval. *Preprint*, arXiv:2502.20936.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems*, volume 37.

Kenneth C. Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, and 1 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First*

Workshop on Scaling Up Multilingual Evaluation, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, and Francisco Guzmán. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 19–35.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

Daniel Hládek, Jan Staš, and Jozef Juhár. 2016. Evaluation set for Slovak news information retrieval. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1913–1916, Portorož, Slovenia. European Language Resources Association (ELRA).

Daniel Hládek, Jan Staš, Jozef Juhár, and Tomáš Koctúr. 2023. Slovak dataset for multilingual question answering. *IEEE Access*, 11:32869–32881.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16024–16036, Torino, Italy. ELRA and ICCL.

Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with Persian. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 89–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2006–2029. Association for Computational Linguistics.

Zach Nussbaum and Brandon Duderstadt. 2025. Training sparse mixture of experts text embedding models. *ArXiv*, abs/2502.07972.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*.

Viktoria Ondrejova and Marek Suppa. 2024. SlovakSum: A large scale Slovak summarization dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14916–14922, Torino, Italia. ELRA and ICCL.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. FineWeb2: A sparkling update with 1000s of languages.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2022. SlovakBERT: Slovak masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7156–7168. Association for Computational Linguistics.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. PL-MTEB: Polish massive text embedding benchmark. *ArXiv*, abs/2405.10138.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525. Association for Computational Linguistics.

Sentence-Transformers. 2024. static-similarity-mrl-multilingual-v1. https://huggingface.co/sentence-transformers/static-similarity-mrl-multilingual-v1/commit/e60353de452a9a9c5616ecf6a1da9b65bff54d18. Static multilingual similarity embedding model.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Aleksandr Abramov. 2025. The russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 236–254, Albuquerque, New Mexico. Association for Computational Linguistics.

Zuzana Sokolová, Maroš Harahus, Daniel Hládek, and Ján Staš. 2025. Annotated slovak datasets for toxicity, hate speech, and sentiment analysis. *Jazykovedný časopis – Journal of Linguistics*, 76(1):279–289.

Michal Štefánik, Marek Kadlčík, Piotr Gramacki, and Petr Sojka. 2023. Resources and few-shot learners for in-context learning in slavic languages. *arXiv preprint arXiv:2304.01922*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *ArXiv*, abs/2409.10173.

Marek Suppa and Jergus Adamec. 2020. A summarization dataset of Slovak news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6725–6730, Marseille, France. European Language Resources Association.

Marek Suppa, Andrej Ridzik, Daniel Hládek, Tomáš Javůrek, Viktória Ondrejová, Kristína Sásiková, Martin Tamajka, and Marian Simko. 2025. skLEP: A Slovak general language understanding benchmark. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26716–26743, Vienna, Austria. Association for Computational Linguistics.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. Efficient multilingual language model compression through vocabulary trimming. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14725–14739, Singapore. Association for Computational Linguistics.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *ArXiv*, abs/2509.20354.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *ArXiv*, abs/2402.05672.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general chinese embedding. *ArXiv*, abs/2309.07597.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292. Association for Computational Linguistics.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *ArXiv*, abs/2506.05176.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. FaMTEB: Massive text embedding benchmark in persian language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11441–11468. Association for Computational Linguistics.

A SkMTEB Pareto Frontier

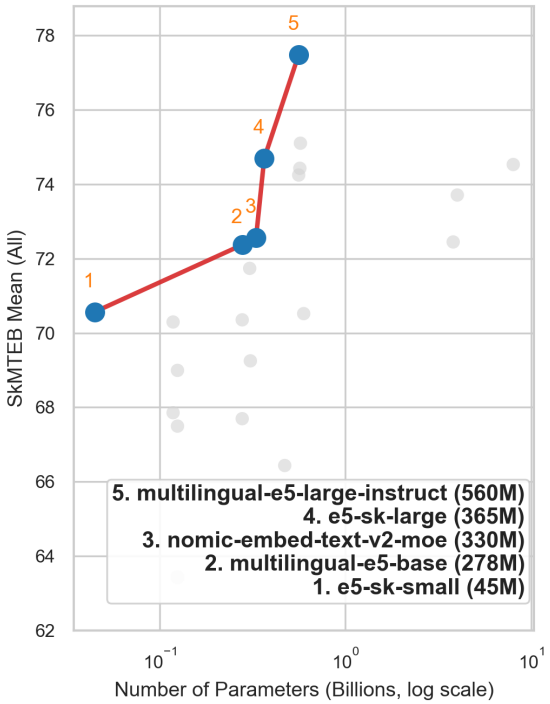

Figure 2: SkMTEB Pareto frontier (All). Each point is a model with mean SkMTEB score (y-axis) versus parameter count in billions (x-axis, log scale). Gray points show all evaluated models; the red polyline connects the Pareto-optimal set (no model simultaneously smaller and higher-scoring). Blue points are the frontier models; orange numbers above each point correspond to the ordered list in the legend (bottom-right), which reports model names and sizes. The y-axis is truncated to start at 62 to emphasize differences among competitive models.

B Slovak Pharmacy Reranking (Dataset Curation)

Data were scraped from two Slovak pharmacy websites, MojaLekaren and DrMax. Each of them has a forum page where customers can ask health-related questions. A certified pharmacist will answer the question. After scraping, the data were cleaned, formatted and anonymized by removing personal names. Duplicate questions were removed. Each dataset includes customer questions as reranking queries and corresponding pharmacist answers as the only positive answer in the reranking dataset. Therefore, a question-answer pair is a query-positive pair.

In addition, the DrMax dataset tags were also scraped. These tags provide a description of the

pair. The most common tags include *interactions*, *pain*, *side effects*, and *pregnancy*. The MojaLekaren dataset contains both tags and categories, which were merged into a single set of tags. Labels *eyes*, *hair*, and *pain* were included in the tags and categories. The most common tags were *advice of a pharmacist*, *digestion*, and *skin*.

For each query-positive pair, four negative answers are assigned. Negative answers are responses to different questions. The process of assigning negatives was as follows. If the query-positive pair has more than four tags, the four most frequent tags were selected. If exactly four tags are present, all of them are used. For each tag, a negative answer is retrieved from a different question-answer pair that shares the same tag.

If the tag is unique, a random answer has been selected. Each selected negative answer is then compared with the original positive answer using cosine similarity, and only answers whose similarity falls within a threshold are retained. If the similarity constraint is not satisfied, a different answer is selected. Sentence embeddings are computed using the paraphrase-multilingual-MiniLM-L12-v2 model, and cosine similarity is calculated using the scikit-learn library.

If fewer than four tags are available for a given query-answer pair, the following procedure is applied. For each available tag, one negative answer is selected as described above. The remaining negative answers are selected from answers that share the most frequent tag. If no suitable answers are found for that tag, the next most frequent tag is considered. If no suitable answers are available, a random answer is selected.

The dataset mojalekaren-reranking contains 738 rows, with a lower similarity threshold of 0.55 and an upper threshold of 0.85. Dataset drmax-reranking contains 4676 rows, with a lower similarity threshold of 0.3 and an upper threshold of 0.9. The upper threshold was selected after manually reviewing several examples to determine whether a negative sample qualifies as a hard negative or a false negative. The lower threshold was established to filter out overly simplistic negatives, thereby ensuring the inclusion of sufficiently challenging samples for the model.

C Slovak STS Synthetic (Dataset Curation)

To construct the dataset, we paired original sentences from the Slovak Summarization Dataset (Ondrejova and Suppa, 2024) with synthetic counterparts generated by the GPT-5 model. For every original sentence, GPT-5 was prompted to produce six distinct variations, each corresponding to one of the six semantic similarity levels (0–5) defined by (Agirre et al., 2013). The STS score definitions used for generation are summarized in Table 4.

STS Score	Description
0	The two sentences are on different topics.
1	The two sentences are not equivalent, but are on the same topic.
2	The two sentences are not equivalent, but share some details.
3	The two sentences are roughly equivalent, but some important information differs or is missing.
4	The two sentences are mostly equivalent, but some unimportant details differ.
5	The two sentences are completely equivalent, as they mean the same thing.

Table 4: Semantic Textual Similarity (STS) score definitions (Agirre et al., 2013)

The following constraints were applied to the original sentences: they had to contain at least 60 characters and no more than 200 characters, and they had to consist of a single sentence (multi-sentence examples were omitted). GPT-5 was required to generate exactly six sentences, one for each similarity score. The dataset generation process was repeated four times, with 20 sentences per iteration, with the prompt refined at each stage to address any generation issues (e.g., similarity between score 4 and 5 or score 1 and 2). GPT-5 was provided with two examples from the authors in the prompt, including explanations for why each sentence received its score and a detailed description of all similarity scores. The temperature parameter was set to 1. A total of 6594 sentence pairs were generated.

The annotation process took place after generation. A random sample of 300 sentence pairs was

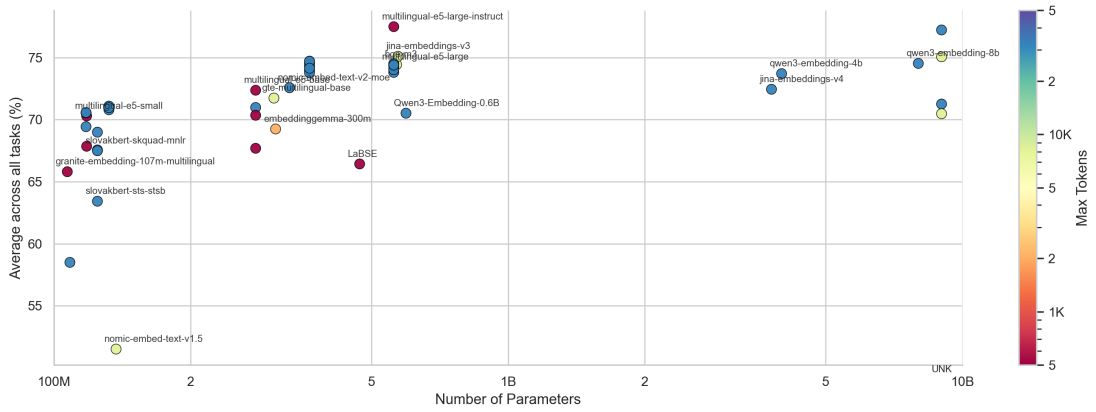

Figure 3: Model Size vs. Average Performance

selected from the full generated dataset, with 50 pairs drawn from each similarity score. The annotation guidelines did not explicitly specify the target distribution of similarity scores. Three native Slovak speakers served as annotators: two are coauthors of this work, and the third is a colleague who volunteered without any expectation of remuneration. All annotators received information about the task (evaluating each pair based on the similarity score), explanations for each STS score, and five examples with explanations. The first two examples were identical to those used in the GPT-5 prompt. The remaining three examples were pairs of original and generated sentences, together with an explanation of why GPT-5 generated that particular sentence for that similarity score. These sentences were selected from the entire STS dataset but were not part of the annotated subset. All three annotators independently annotated the same set of 300 sentence pairs.

After annotation, the results were validated. Pairwise inter-annotator agreement scores, including Cohen's kappa, are reported in Table 5. All Cohen's kappa values exceed 0.5, indicating moderate agreement. Across all annotators, Fleiss' kappa is 0.56 and Krippendorff's alpha is 0.92.

During the annotation process, the annotators found a Hungarian sentence within the original text, although the corresponding generated sentence was in Slovak. After this, the authors manually reviewed the entire dataset to verify language consistency. Two Hungarian sentences were found, and 12 pairs were deleted from the dataset. The dataset was subsequently cleaned and formatted. To prevent data contamination, all original sentences selected for annotation and all their generated sentences were removed from the non-annotated split. This reduced the number of original sentences from 1099 to 821.

The final dataset is divided into two parts: a train split consisting of the non-annotated portion (4926 rows), where the similarity score ranges from 0 to 5, and a test split consisting of the annotated portion (298 rows), where the similarity score is the average score of three annotators.

C.1 Generation Prompt

The full system prompt used for generating sentence variants is provided below. The prompt is written entirely in Slovak to ensure natural, native-quality outputs. It instructs the model to:

1. Read an original Slovak sentence from the SlovakSum corpus.
2. Generate six variant sentences, one for each STS score (0–5).
3. Provide a Slovak explanation justifying each score assignment.
4. Follow detailed score boundary definitions distinguishing between adjacent scores (e.g., score 4 vs. 5, score 1 vs. 2).
5. Maintain natural Slovak language quality suitable for native speakers.
6. Return output in a structured JSON format.

The prompt (in Figure 4) includes two complete worked examples demonstrating expected outputs for different source sentences. Key design choices include:

Annotator pair	Overall Accuracy	Pearson Corr.	Spearman Corr.	Cohen's κ
[1,2]	0.67	0.94	0.94	0.60
[1,3]	0.63	0.91	0.92	0.56
[2,3]	0.61	0.93	0.93	0.53

Table 5: Pairwise agreement between annotators

- **Explicit boundary definitions:** The prompt carefully distinguishes adjacent scores to reduce annotator confusion (e.g., score 4 requires "nearly equivalent with minor generalizations" while score 5 requires "complete paraphrase with all information preserved").
- **Anti-repetition constraints:** For score 0 (different topic), the prompt explicitly prohibits defaulting to common topics like weather or tourism.
- **Quality control:** The prompt instructs the model to internally verify each sentence sounds natural to a native Slovak speaker before outputting.

System Prompt for STS Generation (Slovak) – Abbreviated

```
Si expert na slovenský jazyk, parafrázovanie a anotáciu sémantickej podobnosti.
Tvoja úloha:
1. Prečítať si jednu slovenskú vetu (ORIGINÁL).
2. Pre KAŽDÉ skóre podobnosti 0–5 (STS) vytvor:
  – jednu slovenskú vetu (VARIANT),
  – jedno krátke vysvetlenie v slovenčine.
```

DEFINÍCIE SKÓRE 0–5

Všeobecné pravidlo: „Dôležitá informácia" = hlavná udalosť, aktér, miesto, čas, množstvo, základná pointa. Ak zmeníš alebo vynecháš dôležitú informáciu, skóre NEMÔŽE byť 4 ani 5.

Skóre 0 – ÚPLNE INÁ TÉMA
Vety sú o úplne odlišných veciach. Žiadne zmysluplné tematické prepojenie.

Skóre 1 – ROVNAKÁ ŠIROKÁ TÉMA, ALE TAKMER ŽIADNE DETAILY SPOLOČNÉ
Vety majú spoločnú tému, hovoria však o iných konkrétnych udalostiach.

Skóre 2 – NIE SÚ EKVIVALENTNÉ, ALE ZDIEĽAJÚ NIEKTORÉ DETAILY
Vety sa týkajú podobnej témy a zdieľajú niektoré konkrétne prvky.

Skóre 3 – PRIBLIŽNÁ EKVIVALENCIA, ALE DÔLEŽITÁ INFORMÁCIA SA LÍŠI
Hlavná udalosť je podobná, ALE zmení sa aspoň JEDEN DÔLEŽITÝ aspekt.

Skóre 4 – TAKMER EKVIVALENTNÉ, LÍŠIA SA LEN MENEJ PODSTATNÉ DETAILY
Hlavná udalosť, aktéri, miesto, čas a čísla sú rovnaké. Rozdiely sú minimálne.

Skóre 5 – ÚPLNE ROVNAKÝ VÝZNAM (PARAFRÁZA)
Všetky dôležité informácie sú zachované. VARIANT je parafráza s inými slovami.

VÝSTUPNÝ FORMÁT – JSON objekt:
{"0": {"sentence": "...", "explanation": "..."}, ..., "5": {...}}

Figure 4: Abbreviated system prompt for STS sentence generation (Slovak). The full prompt includes detailed score boundary definitions, two complete worked examples, and additional quality guidelines; see supplementary materials for the complete version.

Model (↓)	Params	Dim
Small models (<130M)		
e5-sk-small	45M	384
granite-embedding-107m-multilingual	107M	384
static-similarity-mrl-multilingual-v1	108M	1024
multilingual-e5-small	118M	384
paraphrase-multilingual-MiniLM-L12-v2	118M	768
slovakbert-skquad-mnlr	125M	768
slovakbert-sts-stsb	125M	768
sturovec-base	125M	768
Base models (>=130M, <350M)		
nomic-embed-text-v1.5	137M	768
granite-embedding-278m-multilingual	278M	768
multilingual-e5-base	278M	768
paraphrase-multilingual-mpnet-base-v2	278M	768
gte-multilingual-base	305M	768
embeddinggemma-300m	308M	768
nomic-embed-text-v2-moe	330M	1024
Large models (>=350M)		
e5-sk-large	365M	1024
LaBSE	471M	768
multilingual-e5-large	560M	1024
multilingual-e5-large-instruct	560M	1024
bge-m3	568M	1024
snowflake-arctic-embed-l-v2.0	568M	1024
jina-embeddings-v3	572M	1024
Qwen3-Embedding-0.6B	596M	1024
jina-embeddings-v4	3.8B	2048
Qwen3-Embedding-4B	4B	2560
Qwen3-Embedding-8B	8B	4096
API access models		
embed-v4.0	-	1536
text-embedding-3-small	-	1536
text-embedding-3-large	-	3072
gemini-embedding-001	-	3072
amazon-titan-embed-text-v2	-	1024

Table 6: Model size and embedding dimension. Models are grouped by parameter-size buckets with API-access models listed last. This table is intended as a companion reference to the main results tables and uses the same model ordering, naming, and highlighting conventions.

D Model Parameter Size and Embedding Dimension

E SkMTEB Task Catalogue

This appendix provides comprehensive descriptions of all tasks included in the SkMTEB benchmark, organized by task type. The benchmark comprises 31 distinct task definitions across 7 categories, with several bitext mining tasks evaluated on multiple language pair subsets. For each task, we briefly describe the data source, creation methodology, temporal coverage, licensing, and evaluation specifics.

B6: Statistics for data. We report dataset sizes and evaluation split details where applicable in the task catalogue below (e.g., evaluation size, temporal coverage, and annotations). For datasets we create or curate, Appendix B and Appendix C provide additional statistics such as row counts, sampling constraints, and annotation sizes.

E.1 Retrieval Tasks

E.1.1 SKQuadRetrieval

Source: SK-QuAD dataset (Hládek et al., 2023)
HuggingFace: `TUKE-KEMT/retrieval-skquad`
Description: A question-answering retrieval task that evaluates Slovak search performance using questions and answers derived from the SK-QuAD dataset. The task measures relevance with scores assigned to answers based on their relevancy to corresponding questions.
Domain: Encyclopaedic
Task Subtype: Question answering
Annotations: Human-annotated relevance scores
License: CC-BY-NC-SA-4.0
Main Metric: nDCG@10

E.1.2 SlovakSumRetrieval

Source: SlovakSum dataset (Ondrejova and Suppa, 2024)
HuggingFace: `NaiveNeuron/slovaksum`
Description: A Slovak news summarization dataset consisting of over 200,000 news articles with titles and short abstracts obtained from multiple Slovak newspapers. Originally intended as a summarization task, reformulated to a retrieval task where article abstracts serve as queries to retrieve full documents.
Domain: News, Social, Web
Task Subtype: Article retrieval
Temporal Coverage: 2015–2022
Annotations: Derived from document structure
License: OpenRAIL
Main Metric: nDCG@10
Evaluation Size: 600 query-document pairs

E.1.3 SMESumRetrieval

Source: SMESum dataset (Suppa and Adamec, 2020)
HuggingFace: `NaiveNeuron/SMESum`
Description: A Slovak news summarization dataset consisting of 80,000 news articles with titles and introductions from the SME news portal. Reformulated as a retrieval task where article introductions serve as queries to retrieve full documents.
Domain: News, Social, Web
Task Subtype: Article retrieval
Temporal Coverage: 2013–2019
Annotations: Derived from document structure
License: Not specified
Main Metric: nDCG@10
Evaluation Size: 600 query-document pairs

E.1.4 BelebeleRetrieval

Source: Belebele benchmark (Bandarkar et al., 2024)
HuggingFace: `facebook/belebele`
Description: A multiple-choice machine reading comprehension dataset spanning 122 language variants. For Slovak (slk_Latn), the task involves retrieving relevant passages given questions.
Domain: Web, News
Task Subtype: Question answering
Annotations: Expert-annotated
License: CC-BY-SA-4.0
Main Metric: nDCG@10

E.1.5 WebFAQRetrieval

Source: WebFAQ corpus (Dinzinger et al., 2025)
HuggingFace: `mteb/WebFAQRetrieval`
Description: A broad-coverage corpus of natural question-answer pairs gathered from FAQ pages on the web, covering 75 languages including Slovak.
Domain: Web
Task Subtype: Question answering
Temporal Coverage: 2022–2024
Annotations: Derived from FAQ structure
License: CC-BY-4.0
Main Metric: nDCG@10

E.2 Semantic Textual Similarity Tasks

E.2.1 SlovakSTS

Source: skLEP benchmark (Suppa et al., 2025)
HuggingFace: `slovak-nlp/sklep` (subset: sts)
Description: Professional Slovak translation of the original GLUE STSb dataset. Contains sentence pairs with human-annotated similarity scores ranging from 0 (completely dissimilar) to 5 (completely similar).
Domain: Blog, News
Task Subtype: Textual Entailment
Temporal Coverage: 2025
Annotations: Human-annotated, machine-translated and verified
License: CC-BY-SA-4.0
Main Metric: Spearman correlation

E.2.2 SlovakSumSTS

Source: Synthetic dataset from SlovakSum
HuggingFace: `slovak-nlp/slovak-sts-synthetic`
Description: Sentence pairs for semantic textual similarity scoring in Slovak. Pairs were generated using text from the SlovakSum dataset, where an LLM created corresponding sentence pairs for each STS score (0–5). The test split pairs were verified by human annotators.
Domain: News
Task Subtype: Textual Entailment
Temporal Coverage: 2025
Annotations: LM-generated and human-reviewed
License: CC-BY-NC-4.0
Main Metric: Spearman correlation

E.3 Pair Classification Tasks

E.3.1 SlovakNLI

Source: Handwritten Slovak NLI dataset
HuggingFace: `natalia-nk/NLI-SK-annotated`
Description: Slovak handwritten annotated Natural Language Inference dataset containing premise-hypothesis pairs. Labels indicate entailment or contradiction relationships.
Domain: News, Web
Task Subtype: Textual Entailment
Temporal Coverage: 2024–2025
Annotations: Human-annotated
License: Not specified
Main Metric: Average Precision (max_ap)

E.3.2 SlovakRTE

Source: skLEP benchmark (Suppa et al., 2025)
HuggingFace: `slovak-nlp/sklep` (subset: rte)
Description: Slovak Recognizing Textual Entailment dataset. Professional translation and human verification of English RTE datasets for Slovak. Binary classification task (entailment vs. not entailment).
Domain: News, Web
Task Subtype: Textual Entailment
Temporal Coverage: 2025
Annotations: Human-annotated, machine-translated and verified
License: CC-BY-SA-4.0
Main Metric: Average Precision (max_ap)

E.3.3 DemagogSKNLI

Source: Demagog.sk fact-checking portal
HuggingFace: `NaiveNeuron/DemagogSK`
Description: Slovak Natural Language Inference dataset created from Demagog.sk fact-checking data. Evidence-claim pairs where professional fact-checkers' analysis (evidence) is paired with political statements (claims). Labels indicate whether evidence supports (Pravda) or refutes (Nepravda) the claim.
Domain: Government, News
Task Subtype: Claim verification
Temporal Coverage: 2010–2025
Annotations: Expert-annotated by fact-checkers
License: Not specified
Main Metric: Average Precision (max_ap)

E.4 Classification Tasks

E.4.1 SlovakHateSpeechClassification.v2

Source: TUKE-KEMT hate speech dataset
HuggingFace: `mteb/slovak_hate_speech`
Description: Social network posts with human annotations for hateful or offensive language in Slovak. Binary classification (toxic vs. not toxic). Version 2 corrects errors from the original dataset.
Domain: Social
Task Subtype: Sentiment/Hate speech
Temporal Coverage: 2024
Annotations: Human-annotated
License: CC-BY-SA-4.0
Main Metric: Accuracy

E.4.2 SlovakMovieReviewSentimentClassification.v2

Source: CSFD movie database (Štefánik et al., 2023)
HuggingFace: `mteb/slovak_movie_review_sentiment`
Description: User reviews of movies from the CSFD movie database with binary sentiment classes (positive, negative). Version 2 corrects errors from the original dataset.
Domain: Reviews
Task Subtype: Sentiment/Hate speech
Temporal Coverage: 2002–2020
Annotations: Derived from user ratings
License: CC-BY-NC-SA-4.0
Main Metric: Accuracy

E.4.3 SIB200Classification

Source: SIB-200 dataset (Adelani et al., 2023)
HuggingFace: `mteb/sib200` (subset: slk_Latn)
Description: The largest publicly available topic classification dataset based on Flores-200, covering 205 languages and dialects. Annotated for 7 topics: science/technology, travel, politics, sports, health, entertainment, and geography. Labels transferred from English via human translation.
Domain: News
Task Subtype: Topic classification
Temporal Coverage: 2023–2024
Annotations: Expert-annotated for English, human-translated
License: CC-BY-SA-4.0
Main Metric: Accuracy

E.4.4 MultilingualSentimentClassification

Source: Cross-lingual sentiment dataset (Mollanorozy et al., 2023)
HuggingFace: `mteb/multilingual-sentiment-classification` (subset: slk)
Description: Sentiment classification dataset with binary labels (positive vs. negative) covering 30 languages and dialects including Slovak.
Domain: Reviews
Task Subtype: Sentiment/Hate speech
Temporal Coverage: 2022
Annotations: Derived
License: Not specified
Main Metric: Accuracy

E.4.5 SlovakParlaSentClassification

Source: ParlaSent corpus (Mochtak et al., 2024)
HuggingFace: `classla/ParlaSent` (subset: SK)
Description: Slovak parliamentary sentiment classification from the ParlaSent corpus. Contains sentences from parliamentary debates with 3-level sentiment annotations (negative, neutral, positive).
Domain: Government, Spoken
Task Subtype: Sentiment/Hate speech
Temporal Coverage: 2018
Annotations: Human-annotated
License: CC-BY-SA-4.0
Main Metric: Accuracy

E.4.6 MultiEupSlovakPartyClassification

Source: Multi-EuP v2 corpus (Yang et al., 2024)
HuggingFace: `unimelb-nlp/MultiEup-v2`
Description: Multi-class classification to predict European Parliament political group from native Slovak

speeches. Uses only speeches originally delivered in Slovak from the Multi-EuP v2 corpus.
Domain: Government, Spoken
Task Subtype: Topic classification
Temporal Coverage: 2020–2024
Annotations: Derived from parliamentary metadata
License: CC-BY-4.0
Main Metric: Accuracy

E.4.7 MultiEupSlovakGenderClassification

Source: Multi-EuP v2 corpus (Yang et al., 2024)
HuggingFace: `unimelb-nlp/MultiEup-v2`
Description: Binary classification to predict gender of Members of the European Parliament from native Slovak speeches. Uses only speeches originally delivered in Slovak.
Domain: Government, Spoken
Task Subtype: Topic classification
Temporal Coverage: 2020–2024
Annotations: Derived from parliamentary metadata
License: CC-BY-4.0
Main Metric: Accuracy

E.5 Reranking Tasks

E.5.1 SkQuadReranking

Source: SK-QuAD dataset (Hládek et al., 2023)
HuggingFace: `TUKE-KEMT/reranking-skquad`
Description: Article retrieval reranking task derived from SK-QuAD. Given a query and candidate documents, the model must rank documents by relevance.
Domain: Encyclopaedic
Task Subtype: Article retrieval
Annotations: Derived
License: CC-BY-SA-4.0
Main Metric: MAP@1000

E.5.2 SlovakPharmacyDrMaxReranking

Source: DrMax pharmacy website
HuggingFace: `slovak-nlp/slovak-pharmacy-drmax-reranking`
Description: Reranking dataset from Q&A content on the DrMax pharmacy website. Questions about medications, health conditions, and pharmaceutical advice with answers from qualified pharmacists.
Domain: Medical, Web
Task Subtype: Article retrieval
Temporal Coverage: 2025
Annotations: Derived
License: CC-BY-NC-ND-4.0
Main Metric: MAP@1000

E.5.3 SlovakPharmacyMojaLekarenReranking

Source: MojaLekaren pharmacy website
HuggingFace: `slovak-nlp/slovak-pharmacy-mojalekaren-reranking`
Description: Reranking dataset from Q&A content on the MojaLekaren pharmacy website. Questions about medications, health conditions, and pharmaceutical advice with answers from qualified pharmacists.
Domain: Medical, Web
Task Subtype: Article retrieval
Temporal Coverage: 2025

Annotations: Derived
License: CC-BY-NC-ND-4.0
Main Metric: MAP@1000

E.6 Clustering Tasks

E.6.1 SIB200ClusteringS2S

Source: SIB-200 dataset (Adelani et al., 2023)
HuggingFace: `mteb/sib200` (subset: slk_Latn)
Description: Clustering variant of the SIB-200 topic classification dataset. Up to 1,004 documents clustered into 7 thematic topics covering science/technology, travel, politics, sports, health, entertainment, and geography.
Domain: News
Task Subtype: Thematic clustering
Annotations: Expert-annotated for English, human-translated
License: CC-BY-SA-4.0
Main Metric: V-measure

E.6.2 PravdaSKTagClustering

Source: Pravda.sk news portal
HuggingFace: `NaiveNeuron/pravda-sk-tag-clustering`
Description: Clustering of Slovak news articles from Pravda.sk based on article tags. Articles grouped into 50 thematic categories including Slovak politics, international affairs, events, and various topics. Uses title + summary as input.
Domain: News
Task Subtype: Thematic clustering, Topic classification
Temporal Coverage: 2014–2024
Annotations: Derived from article tags
License: Not specified
Main Metric: V-measure
Maximum Documents: 2,048

E.6.3 PravdaSKURLClustering

Source: Pravda.sk news portal
HuggingFace: `NaiveNeuron/pravda-sk-url-clustering`
Description: Clustering of Slovak news articles from Pravda.sk based on URL structure. Articles organized into 50 editorial categories reflecting the portal's content organization, including news, sports, culture, economy, health, travel, celebrity, and science sections.
Domain: News
Task Subtype: Thematic clustering, Topic classification
Temporal Coverage: 2014–2024
Annotations: Derived from URL structure
License: Not specified
Main Metric: V-measure

E.6.4 SlovakSumURLClustering

Source: SlovakSum dataset (Ondrejova and Suppa, 2024)
HuggingFace: `kiviki/slovaksum-url-clustering`
Description: Clustering of Slovak news articles from SlovakSum based on URL structure. Articles organized into 12 editorial categories including sports, culture, economy, health, travel, politics, and technology.
Domain: News

Task Subtype: Thematic clustering, Topic classification
Temporal Coverage: 2015–2022
Annotations: Derived from URL structure
License: Not specified
Main Metric: V-measure

E.6.5 SMESumCategoryClustering

Source: SMESum dataset (Suppa and Adamec, 2020)
HuggingFace: `NaiveNeuron/SMESum`
Description: Clustering of Slovak news articles from SMESum based on news categories. Articles organized into 11 thematic categories covering politics, economy, sports, culture, and other news domains. Articles with "none" category are excluded.
Domain: News
Task Subtype: Thematic clustering, Topic classification
Temporal Coverage: 2013–2019
Annotations: Derived from category metadata
License: Not specified
Main Metric: V-measure

E.7 Bitext Mining Tasks

E.7.1 OpusSlovakEnglishBitextMining

Source: OPUS-100 corpus (Zhang et al., 2020)
HuggingFace: `Helsinki-NLP/opus-100` (subset: en-sk)
Description: Slovak-English parallel sentences from OPUS-100, a multilingual dataset with 100 languages for evaluating massively multilingual neural machine translation.
Domain: Web, Subtitles, Fiction, Non-fiction
Temporal Coverage: 2000–2020
Annotations: Derived from parallel corpora
License: Not specified
Main Metric: F1

E.7.2 TatoebaBitextMining

Source: Tatoeba corpus
HuggingFace: `mteb/tatoeba-bitext-mining` (subset: slk-eng)
Description: 1,000 English-aligned sentence pairs for Slovak based on the Tatoeba corpus, a community-contributed collection of sentences and translations.
Domain: Written
Temporal Coverage: 2006–2021
Annotations: Human-annotated by community contributors
License: CC-BY-2.0
Main Metric: F1

E.7.3 FloresBitextMining

Source: FLORES benchmark (Goyal et al., 2022)
HuggingFace: `mteb/FloresBitextMining`
Subsets: eng_Latn-slk_Latn, ces_Latn-slk_Latn
Description: Benchmark dataset for machine translation between English and low-resource languages. Slovak pairs include English-Slovak and Czech-Slovak alignments.
Domain: Non-fiction, Encyclopaedic
Temporal Coverage: 2022
Annotations: Human-annotated

License: CC-BY-SA-4.0
Main Metric: F1

E.7.4 NTREXBitextMining

Source: NTREX-128 dataset (Federmann et al., 2022)
HuggingFace: `mteb/NTREXBitextMining`
Subsets: eng_Latn-slk_Latn, ces_Latn-slk_Latn
Description: News Test References for MT Evaluation covering 128 languages. Slovak pairs include English-Slovak and Czech-Slovak alignments with 1,997 parallel sentences each.
Domain: News
Temporal Coverage: 2019–2022
Annotations: Expert-annotated, human-translated
License: CC-BY-SA-4.0
Main Metric: F1

E.7.5 WebFAQBitextMiningQuestions

Source: WebFAQ corpus (Dinzinger et al., 2025)
HuggingFace: `PaDaS-Lab/webfaq-bitexts`
Subsets: eng-slk, ces-slk
Description: Natural FAQ-style question-answer pairs aligned across languages. This task uses questions from aligned Q&A pairs for cross-lingual retrieval.
Domain: Web
Temporal Coverage: 2022–2024
Annotations: Human-annotated, human-translated
License: CC-BY-4.0
Main Metric: F1

E.7.6 WebFAQBitextMiningQAs

Source: WebFAQ corpus (Dinzinger et al., 2025)
HuggingFace: `PaDaS-Lab/webfaq-bitexts`
Subsets: eng-slk, ces-slk
Description: Natural FAQ-style question-answer pairs aligned across languages. This task uses concatenated question-answer pairs for cross-lingual retrieval.
Domain: Web
Temporal Coverage: 2022–2024
Annotations: Human-annotated, human-translated
License: CC-BY-4.0
Main Metric: F1

F Reproducibility Details

To ensure reproducibility, we provide complete training and evaluation details below.

F.1 Training Hyperparameters

Table 8 summarizes all hyperparameters used for fine-tuning our Slovak embedding models.

F.2 Computational Resources

All training experiments were conducted on a single NVIDIA H100 GPU (80GB). Training time per model:

- `e5-sk-small` (45M params): ∼30 minutes
- `e5-sk-large` (365M params): ∼50 minutes
- `sturovec-base` (125M params): ∼45 minutes

Benchmark evaluation of all 25+ models required approximately 48 GPU-hours on NVIDIA H100.

F.3 Vocabulary Trimming Procedure

We apply Pre-FT Vocabulary Trimming following Ushio et al. (2023):

1. Compute token frequencies on FineWeb2-Slovak corpus
2. Rank tokens by frequency in target language
3. Retain top 60K tokens (recommended threshold from original paper)
4. Resize embedding and output projection matrices
5. Fine-tune on downstream tasks

The 60K threshold was chosen based on the original vocabulary trimming paper's recommendation, which found this value provides good coverage while maximizing compression. For comparison, MTEB-NL (Banar et al., 2025) uses 50K tokens for Dutch.

F.4 Model and Data Availability

All released models and datasets are collected on Hugging Face at `https://huggingface.co/collections/slovak-nlp/skmteb`, and the code is available at `https://github.com/slovak-nlp/skmteb`.

Getting started. SkMTEB is registered as `MTEB(slk, v1)` in the `mteb` library; any embedding model can be evaluated end-to-end with a single command:

```
mteb run -m <model-id> -b "MTEB(slk, v1)"
```

Maintenance and versioning. Because SkMTEB lives inside the `mteb` framework, the benchmark remains runnable as the framework evolves. Each task is pinned to a specific Hugging Face dataset revision hash in its definition, so evaluation scores are reproducible across time. We plan to maintain a public leaderboard for community submissions on Hugging Face.

F.5 Evaluation Protocol

We use the MTEB evaluation framework (Muennighoff et al., 2023) with default settings. For classification tasks, we train a logistic regression classifier on embeddings with default scikit-learn parameters. All results are single-run evaluations with seed 42.

G SkMTEB Classification Results

Dataset	Task	Domain	Origin	License	Splits (train/valid/test)
Retrieval (5)					
BelebeleRetrieval	Rtrvl	Web, News	Native	CC-BY-SA-4.0	— / — / 900
SKQuadRetrieval	Rtrvl	Encyclopaedic	Native	CC-BY-NC-SA-4.0	— / — / 1,134
SlovakSumRetrieval	Rtrvl	News, Social	Native	OpenRAIL	— / — / 600
SMESumRetrieval	Rtrvl	News, Social	Native	N/A	— / — / 600
WebFAQRetrieval	Rtrvl	Web	Native	CC-BY-4.0	— / — / 3,153
Reranking (3)					
SkQuadReranking	Rrnk	Encyclopaedic	Native	CC-BY-SA-4.0	— / — / 1,133
⋆ SlovakPharmacyDrMaxReranking	Rrnk	Medical, Web	Author-created	CC-BY-NC-ND-4.0	— / — / 4,676
⋆ SlovakPharmacyMojaLekarenReranking	Rrnk	Medical, Web	Author-created	CC-BY-NC-ND-4.0	— / — / 738
Classification (7)					
MultiEupSlovakGenderClassification	Clf	Government	Native	CC-BY-4.0	508 / — / 128
MultiEupSlovakPartyClassification	Clf	Government	Native	CC-BY-4.0	502 / — / 126
MultilingualSentimentClassification	Clf	Reviews	Native	N/A	3,560 / 522 / 1,042
SIB200Classification	Clf	News	Native	CC-BY-SA-4.0	701 / 99 / 204
SlovakHateSpeechClassification.v2	Clf	Social	Native	CC-BY-SA-4.0	11,301 / — / 1,237
SlovakMovieReviewSentimentClassification.v2	Clf	Reviews	Native	CC-BY-NC-SA-4.0	20,181 / 2,083 / 2,008
SlovakParlaSentClassification	Clf	Government	Native	CC-BY-SA-4.0	2,080 / — / 520
Clustering (5)					
⋆ PravdaSKTagClustering	Clust	News	Author-created	N/A	— / — / 15,000
⋆ PravdaSKURLClustering	Clust	News	Author-created	N/A	— / — / 15,000
SIB200ClusteringS2S	Clust	News	Native	CC-BY-SA-4.0	— / — / 204
SMESumCategoryClustering	Clust	News	Native	N/A	— / — / 7,233
SlovakSumURLClustering	Clust	News	Native	N/A	— / — / 10,871
Bitext Mining (6)					
FloresBitextMining	Btxt	Non-fiction, Encyclopaedic	Native	CC-BY-SA-4.0	— / — / 1,012
NTREXBitextMining	Btxt	News	Native	CC-BY-SA-4.0	— / — / 1,997
OpusSlovakEnglishBitextMining	Btxt	Web, Subtitles	Native	N/A	1,000,000 / 2,000 / 2,000
Tatoeba	Btxt	Written	Native	CC-BY-2.0	— / — / 1,000
WebFAQBitextMiningQAs	Btxt	Web	Native	CC-BY-4.0	— / — / 2,551+1,823
WebFAQBitextMiningQuestions	Btxt	Web	Native	CC-BY-4.0	— / — / 2,551+1,823
Pair Classification (3)					
⋆ DemagogSKNLI	PrClf	Government, News	Author-created	N/A	— / — / 3,085
⋆ SlovakNLI	PrClf	News, Web	Author-created	N/A	— / — / 382
SlovakRTE	PrClf	News, Web	Translated	CC-BY-SA-4.0	2,490 / 277 / 1,660
STS (2)					
SlovakSTS	STS	News, Blog	Translated	CC-BY-SA-4.0	5,604 / 1,481 / 1,352
⋆ SlovakSumSTS	STS	News	Author-created	CC-BY-NC-4.0	4,926 / — / 298

Table 7: Datasheet for the 31 SkMTEB datasets. **Origin**: Native = natively Slovak; Translated = translated from another language; Author-created (⋆) = introduced in this work. **Splits** lists the sizes of the train / valid / test splits as provided by the task definition; "—" means the split is not provided. SkMTEB evaluation uses the test split in every case. **Task** codes: Rtrvl = Retrieval, Rrnk = Reranking, Clf = Classification, Clust = Clustering, Btxt = Bitext Mining, PrClf = Pair Classification, STS = Semantic Textual Similarity. For bitext mining, test sizes are per language pair; Flores and NTREX are evaluated on both eng–slk and ces–slk pairs, WebFAQ variants list both pair sizes (eng–slk / ces–slk), and Tatoeba and Opus use slk–eng only.

Hyperparameter	**Value**
Base models	mE5-small, mE5-large
Pooling strategy	Mean pooling
Max sequence length	256 tokens
Batch size	32
Learning rate	2×10^{-5}
LR scheduler	Linear with warmup
Warmup ratio	10% of steps
Training epochs	3
Optimizer	AdamW
Weight decay	0.01
Random seed	42
Precision	FP32
Loss Functions	
STS task	Cosine Similarity Loss
Other tasks	Multiple Negatives Ranking Loss

Table 8: Training hyperparameters for e5-sk models.

Model (↓)	MultiEupSlovakGenderClassification	MultiEupSlovakPartyClassification	MultilingualSentimentClassification	SIB200Classification	SlovakHateSpeechClassification.v2	SlovakMovieReviewSentimentClassification.v2	SlovakParlaSentClassification	Avg
Small models (<130M)								
e5-sk-small (45M)	58.52	45.24	84.45	73.04	54.92	60.48	49.23	60.84
granite-embedding-107m-multilingual (107M)	53.98	38.02	74.73	70.93	54.94	56.99	42.67	56.04
static-similarity-mrl-multilingual-v1 (108M)	51.72	48.49	56.86	41.86	48.82	55.33	39.02	48.87
multilingual-e5-small (118M)	56.09	43.57	83.53	71.72	54.65	60.31	48.81	59.81
paraphrase-multilingual-MiniLM-L12-v2 (118M)	52.81	39.52	89.01	71.18	53.99	64.12	55.17	60.83
slovakbert-skquad-mnlr (125M)	58.28	48.81	85.87	67.75	55.80	66.78	54.13	62.49
slovakbert-sts-stsb (125M)	62.42	49.44	86.66	62.70	52.55	70.64	56.90	63.05
sturovec-base (125M)	55.16	44.29	86.45	71.76	55.46	69.98	52.23	62.19
Base models (>=130M, <350M)								
nomic-embed-text-v1.5 (137M)	57.58	39.52	74.31	42.70	51.62	55.78	38.27	51.40
granite-embedding-278m-multilingual (278M)	56.95	42.14	76.76	70.64	55.02	59.09	44.08	57.81
multilingual-e5-base (278M)	61.88	44.68	85.88	70.10	55.93	65.28	52.67	62.35
paraphrase-multilingual-mpnet-base-v2 (278M)	59.30	41.51	89.74	75.69	54.68	68.60	**58.85**	64.05
gte-multilingual-base (305M)	57.42	43.49	85.14	75.74	53.19	65.72	51.21	61.70
embeddinggemma-300m (308M)	53.59	46.43	88.83	71.32	53.48	72.15	48.77	62.08
nomic-embed-text-v2-moe (330M)	60.86	**49.84**	84.96	76.08	54.37	70.99	50.79	63.98
Large models (>=350M)								
e5-sk-large (365M)	63.12	49.76	90.41	74.61	58.25	72.44	55.81	66.34
LaBSE (471M)	59.22	46.90	84.17	57.11	57.34	59.44	46.90	58.73
multilingual-e5-large (560M)	61.09	47.94	88.98	75.69	57.11	70.70	55.88	65.34
multilingual-e5-large-instruct (560M)	62.81	46.51	94.36	**83.48**	59.86	86.15	58.77	70.28
bge-m3 (568M)	58.59	45.48	93.24	72.55	57.71	81.97	58.10	66.81
snowflake-arctic-embed-l-v2.0 (568M)	58.91	44.76	89.74	75.00	57.96	67.58	49.87	63.40
jina-embeddings-v3 (572M)	57.27	45.00	93.48	76.23	55.16	83.54	57.58	66.89
Qwen3-Embedding-0.6B (596M)	52.42	41.51	83.98	79.31	55.55	76.25	50.83	62.84
jina-embeddings-v4 (3.8B)	60.08	44.52	87.19	77.89	55.33	66.83	49.12	62.99
Qwen3-Embedding-4B (4B)	58.67	43.10	87.94	78.19	55.81	72.96	53.17	64.26
Qwen3-Embedding-8B (8B)	62.03	44.21	87.67	78.92	56.94	76.50	55.33	65.94
API access models								
embed-v4.0	57.66	45.48	81.02	72.75	54.79	62.42	46.73	60.12
text-embedding-3-small	57.11	47.78	83.46	74.41	56.35	67.34	48.46	62.13
text-embedding-3-large	61.09	47.70	90.62	79.17	56.87	79.64	53.31	66.91
gemini-embedding-001	**66.64**	48.97	**94.99**	75.78	**62.83**	**91.08**	56.77	**71.01**
amazon-titan-embed-text-v2	55.62	38.49	85.09	63.73	52.85	69.61	49.44	59.26

Table 9: SkMTEB classification results (percent). Columns correspond to each classification dataset; the final **Avg** column is the unweighted mean across available classification tasks for a model. Models are grouped by size bucket (Small/Base/Large), followed by API-access models. The best result per task is **bolded** with the runner-up underlined.