TR-TEB: Turkish Text Embedding Benchmark

Ömer Arslan, Atalay Çelik, Yusuf Aslan, Hasan Fatih Durkaya,
Mustafa Furkan Zenginoğlu, Musa Alperen Yılmaz,
Merve Gül Kantarcı, Mehmet Haklıdır
TÜBİTAK BİLGEM
Kocaeli, Türkiye
{omer.arslan, atalay.celik, aslan.yusuf, hasan.durkaya, mustafa.zenginoglu,
alperen.yilmaz, merve.kantarci, mehmet.haklidir}@tubitak.gov.tr

Abstract

Text embeddings are central to modern natural language processing, enabling several downstream tasks. Despite their significance, existing evaluation frameworks primarily target English and other high-resource languages, leaving critical gaps for languages such as Turkish. To address this, we present TR-TEB (Turkish Text Embedding Benchmark), the first comprehensive, standardized, and reproducible benchmark for Turkish text embeddings. TR-TEB spans five core task categories: classification, pair classification, clustering, retrieval, and semantic textual similarity. It is supported by a diverse dataset portfolio that integrates 14 curated open-source resources, 26 high-quality translated datasets, and 7 newly constructed Turkish-specific datasets designed to capture the language's unique characteristics. We test our framework by comparing 45 well-known open-source embedding models. As the first unified evaluation suite, TR-TEB serves as a core tool for the Turkish embedding research community, establishing a systematic basis for model comparison and improvement. Furthermore, its benchmarking methodology and dataset creation process provide a blueprint for extending robust embedding evaluation to other low-resource languages.

Keywords: Benchmark, Text Embeddings, Natural Language Processing, Turkish

1. Introduction

Even in the era of large language models, embedding-based systems continue to play a central role in many natural language processing (NLP) applications due to their effectiveness. They form the foundation of tasks such as recommendation systems, search engines, and retrieval-augmented generation (RAG), where capturing semantic relationships between texts is crucial. However, evaluating these embeddings remains inherently challenging, as it involves representing variable-length texts as fixed-size numerical vectors and ensuring that these representations capture meaningful linguistic nuances. Since embedding quality is closely tied to the linguistic characteristics of each language, evaluation frameworks must be adapted accordingly. Yet current systems often fail to meet these language-specific needs. A prominent effort in this direction, the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023), provides a valuable foundation with its diverse set of tasks, clear methodology, and extensible, community-friendly framework. However, despite its claim of being multilingual, MTEB primarily focuses on English and a few other high-resource languages, lacking true linguistic balance. This imbalance limits the assessment of model performance in languages with distinct morphological, syntactic, and semantic characteristics, such as Turkish, which presents unique challenges due to its agglutinative structure, rich morphology, and flexible word order. Although MTEB is not the only initiative targeting Turkish, existing evaluations remain fragmented, task-specific, and limited in scale, hindering systematic comparison. The lack of a unified, high-quality evaluation framework for Turkish embeddings creates a critical gap for both research and practical applications. In this work, we introduce the Turkish Text Embedding Benchmark (TR-TEB) to address this gap. Specifically, we propose:

- A uniform evaluation suite that covers all core tasks for monolingual embedding evaluation, including classification, pair classification, clustering, retrieval, and semantic textual similarity (STS).
- A diverse Turkish dataset portfolio spanning various sources, text lengths, and domains, curated from high-quality translations of large-scale English datasets within the MTEB framework, open-source Turkish resources, and newly curated datasets designed to fill existing gaps.
- A standardized framework that enables systematic benchmarking and allows future research to test new embedding models.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11028–11044
11–16 May 2026.

2. Related Work

In early approaches, the evaluation of language models has been largely driven by general Natural Language Understanding (NLU) benchmarks. Frameworks like GLUE (Wang et al., 2019) and SuperGLUE (Wang et al., 2020) were instrumental in this progress. They primarily target classification, inference, and coreference. While foundational for general NLU assessment, several studies (Poświata et al., 2024; Pham et al., 2025) note that these frameworks are unsuitable for direct embedding evaluation, as they offer limited insight into embedding generalization and focus solely on English (Zinvandi et al., 2025). To address these gaps, dedicated benchmarks have emerged. SentEval (Conneau and Kiela, 2018) introduced a toolkit for evaluating universal sentence representations, while BEIR (Thakur et al., 2021) provided a detailed benchmark for zero-shot retrieval. The increasing capacity of models led to performance saturation on earlier benchmarks like GLUE, surpassing non-expert human levels and reducing their usefulness. Although specialized benchmarks such as BEIR became retrieval standards, they lack holistic coverage across tasks like clustering or reranking (Muennighoff et al., 2023), leaving the evaluation landscape fragmented.

MTEB marked a pivotal step in unifying the previously fragmented evaluation landscape. It offers a comprehensive and standardized suite comprising 58 datasets across eight core task categories, establishing a more holistic standard for assessing the general performance of text embedding models. Despite its high coverage, MTEB remains predominantly English-centric, with limited dataset diversity for other languages. This gap inspired benchmarks tailored to specific linguistic characteristics similar to MTEB. Notable examples include efforts for Polish (Poświata et al., 2024), Persian (Zinvandi et al., 2025), French (Ciancone et al., 2024), Vietnamese (Pham et al., 2025), German (Wehrli et al., 2024), and Chinese (Xiao et al., 2024) languages. For instance, C-MTEB represented a major milestone for the Chinese language by consolidating and standardizing 35 public datasets to provide a comprehensive evaluation capability previously unavailable. These initiatives equip their NLP communities with tools for fairer, accurate model comparisons. In parallel, MTEB's large-scale and community-driven successor, the Massive Multilingual Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025), extended coverage to over 500 tasks in more than 250 languages. However, even wide-ranging multilingual benchmarks like MMTEB fall short for deep, language-specific evaluation like Turkish. Much of the Turkish data in these frameworks relies on unchecked machine translations, which fail to capture the language's morphological richness, cultural context, and regional nuances. Additionally, the scarcity of native Turkish datasets and the presence of mixed English–Turkish content hinder reliable performance measurement.

The evaluation landscape for the Turkish language remains fragmented and ad hoc. The scarcity of high-quality, curated datasets poses a major challenge, forcing researchers to create custom evaluation sets or rely on inconsistent methods. Early efforts in intrinsic evaluation underscored this gap, leading to studies that introduced Turkish datasets by manually translating English benchmarks such as SemEval or by generating resources addressing Turkish-specific problems like morphological relations (Agun and Yılmazel, 2020; Arslan et al., 2023). On the extrinsic side, models are often tested on disparate, task-specific datasets, hindering systematic comparison. For instance, EmbedTurk (Oytac et al., 2025) evaluated its model on a subset of MTEB tasks such as STS and retrieval, while another study (Ezerceli et al., 2025) focused on Turkish information retrieval and reported achieving state-of-the-art results, yet the absence of a unified evaluation protocol and comprehensive evaluation across core tasks continues to hinder progress in the field.

Recent concurrent work has introduced TR-MTEB (Baysan and Gungor, 2025), which establishes a Turkish embedding benchmark featuring 26 datasets and contributes newly trained Turkish embedding models. While TR-MTEB provides a highly valuable contribution to the field, our work TR-TEB, offers a more extensive and novel evaluation framework. A primary distinction lies in the overall scope; TR-TEB evaluates a significantly broader portfolio of 47 datasets, compared to the 26 datasets featured in TR-MTEB. Furthermore, for the machine-translated portions of the benchmark, TR-TEB utilizes the highly capable GPT-4o-mini model to ensure strict structural consistency and linguistic accuracy. In contrast, TR-MTEB relied on a smaller 8-billion parameter model (Aya-expanse) (Dang et al., 2024) to translate its datasets due to computational constraints. While both benchmarks utilize machine-translated data to expand task coverage, TR-TEB addresses the limitations and cultural biases of translated texts by curating 7 entirely novel, native Turkish datasets. These datasets were built from scratch to capture the unique morphological rules and cultural context of the language, an aspect that is harder to measure when relying primarily on translated resources. Crucially, TR-TEB tests much larger and more complex architectures, scaling up to 9.2 billion parameters with models like BGE-Multilingual-Gemma2 and Qwen3-Embedding-8B. By testing these larger, top-tier models and utilizing a more advanced LLM

for dataset curation, TR-TEB is able to accurately capture scaling laws and provide a more comprehensive evaluation standard for Turkish text embeddings.

3. Methodology

Benchmarks evaluate models across multiple aspects of their capabilities. A robust benchmark carefully designs its methodology and ensures dataset diversity to function as a general evaluation framework. Accordingly, TR-TEB standardizes the evaluation of text embedding models through a systematic approach encompassing task selection, metric design, and dataset curation. Tasks are selected to comprehensively capture monolingual evaluation, and appropriate metrics are defined for each. Identified issues in existing Turkish datasets are addressed by standardizing formats, curating new resources when necessary, and translating additional open datasets with an error-checking mechanism. To prevent data contamination, only the test subsets of open-source datasets are used for evaluation. The embedding models are selected based on their significance to the Turkish AI community to demonstrate their practical value and feasibility. All sentence embeddings are computed using the SentenceTransformers (Reimers and Gurevych, 2019) library.

3.1. Tasks

TR-TEB assesses performance across five types of tasks: classification, pair classification, clustering, retrieval, and STS. Each task is structured for a specific objective, accepts input in a predefined format, and produces discrete labels or continuous scores. A primary evaluation metric is assigned to each task to ensure results are easily comparable across models. This task and metric selection aligns well with established literature and other benchmarks such as MTEB (Muennighoff et al., 2023). Metrics other than the primary metric are also calculated; however, only the primary metric is taken as the decisive measure.

3.1.1. Classification

This task assesses the quality of sentence embeddings for single-text categorization. The evaluation is conducted using Supervised Probing with Logistic Regression, which is a form of lightweight supervised assessment. It measures how discriminative the embeddings are by using them as features to train a simple Logistic Regression model on the task's training split. This approach serves as an indicator of the general-purpose representational power of the embeddings. Accuracy is chosen as the primary performance metric, as it emphasizes the overall rate of correct predictions across all classes. This choice is especially suitable for benchmark datasets in which class distributions are treated as balanced or where maximizing total correctness is the main objective.

3.1.2. Pair Classification

This task assesses a model's ability to capture the binary relationship between two distinct texts: related or unrelated. It is crucial for paraphrase identification and Natural Language Inference (NLI) applications. Performance is evaluated using two distinct approaches to provide a comprehensive view of the embedding quality. First, logistic regression is used, where the embeddings of sentence pairs are concatenated into single feature vectors in the $[u, v, |u - v|]$ format, and a Logistic Regressor is trained on the task's training split. Concurrently, a zero-shot threshold accuracy method is employed, where we calculate the cosine similarity between the two sentence embeddings and compute the maximum accuracy achievable by selecting the optimal threshold on the development split. This zero-shot approach measures the intrinsic semantic distance captured by the embeddings without any task-specific training. The primary reported metric for this task is the threshold accuracy, as it directly reflects the model's intrinsic power to separate the two classes based purely on cosine similarity, which aligns with standard zero-shot embedding benchmarks.

3.1.3. Clustering

This task assesses the quality of embeddings in an unsupervised setting by evaluating how effectively a model can group semantically similar texts without relying on explicit label information. This is a critical test of a model's intrinsic ability to induce meaningful structure in the vector space, where the distance between embeddings should directly reflect the semantic similarity between texts. In the evaluation, MiniBatchKMeans (Sculley, 2010) finds k clusters, k being the number of ground-truth classes. Then V-Measure (Rosenberg and Hirschberg, 2007) takes the harmonic mean of homogeneity (where clusters contain only members of a single class) and completeness (where all members of a given class are assigned to the same cluster), thereby providing a balanced and robust assessment of clustering quality.

3.1.4. Retrieval

The retrieval task evaluates the zero-shot information retrieval capability of the models, which

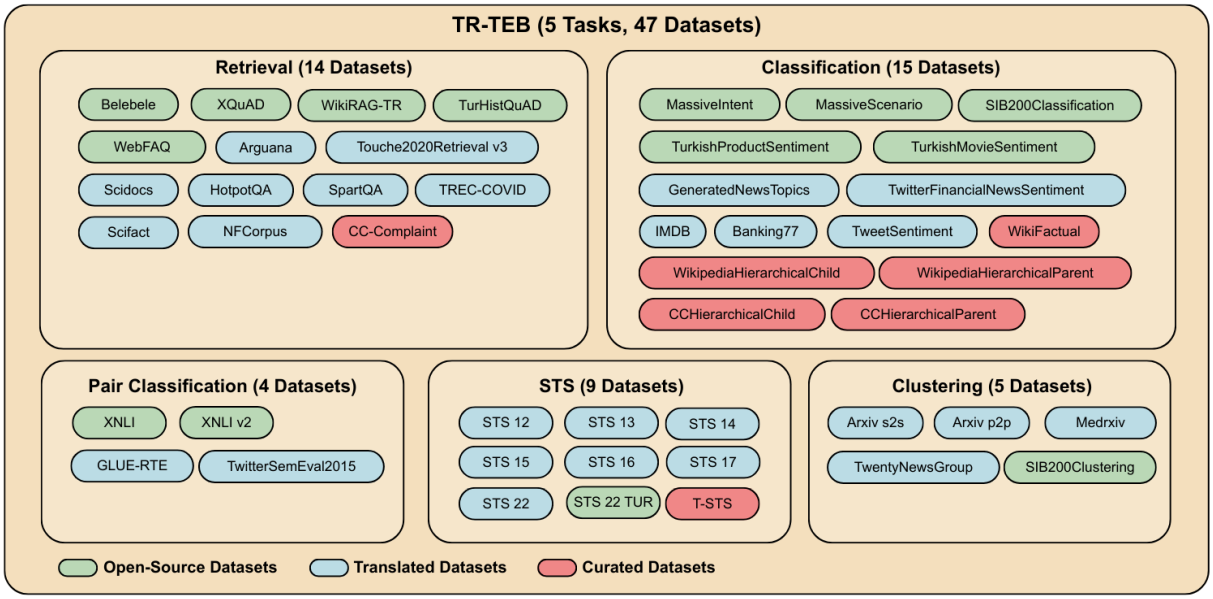

Figure 1: Overview of the TR-TEB evaluation framework with task distribution and dataset taxonomy.

is essential for applications such as search engines and question answering systems. Normalized Discounted Cumulative Gain at k=10 (nDCG@10)(Järvelin and Kekäläinen, 2002) is selected as the primary metric. It measures how accurately the model can rank a set of documents based on a given query, using only the cosine similarity between the query embedding and the document embeddings. It is a rank-aware metric that penalizes retrieving relevant documents at lower ranks more severely than at higher ranks, thus capturing the true performance of a search system where users typically only examine the top few results.

3.1.5. Semantic Textual Similarity (STS)

This task requires models to predict the continuous degree of semantic equivalence between two sentences, often scored on a scale from 0 to 5. This is the most fine-grained semantic task, measuring the model's ability to capture nuanced semantic relationships beyond simple categorical judgments. The evaluation is conducted by correlating the model's predicted similarity scores (derived from the cosine similarity of the sentence embeddings) against the human-annotated ground-truth scores. Spearman's Rank Correlation Coefficient (Spearman Correlation) is selected as the primary metric because it is a non-parametric statistic that measures the strength and direction of the monotonic relationship between the predicted and true scores. It is ideal as it focuses on whether the model correctly ranks the sentence pairs by similarity, rather than the absolute difference between the predicted and true scores.

3.2. Datasets

One of the fundamental steps in making TR-TEB a comprehensive benchmark is establishing a robust dataset portfolio that reflects the specific challenges of the Turkish language and still aligns with the globally recognized methodologies. TR-TEB datasets are selected based on the tasks and benchmarks commonly used in evaluations of state-of-the-art (SoTA) embedding models with high impact, such as BGE (Chen et al., 2024), Jina (Sturua et al., 2024), Qwen (Zhang et al., 2025), GTE (Zhang et al., 2024), E5 (Wang et al., 2024) and EmbeddingGemma (Vera et al., 2025). This approach ensures that the results are comparable with the literature. Prior to dataset inclusion, each dataset's license was carefully reviewed for compatibility with open research and redistribution. Datasets with unclear, restrictive, or non-permissive licenses were excluded from TR-TEB to maintain ethical and legal integrity. Afterwards, following dataset selection based on license considerations, the dataset portfolio is divided into three primary categories: open-source datasets, translated datasets, and curated datasets. A statistical breakdown of the datasets, including the number of instances and sentence lengths, is provided in Table 1. Also, the selection targets provide comprehensive coverage for each task. An overview of the dataset and task categorization is given in Figure 1.

3.2.1. Open Source Datasets

A Turkish subset of open-source MMTEB and Hugging Face datasets is referred to as the open-source datasets in this study. These datasets are mostly included as is since MMTEB accepts

Dataset Name	Number of Instances	Average Character Lengths	Translation Error Rate
Belebele (Bandarkar et al., 2024)	900 ; 488	71.56 ; 489.04	-
XQuad (Artetxe et al., 2019)	1,184 ; 240	60.88 ; 788.30	-
WebFAQ (Dinzinger et al., 2025)	10,000 ; 144,846	41.72 ; 254.48	-
TurHistQuAD (Soygazi et al., 2021)*	14,221 ; 2,149	66.87 ; 714.53	-
WikiRAG-TR (Usta, 2024)	5,725 ; 23,012	74.24 ; 699.16	-
Arguana (Wachsmuth et al., 2014)*	1,394 ; 1,394	1179.04 ; 426.25	0.64%
Touche2020Retrieval v3 (Thakur et al., 2024)*	49 ; 2,729	45.31 ; 2885.80	0.11%
Scidocs (Cohan et al., 2020)*	29,592; 29,128	75.01 ; 621.08	1.63%
HotpotQA (Yang et al., 2018)*	14,777 ; 14,777	92.15 ; 292.86	4,59%
SpartQA (Mirzaee et al., 2021)*	3,486 ; 489	612.46 ; 33.17	0.01%
TREC-COVID (Roberts et al., 2021)*	50 ; 64,756	69.68 ; 708.59	1.21%
Scifact (Wadden et al., 2020)*	1,107 ; 10,331	92.92 ; 766.94	0.33%
NFCorpus (Boteva et al., 2016)*	2,175 ; 3,556	25.58 ; 807.80	1.11%
CC-Complaint	44,098 ; 44,098	53.08 ; 561.96	-
MassiveIntent (FitzGerald et al., 2022)*	2,974	34.81	0.32%
Massive Scenario (FitzGerald et al., 2022)*	2,974	34.81	0.30%
SIB200Classification (Adelani et al., 2024)	701	134.51	-
TurkishProductSentiment (Ayhan, 2022)	4,800	247.75	-
TurkishMovieSentiment (Demirtas and Pechenizkiy, 2013)*	7,972	141.03	-
GeneratedNewsTopics (Diaz, 2024; Wang et al., 2023)	3,595	424.00	0%
TwitterFinancialNewSentiment (zeroshot, 2022)	9,543	85.82	10.93%
IMBD (Maas et al., 2011)*	24,904	1,326.45	0.09%
Banking77 (Casanueva et al., 2020)*	9,993	59.50	0.31%
TweetSentiment (Maggie, 2020)*	26,732	73.31	1.76%
WikiFactual	50,000	332.50	-
WikipediaHierarchicalParent	49,989	1,368.70	-
WikipediaHierarchicalChild	49,989	1,368.70	-
CCHierarchicalParent	4,903	5,031.06	-
CCHierarchicalChild	4,903	5,031.06	-
XNLI (Conneau et al., 2018)	1,365	76.43	-
XNLI v2 (Upadhyay and Upadhya, 2023)*	1,365	75.72	-
TwitterSamEval2015 (Xu et al., 2015)*	16,777	44.69	8.30%
GLUE-RTE (Wang et al., 2019)	2,490	271.45	0.33%
STS 12 (Agirre et al., 2012)	2,997	65.77	2.19%
STS 13 (Agirre et al., 2013)	1,472	58.76	1.27%
STS 14 (Agirre et al., 2014)	3,627	59.48	1.95%
STS 15 (Agirre et al., 2015)	2,959	59.57	0.77%
STS 16 (Agirre et al., 2016)	1,143	66.34	1.94%
STS 17 (Cer et al., 2017)	244	43.32	0.00%
STS 22 (Chen et al., 2022)	196	2703.80	1.14%
STS 22 TUR (Chen et al., 2022)	208	1979.98	-
T-STS	522	61.80	-
Arxiv s2s (arXiv.org submitters, 2024)*	726,569	74.16	0.53%
Arxiv p2p (arXiv.org submitters, 2024)*	732,423	1011.99	0.04%
Medrxiv*	17,606	114.98	0.21%
TwentyNewsGroup (Mitchell, 1997)	52,536	36.45	3.60%
SIB200Clustering (Adelani et al., 2024)	1,004	133.61	-

Table 1: Dataset properties. The statistics of the retrieval task datasets are given in pairs, respectively, as query and corpus. *Note.* * Indicates the dataset is also referenced to MTEB and MMTEB (Enevoldsen et al., 2025; Muennighoff et al., 2023).

datasets with clear task definitions, high-quality standards, and community validation, making it the most suitable foundation for a Turkish-specific benchmark. In the selection, MMTEB datasets already containing verified Turkish subsets are prioritized. The selection is grounded in MTEB's rigorous standards, robust evaluation metrics, balanced splits, broad linguistic/domain coverage, reproducibility, fairness, meaningful benchmarking across model scaling laws, and varied task difficulty. Since the Turkish subsets in multilingual datasets play a critical role in producing valid and reliable results, this study implements an additional layer of quality control. This involves carefully checking content appropriateness, linguistic accuracy, and alignment with the original dataset intent. Any low-quality, misaligned, or inconsistent subsets are either filtered out entirely or carefully re-worked, translated, or adjusted to maintain the integrity of the evaluation. Texts that meet strict quality criteria are incorporated directly into the dataset, ensuring the benchmark remains high-quality and representative of Turkish language and domain-specific characteristics. This approach ensures that evaluations accurately reflect model performance without noise or inconsistencies.

3.2.2. Translation

Similar to the open-source datasets, this subset is drawn from widely used open-source resources and MMTEB datasets. However, the language crite-

ria are not applied in this subset. The datasets are carefully selected based on task diversity and their high impact among SoTA embedding models, and then translated into Turkish. This step is crucial for incorporating resources from other languages into TR-TEB, and a total of 26 datasets were translated across all defined tasks to ensure high coverage. The translation process is performed using GPT-4o-mini (OpenAI, 2024) model and applied across all text fields excluding text formatted labels, resulting in approximately 300 million translated tokens. Multiple large language models (LLMs) with established multilingual capabilities are evaluated to identify the model that achieves the best performance while maintaining minimal computational time and cost overhead. All samples are translated while preserving their original structure, labels, and relational consistency. To ensure the quality and fidelity of the translated data, TR-TEB implements a dedicated Translation Quality Checker module. Each translation undergoes a multi-stage automated validation process. This process detects erroneous translations with respect to the defined six categories: word count error, wrong language error, unwanted character error, and translation artifact. The module flags translations that fail these checks for re-translation or human review, ensuring that the final datasets remain structurally consistent and linguistically accurate. The resulting Turkish datasets closely mirror their English counterparts, and the high quality of the translations ensures reliable evaluation. Further details on the module and translator model selection are provided in Appendix A.

3.2.3. Curated Datasets

Even though the available open source datasets, both taken as-is and translated, are standard and high quality, there are areas in which they are inadequate. Cultural differences and biases, the information gaps in Turkish texts and English texts create a need for language-specific, naturally occurring datasets that represent the targeted language better than other options. Certain datasets include examples of this type, which are collected specifically for the Turkish language as a subset. However, more examples are needed to have a comprehensive benchmark. To compensate for this, an attempt is made to create datasets for different tasks. In this part, available data sources, Wikipedia and Common Crawl (Crawl), are used as a seed mechanism to generate tailored datasets with custom methods, which resulted in seven unique datasets. For all of the datasets, a novel contribution is made, such as class generation or expert labeling, to ensure that the contamination rate for datasets, including the common training corpora, is minimized.

Turkish Wikipedia Factual Classification Dataset: The factual understanding of the language-specific information is determined as one of the improvement areas in the benchmark study. To test the model's factual knowledge, a binary classified dataset is formed with partial replacement of the named entity recognition (NER) components, which are taken from random Turkish Wikipedia article chunks. For each chunk, 23 different categories of components are manually chosen and then GliNER (Zaratiana et al., 2023) is used to detect and remove these components. A chosen percentage of the chunks' NER components are partially replaced with a random NER within the same component category. The replaced chunks are labeled as false, to indicate there is false information, and non-replaced as true. One of the potential drawbacks of this dataset is that some information in Wikipedia pages can be too specific, and hence, the maximum success rate might be lower than a full success rate.

Turkish Wikipedia Hierarchical Topic Classification Dataset: Another dataset is designed for topic classification based on the available topics that are present in the user-labeled categories in Wikipedia pages. Instead of deciding on the topic with the text itself, these user-defined categories are structured so that all of them are hierarchically dependent on one or more main topics. After a rigorous filtering of categories not relevant to the text, such as "Alive People", is done, the page title and the remained categories are given Gemma 3 12B (Team et al., 2025) LLM to classify the page into a predefined main and sub-categories. The predefined categories are decided from Wikipedia's own main and one level subcategories, which results in a total of 8 main and 117 sub-categories. Main and sub-category labels are treated as separate labels in the final dataset.

Turkish Common Crawl Hierarchical Topic Classification Dataset: A similar method is used to create a dataset of Turkish Common Crawl texts. Common Crawl datasets include a wide variety and a large number of documents across the web. Since these datasets are not labeled with any user-defined categories, a synthetic labeling process is applied to create main and sub-categories. The labeling process is done with Gemma 3 12B (Team et al., 2025) LLM. The prompt is designed so that the model remembers the past assigned parent and child categories and decides on if there is a need to create a new category for each new text or to use an existing category. As a result, a randomly sampled subset of Common Crawl documents is chunked and labeled with a parent and a related child category synthetically. It is expected that the

uniqueness of categorizations in the dataset will prevent any contamination effects, since Common Crawl is one of the main training data sources for most of the models in the masked language modeling (MLM) training phase.

Turkish-STS (T-STS): The scarcity of high-quality Turkish evaluation resources creates a need for a novel STS dataset. This curated dataset was created through a seed sentence selection process from a diverse Turkish web corpus. Initially, raw text is segmented into candidate sentences. Each candidate was then filtered for semantic integrity using a protocol that enforced strict constraints on word count (4-10), punctuation, and capitalization. Furthermore, heuristics were applied to filter out malformed or boilerplate content by checking character ratios and forbidden patterns. The final high-quality seeds were then sampled randomly across 21 distinct topics to ensure broad thematic coverage.

A hybrid generation strategy was employed for the dataset's creation. After a comparative evaluation of several models, GPT-4.1 (OpenAI, 2025) was selected for its ability to generate pairs with high linguistic quality and semantic accuracy in line with the target scores. For nuanced semantic relationships (scores 1-5), the model is prompted to generate a counterpart for a given seed sentence, based on a detailed rubric with annotated examples and diversity constraints. Conversely, dissimilar pairs (score 0) were generated deterministically by pairing sentences from mutually exclusive topics, ensuring semantic opposition.

The entire corpus was validated by at least three human experts per pair. An analysis was performed to measure the Pearson correlation between the model's generated STS score and the resulting human annotations. After a quality control phase to discard pairs with grammatical or semantic flaws, the final ground-truth score was determined by averaging the remaining human ratings. Even though the contents are generated, personally identifiable information (PII) still exists in the text if the seed text included it. The annotators removed any PII-related information in the process. This comprehensive process certifies the dataset's fidelity, establishing it as a reliable resource for model evaluation.

CommonCrawl Complaints Dataset: Most of the complaint sites are a great source of text pairs structured as the complaint title and the text of the complaint. To curate a novel retrieval dataset, the crawls of known Turkish complaint sites in Common Crawl are used to construct a dataset. The titles are marked as queries, and the complaint texts are taken as the corpus. Several checks are done, such as NER detection and regex matching to ensure PII does not appear in the text. One downside of this dataset is that the complaints are, in some cases, too similar to each other when the given context in the complaint is shallow. Since the dataset consists of a summary-like title and a main body, it evaluates models from a unique perspective that title-abstract datasets cannot capture.

3.3. Models

Embedding models are a derivation of the Transformer architecture for the specific use case of turning text into numerical representations. Since the appearance of BERT (Devlin et al., 2019) model family in the literature, encoder-only models have emerged as a usable architecture in several areas. These models use the encoder module in the architecture to extract features, which are then used for downstream tasks such as retrieving, searching, or classifying text. BERT models are trained with MLM to understand the token similarities. The model training steps usually follow a similar approach. First, the base embedding models are trained with MLM or similar methodologies with raw text from sources like Common Crawl. Then, to increase the semantic understanding capabilities of the model, a contrastive learning approach is used with pairs or triplets of text. This stage helps model in downstream tasks greatly and has been the main driver of trsentence embedding models. After this stage, models like BGE-M3 use a fine-tuning approach specific to downstream tasks. Additionally, architectures like Jina Embeddings v3 use adapters for each downstream task to increase the performance after the finetuning stage. Model's maximum context length, embedding dimensions and the total number of parameters are commonly used to represent the general model architecture and performance capability. For most applications, the maximum context length of models plays a crucial role, as it directly limits the length of the text that can be processed. Models usually have between 384 and 8192 tokens of context size, though this number increases as the model landscape develops. Embedding dimensions determine the final dimension of the representation vector. A new methodology called the Matryoshka Embedding (Kusupati et al., 2024) can introduce new loss components for lower embedding dimensions to be usable. For this benchmark study, base embedding models and sentence embedding models that are trained on Turkish data are used for comparison. Base embedding models, as expected, perform worse in these tasks than sentence embedding models. Even though embedding models are not LLMs, they still can include prompt engineering in their training cycles to route the model for downstream tasks, and are suggested to be used by the model providers. To be compatible with provider

Model	Params	Max Seq	Embed Dim	Mean (Task Type)	Classification	Clustering	STS	Retrieval	Pair Class.
Qwen3-Embedding-8B	7.6B	40960*	4096	**64.58**	80.86	37.74	76.30	**59.62**	68.39
Jina-Embeddings-v3	572M	8194	1024	64.18	80.17	36.76	**79.31**	57.90	66.73
BGE-m3	568M	8192	1024	63.76	80.15	30.24	77.08	59.30	72.04
Qwen3-Embedding-4B	4.0B	40960*	2560	63.72	80.13	36.83	75.53	57.88	68.22
BGE-Multilingual-Gemma2	9.2B	8192*	3584	63.02	**81.81**	**38.73**	76.76	48.24	69.56
GTE-multilingual-base	305M	8192	768	62.87	77.96	35.38	76.09	55.44	69.50
Multilingual-E5-Large	560M	512	1024	62.61	75.48	34.28	75.85	58.57	68.85
Snowflake-Arctic-Embed-L-v2.0	568M	8192	1024	61.06	78.52	30.48	74.18	56.78	65.34
Multilingual-E5-Base	278M	512	768	60.39	73.82	29.70	73.58	56.79	68.09
Jina-Colbert-v2	559M	8194	1024	59.64	78.20	27.25	76.82	41.71	**74.24**
Multilingual-E5-Small	118M	512	384	59.20	71.53	30.16	72.69	55.40	66.19
Qwen3-Embedding-0.6B	596M	32768**	1024	59.13	75.95	33.18	70.21	50.81	65.48
Granite-Embedding-278M-Multilingual	278M	512	768	56.36	72.68	28.75	67.95	49.87	62.52
YTU-CE-Cosmos-Turkish-Large-Bert-Cased	337M	1024	1024	45.95	67.70	21.74	54.08	23.70	62.54
XLM-Roberta-Large	560M	512	1024	33.02	40.35	13.49	42.47	8.54	60.24
XLM-Roberta-Base	278M	512	768	31.74	40.13	12.25	41.35	3.59	61.38

Table 2: Turkish Embedding Model Benchmark Results. Scores are averaged across all tasks in each category. Models are sorted by overall score. Entries marked (*) and (**) were evaluated with context size limitations of 4,096 and 8,192 tokens, respectively. Model names correspond to their Hugging Face identifiers.

implementations, task-specific prompts or adapter activations are done for each task separately in the benchmark runs. This ensures that each model uses the correct settings for model inference. In some instances, the prompts or adapters can also be separately defined for the embedding of queries and documents in the retrieval task. The benchmark currently lacks native support for ColBERT-type models, which produce multi-vector representations per input. As a representative of this model type, Jina-ColBERT V2 (Jha et al., 2024) is included in this study; however, the implementation relies on single-vector representations with cosine similarity rather than the model's intended multi-vector inference procedure. Because of the resource constraints three of largest model's (BGE-Multilingual-Gemma2, Qwen3-Embedding-4B and Qwen3-Embedding-8B) context sizes are limited to 4096 tokens.

4. Observations and Findings

To ensure TR-TEB provides a reliable evaluation, its curated dataset portfolio underwent a rigorous data and model-driven validation. The analysis confirms that the benchmark is sensitive to key model characteristics such as scale, context length, and task specialization, making it suitable for nuanced evaluation of Turkish text embedding models. The results are presented in Table 4. Extensive results that include other known models are included in Appendix B.

4.1. Correlation with Model Scale and Complexity

A fundamental indicator of a benchmark's validity is its ability to reflect performance gains from increased model complexity, a principle known as scaling laws (Kaplan et al., 2020). The TR-TEB results consistently demonstrate this correlation across different model families, confirming that the benchmark is sufficiently challenging to reward the enhanced representational capacity of larger models.

For instance, within the Qwen3 series, a clear scaling trend is evident. Qwen3-Embedding-4B (4B parameters) scores 63.72, and increasing the model size to the 7.6B-parameter Qwen3-Embedding-8B results in a higher score of 64.58. Both of these models significantly exceed the 0.6B variant (596M parameters, 59.13), creating a substantial gap of 4.59% between the largest and smallest models. A similar pattern emerges in the Multilingual-E5 family: the largest variant, Multilingual-E5-Large (560M parameters), achieves 62.61, slightly above Multilingual-E5-Base (278M parameters, 60.39), while both notably outperform Multilingual-E5-Small (118M parameters, 59.20). These hierarchical scores based on parameter count consistently reinforce the sensitivity of TR-TEB to model scale.

4.2. Sensitivity to Context Length and Architecture

The benchmark includes long-text datasets, STS22 (avg. 350 words), Touche 2020 Retrieval v3 (avg. 234 words), and IMDB Classification (avg. 234 words), to probe the context window size of embedding models. Models with extended contexts consistently outperform context-limited ones (512 tokens) on long-text datasets, regardless of model size, with smaller models gaining substantially and larger models benefiting from both scale and extended context size. Notably, BGE-m3 (568M parameters, 8K context) and Multilingual-E5-Large (560M parameters, 512 context) have similar scale but vastly different context windows, highlighting context size as one of the key performance factors. For another perspective on model design, Jina-Embeddings-v3 (64.18) and BGE-m3 (63.76) are

compared: both use XLM-RoBERTa-large (Conneau et al., 2020) with 8192 input length but differ in positional embeddings (Rotary vs. Absolute respectively). Their similar performance shows that the benchmark evaluates architectural differences and holistic quality (training data, methodology).

4.3. Task-Specific Differentiation

The benchmark results reveal that while Qwen3-Embedding-8B (Zhang et al., 2025) emerges as the top overall performer, no single model dominates across all task categories, highlighting distinct model specializations. Qwen3-Embedding-8B leads with an overall score of 64.58 and secures the highest score in Retrieval (59.62) indicating its robust embedding capabilities.

A clear performance trade-off is evident between symmetric and asymmetric tasks. For instance, Jina-Embeddings-v3 is the leader in the symmetric STS task (79.31) but performs moderately in asymmetric retrieval (57.90). This specialization is even more pronounced in Jina-ColBERT V2 (Jha et al., 2024), which excels at Pair Classification (74.24) while recording one of the weakest retrieval scores (41.71). However, it should be noted that the approximated single-vector implementation used for ColBERT-type models may partly account for these divergent results. Conversely, top retrieval models like Qwen3-Embedding-8B underperform on symmetric tasks. This persistent pattern suggests that architectural and training optimizations for symmetric similarity differ significantly from those for asymmetric search, underscoring that improving performance in one often compromises generalization to the other.

No single model dominates across all tasks, which shows the benchmark effectively evaluates different dimensions of embedding quality. This underscores the importance of selecting models based on their performance in the tasks that best reflect an application's nature, rather than relying solely on overall scores. These performance patterns reveal how model architectures and training objectives shape embeddings for particular downstream uses, further reinforcing the need for comprehensive, language-specific evaluation tools.

5. Conclusion

This study introduced TR-TEB, the first comprehensive benchmark designed to evaluate Turkish text embedding models across five fundamental NLP tasks. Through rigorous data and model-driven validation, findings confirm that TR-TEB reliably captures key model characteristics, including scaling behavior, architectural choices, and context length capabilities. Larger models consistently outperform smaller ones, demonstrating that the benchmark is sensitive to representational capacity, while long-text datasets effectively differentiate models based on context window size. Moreover, task-specific results reveal clear distinctions in model strengths and weaknesses, highlighting that no single model excels across all tasks and underscoring the benchmark's ability to probe complementary embedding capabilities. These insights establish TR-TEB as a robust, versatile tool for evaluating and comparing Turkish text embeddings, supporting both model development and applied NLP research. Beyond its immediate contributions, this benchmark represents a foundational step toward scalable, language-specific evaluation frameworks for low-resource languages, providing a template that can be extended to other linguistic contexts and encouraging more equitable progress in global NLP research.

6. Ethical Considerations, Bias and Limitations

Open-source and API-based large language models are used within this study to translate existing datasets and generate samples for benchmarking purposes. Due to the nature of these models, several ethical considerations must be made to address potential inherent biases. The models exhibit biases stemming from their training data and methodologies. Model providers attempt to mitigate these biases through dataset curation, output filtering, and alignment techniques such as RLHF; however, the scale and diversity of training sources makes this a challenging task. Furthermore, model providers do not share full details about their data sources, training methodologies, or safety precautions. Therefore, generated content may include biases that are difficult to detect using rule-based filtering methods. While detection by human experts is theoretically possible, it is infeasible given the large volume of samples.

Another implication of these biases in benchmarking is the potential advantage of models with similar architectures or training data distributions, although this effect may be limited given the methodological diversity among embedding models and LLMs. Translations were performed using a proprietary model, while this work evaluates only open-source models, which may reduce the risk of architectural overlap and associated bias. Data contamination also poses a risk, as the original untranslated datasets may have been included in the training data of the evaluated models. Since training runs sometimes incorporate publicly available test sets and providers may not fully disclose their training data sources, the extent of such contamination cannot be verified. Although curated

datasets were incorporated to mitigate these effects and better approximate the target distribution, these potential sources of bias should be considered when interpreting the benchmark results.

7. Bibliographical References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

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, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hayri Volkan Agun and Ömer Faruk Yılmazel. 2020. Intrinsic evaluation of word embeddings for turkish. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Oğuz Ali Arslan, Berfin Duman, Hakan Erdem, Can Günyel, Bike Sönmez, and Doğukan Arslan. 2023. Towards turkish word embeddings: An intrinsic evaluation. In *2023 8th International Conference on Computer Science and Engineering (UBMK)*, pages 564–568.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

arXiv.org submitters. 2024. arxiv dataset.

Batuhan Ayhan. 2022. turkish-sentiment-analysis-dataset. `https://huggingface.co/datasets/winvoker/turkish-sentiment-analysis-dataset`. Accessed: 2025-07-15.

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, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders.

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, Vancouver, Canada. 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.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *ACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Common Crawl. Common crawl corpus. `https://commoncrawl.org`. Accessed: 2025-02-10.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *wisdom*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Sara Han Diaz. 2024. text-classification-news-topics. `https://huggingface.co/datasets/sdiazlor/text-classification-news-topics`. Accessed: 2025-07-15.

Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. 2025. Webfaq: A multilingual collection of natural qamp;a datasets for dense retrieval.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan

Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark.

Ö. Ezerceli, G. Gümüşçekiçci, T. Erkoç, and B. Özenç. 2025. Turkembed4retrieval: Turkish embedding model for retrieval task. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *Information Processing & Management*, 39(1):361–384.

Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Wang, Nan Wang, and Han Xiao. 2024. Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 159–166, Miami, Florida, USA. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2024. Matryoshka representation learning.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Wei Chen Maggie, Phil Culliton. 2020. Tweet sentiment extraction.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

Tom Mitchell. 1997. Twenty Newsgroups. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5C323.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. OpenAI blog — introduces GPT-4o mini, a cost-effective AI model with enhanced reasoning and multimodal capabilities.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/.

Doga Oytac, Ahmet Emre Ergun, Tuğba Çelikten, and Aytuğ Onan. 2025. Embedturk: Leveraging large language models as text encoders for turkish language. In *Proceedings of the 7th International Conference on Intelligent and Fuzzy Systems (INFUS 2025)*, pages 593–600, İstanbul, Turkey. Springer.

Loc Pham, Tung Luu, Thu Vo, Minh Nguyen, and Viet Hoang. 2025. Vn-mteb: Vietnamese massive text embedding benchmark.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark.

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*. Association for Computational Linguistics.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of trec-covid.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178, Raleigh, North Carolina, USA. ACM.

Fatih Soygazi, Okan Çiftçi, Uğurcan Kök, and Soner Cengiz. 2021. Thquad: Turkish historic question answering dataset for reading comprehension. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 215–220.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic evaluation of neural retrieval models on the Touché 2020 argument retrieval subset of BEIR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Ankit Kumar Upadhyay and Harsit Kumar Upadhya. 2023. Xnli 2.0: Improving xnli dataset and performance on cross lingual understanding (xlu). In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE.

Metin Usta. 2024. Wikirag-tr. https://huggingface.co/datasets/Metin/WikiRAG-TR. Accessed: 2025-06-15.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghu-

ram 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, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. Embeddinggemma: Powerful and lightweight text representations.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *Computational Linguistics and Intelligent Text Processing*, pages 115–127, Berlin, Heidelberg. Springer Berlin Heidelberg.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2024. German text embedding clustering benchmark.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer.

zeroshot. 2022. twitter-financial-news-sentiment. https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment. Accessed: 2025-07-15.

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.

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.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. Famteb: Massive text embedding benchmark in persian language.

Original	GPT 4o-mini	Qwen 3 32B	Llama 3.1 8B	Gemma 3 27B	Aya Expanse 32B	Aya Expanse 8B
do i have enough gas in my car to make it to dallas	**Arabamda Dallas'a ulaşmak için yeterince benzin var mı?**	arabamda yeterli benzin var mı Dallas'a varabilmek için	**arabamda dallas'a varmak için yeterli benzin var mı?**	Arabamda Dallas'a gitmek için yeterli benzin var mı?	Arabamda Dallas'a gitmek için yeterli benzinim var mı?	Arabamın Dallas'a gitmek için yeterli benziyor mu?
can i order takeaway dinner from byron's	**Byron's'tan paket yemek siparişi verebilir miyim?**	byron'dan akşam yemeği alabilir miyim	Byron'dan takeaway akşam yemeği sipariş edebilir miyim?	Byron'dan paket akşam yemeği sipariş edebilir miyim?	Byron'dan yemek siparişi verebilir miyim?	Byron'dan yemek siparişi verebilir miyim?
show me the answer for this problem two hundred and forty six plus six hundred and fifty four	**Bu problemin cevabını göster: iki yüz kırk altı artı altı yüz elli dört.**	**bu sorunun cevabını göster, iki yüz kırk altı artı altı yüz elli dört**	iki yüz kırklı altı plus altı yüz elli dört	**Bu problemin cevabını göster: iki yüz kırk altı artı altı yüz elli dört**	iki yüz kırk altı artı altı yüz elli dört	246 + 654 = 900
i feel intimidated nervous and overwhelmed and i shake like a leaf	**Kendimi tehdit altında, gergin ve bunalmış hissediyorum ve bir yaprak gibi titriyorum.**	korkuyorum, sinirliyim ve bastırıcı hissediyorum ve yaprak gibi titriyorum	Korku duyuyorum, sinirleniyorum ve üzerime bir yük var. Öfkeye kapılmış gibi hissediyorum.	Kendimi sindirilmiş, gergin ve bunalmış hissediyorum ve yaprak gibi titriyorum.	Kendimi sindirilmiş, sinirli ve üstümden gelen bir baskı hissediyorum ve bir yaprak gibi titriyorum.	Kendimi tehdit altında, gergin ve aşırı yüklenmiş hissediyorum ve yaprak gibi titriyorum.
i thought maybe once i started running i would feel ok	**Koşmaya başladıktan sonra belki iyi hissederim diye düşündüm.**	koşmaya başladığım anda iyi hissedeceğimi düşünmüştüm	koşmaya başladığım zaman iyi hissedeceğimi düşünüyordum	Koşmaya başladığımda iyi hissedeceğimi düşündüm belki.	Belki koşmaya başladığım zaman kendimi iyi hissedeceğimi düşündüm	Koşmaya başladığım zaman iyi hissetmeye başlayacağımı düşündüm.

Table 3: Full Turkish translation outputs from each model. Bold notated cells denote higher translation quality with respect to other models.

8. Appendices

A. Translation Quality Checker and Model Selection

Translation quality checker evaluates translations using a set of predefined rules designed to detect common errors without relying on semantic analysis, which would incur higher computational cost. The evaluation consists of the following checks: (1) language detection to verify that the translation matches the intended target language (2) the presence of non-ASCII characters (3) the ratio of words in the translated text. If this ratio is below 0.5, the same threshold is recalculated after removing auxiliary words from the source text. This adjustment accounts for the agglutinative structure of Turkish and the higher use of auxiliary words in English (4) common artifact patterns introduced by LLMs despite prompt constraints.

A violation of any rule indicates low translation quality. Such cases are retried up to two times, after which the translation is discarded if it continues to fail.

Among the highest-performing multilingual models, six candidates of varying sizes were selected for evaluation as translation models. Based on these results, along with additional examples assessed through human review, GPT-4o mini demonstrated the strongest overall performance.

Although certain models achieved comparable or occasionally superior results on specific examples, GPT-4o mini consistently outperformed the others on average, maintaining a clear overall advantage. The comparison is reported in Table 3.

B. Extensive Benchmark Results

Beyond the results reported in the paper, popular Hugging Face models, including both embeddings and BERT-based models, are evaluated across all datasets. This provides a comprehensive outlook on the general performance of models on Turkish language. The complete benchmark results and a scatter plot comparing the model size and performance are presented in Table 4 and Figure 2.

Model	Parameters	Max Seq Length	Embedding Dim	Classification	Clustering	STS	Retrieval	Pair Classification	Mean (Task-Type)	Mean (Dataset)
Qwen3-Embedding-8B	7.6B	40960*	4096	80.87	37.75	76.30	59.62	68.39	64.58	68.01
jina-embeddings-v3	572M	8194	1024	80.17	36.76	79.31	57.90	66.73	64.18	67.61
bge-m3	568M	8192	1024	80.15	30.24	77.08	59.30	72.04	63.76	67.35
Qwen3-Embedding-4B	4.0B	40960*	2560	80.13	36.83	75.53	57.88	68.22	63.72	67.00
bge-multilingual-gemma2	9.2B	8192*	3584	81.81	38.73	76.76	48.24	69.56	63.02	65.22
gte-multilingual-base	305M	8192	768	77.96	35.38	76.09	55.44	69.50	62.87	65.64
multilingual-e5-large	560M	512	1024	75.48	34.28	75.85	58.57	68.85	62.61	65.57
snowflake-arctic-embed-l-v2.0	568M	8192	1024	78.52	30.48	74.18	56.78	65.34	61.06	64.98
multilingual-e5-base	278M	512	768	73.82	29.70	73.58	56.79	68.09	60.39	63.52
jina-colbert-v2	559M	8194	1024	78.20	27.25	76.82	41.71	74.24	59.64	61.31
multilingual-e5-small	118M	512	384	71.53	30.16	72.69	55.40	66.19	59.20	62.09
Qwen3-Embedding-0.6B	596M	32768**	1024	75.95	33.18	70.21	50.81	65.48	59.13	61.92
embeddingmagibu-152m	157M	2048	768	76.00	28.05	69.92	49.90	64.00	57.57	60.94
granite-embedding-278m-multilingual	278M	512	768	72.68	28.75	67.95	49.87	62.52	56.36	59.45
bert-base-turkish-cased-mean-nli-stsb-tr	111M	75	768	74.30	22.02	76.99	39.71	68.51	56.30	58.46
bge-large-en-v1.5	335M	512	1024	65.64	22.27	60.26	32.76	62.92	48.77	49.97
embeddinggemma-300m	303M	2048	768	65.58	16.64	46.24	51.53	62.41	48.48	52.22
bert-base-turkish-uncased	111M	512	768	69.34	22.27	57.87	26.66	64.51	48.13	49.01
bge-base-en-v1.5	109M	512	768	65.34	20.15	58.99	32.11	63.32	47.98	49.25
mxbai-embed-large-v1	335M	512	1024	65.75	21.73	57.39	31.87	62.68	47.89	49.11
bge-large-en	335M	512	1024	63.63	22.88	60.44	28.68	63.50	47.83	48.26
bge-small-en-v1.5	33M	512	384	65.68	18.76	59.75	31.81	62.98	47.80	49.24
bge-base-en	109M	512	768	62.99	21.35	59.04	27.47	63.15	46.80	47.24
bge-small-en	33M	512	384	63.62	18.21	60.41	28.68	62.73	46.73	47.69
bert-base-turkish-cased	111M	512	768	68.32	21.44	55.17	25.38	63.26	46.71	47.59
turkish-large-bert-cased	337M	1024	1024	67.70	21.74	54.08	23.70	62.54	45.95	46.66
mxbai-embed-2d-large-v1	335M	512	1024	64.07	21.78	51.72	26.29	60.47	44.87	45.65
modernbert-base-tr-uncased	135M	8192	768	59.68	21.20	56.83	22.40	63.85	44.79	44.29
bert-base-multilingual-uncased	167M	512	768	66.25	21.02	51.55	19.03	61.16	43.80	44.13
mmBERT-base	307M	8192	768	65.35	21.54	52.37	13.62	61.70	42.92	42.48
turkish-base-bert-uncased	111M	512	768	67.33	16.03	48.76	16.60	62.43	42.23	42.79
mxbai-embed-xsmall-v1	24M	512	384	63.08	15.69	50.12	17.60	61.50	41.60	41.88
bert-base-multilingual-cased	178M	512	768	65.03	18.71	45.98	15.23	60.85	41.16	41.27
all-MiniLM-L6-v2	23M	256	384	62.64	15.91	49.40	16.36	61.18	41.10	41.22
turkish-mini-bert-uncased	12M	512	256	63.64	13.72	51.92	10.73	63.99	40.80	40.35
turkish-small-bert-uncased	30M	512	512	62.60	14.53	51.56	10.49	63.66	40.57	39.94
turkish-tiny-bert-uncased	5M	512	128	62.19	13.12	53.31	10.50	62.37	40.30	39.89
turkish-medium-bert-uncased	42M	512	512	62.47	16.25	49.08	11.67	61.94	40.28	39.81
TabiBERT	149M	8192	768	58.55	16.85	49.61	13.14	62.73	40.18	39.23
mmBERT-small	140M	8192	384	52.12	13.65	47.61	11.44	60.59	37.08	35.77
xlm-roberta-large	560M	512	1024	40.35	13.49	42.47	8.54	60.24	33.02	30.11
ModernBERT-large	395M	8192	1024	50.53	8.12	37.29	1.60	62.08	31.92	29.89
xlm-roberta-base	278M	512	768	40.13	12.25	41.35	3.59	61.38	31.74	28.32
ModernBERT-base	149M	8192	768	48.22	7.68	37.83	1.97	59.75	31.09	29.12
turkish-sentiment-modern-bert	149M	8192	768	38.43	3.31	23.90	0.90	60.10	25.33	22.58
roberta-large-mnli	355M	512	1024	38.59	3.63	20.18	1.04	59.71	24.63	21.96

Table 4: Turkish Embedding Model Benchmark Results for all selected models. Models are sorted by Mean (Task) scores. Entries marked (*) and (**) were evaluated with context size limitations of 4,096 and 8,192 tokens, respectively. Model names correspond to their Hugging Face identifiers.

Figure 2: Model size and benchmark performance comparison for all benchmarked models.