JMTEB and JMTEB-lite: Japanese Massive Text Embedding Benchmark and Its Lightweight Version

Shengzhe Li[1,2] **Masaya Ohagi**[1] **Ryokan Ri**[1] **Akihiko Fukuchi**[1]
Tomohide Shibata[1] **Daisuke Kawahara**[2]
[1] SB Intuitions Corp. [2] Waseda University
shengzhe.li@{sbintuitions.co.jp, asagi.waseda.jp}, dkw@waseda.jp
{akihiko.fukuchi, tomohide.shibata}@sbintuitions.co.jp

Abstract

We present JMTEB, a large-scale evaluation suite for Japanese text embedding models, designed to provide comprehensive coverage across multiple task types. The benchmark integrates 28 datasets across 5 tasks, enabling broad and challenging evaluation of model performance in diverse scenarios. While the full benchmark delivers thorough assessment, its scale poses practical challenges in terms of computation time and resource requirements. To address this, we construct JMTEB-lite, a lightweight version of JMTEB, by substantially reducing corpus size in retrieval-related tasks. JMTEB-lite significantly accelerates evaluation while maintaining high fidelity to the full benchmark. Together, JMTEB and JMTEB-lite form a flexible evaluation framework: the full version serves as a comprehensive standard for exhaustive benchmarking, while the lightweight version enables rapid iteration and efficient model selection. This dual approach facilitates both rigorous evaluation and practical development workflows, supporting the advancement of Japanese text embedding research.

Keywords: Japanese, embeddings, benchmark

1. Introduction

The advent of dense vector representations has fundamentally reshaped the field of modern natural language processing (NLP). This paradigm, which began with foundational work on word embeddings (Mikolov et al., 2013) and has been supercharged by the power of deep contextual models based on the Transformer architecture (Vaswani et al., 2017), has largely replaced traditional sparse, keyword-based methods. These dense embeddings capture the rich semantic and syntactic nuances of language, enabling a more profound level of text understanding. Today, they serve as the critical backbone for a vast array of advanced applications, from enhancing the factual grounding of large language models through retrieval-augmented generation (RAG) (Lewis et al., 2020) to powering sophisticated semantic search engines, document clustering systems, and classification models. The performance and reliability of these downstream systems are inextricably linked to the quality of their underlying text embeddings, making their rigorous and standardized evaluation a paramount concern for both researchers and practitioners.

The development of standardized evaluation suites, such as the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) for English and its adaptations for multilingual (MMTEB) (Enevoldsen et al., 2025) and Chinese (C-MTEB) (Xiao et al., 2023) contexts, has been instrumental in advancing text embedding research. Despite this global trend, the Japanese NLP community has lacked a comparable, unified benchmark for evaluating language-specific models. As a result, researchers have often evaluated models on disparate, custom-tailored tasks and datasets. This absence of a common "proving ground" makes direct model comparisons difficult, hinders reproducible scientific progress, and leaves practitioners without clear guidance on selecting the optimal model for their specific needs. This situation highlights a critical need for a standardized evaluation framework within the Japanese NLP community.

To address this critical gap in the Japanese NLP ecosystem and establish a much-needed standard, we introduce **JMTEB**: **J**apanese **M**assive **T**ext **E**mbedding **B**enchmark. It is a comprehensive evaluation suite designed from the ground up to serve as a definitive and lasting resource for the community. JMTEB integrates 28 diverse datasets carefully selected across 5 core task families: Retrieval, Reranking, Classification, Clustering, and Semantic Textual Similarity (STS). Each task family probes a different facet of an embedding model's capabilities, from large-scale information seeking (Retrieval) to fine-grained semantic comparison (STS). The datasets have been meticulously reviewed for linguistic quality and contextual relevance, spanning a wide spectrum of real-world domains. We built JMTEB to be a high-quality, comprehensive resource that can serve as a standard foundation for Japanese text embedding evaluation. Notably, a part of JMTEB has already been incorporated into the global MMTEB (Enevoldsen et al., 2025) benchmark, establishing it as the de facto standard for Japanese.

While the comprehensive nature of JMTEB es-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7423–7434
11–16 May 2026.

tablishes a new gold standard, we also recognize that the scale required for such rigor can present practical challenges. The computational cost and time needed to run a large-scale benchmark can create a bottleneck in the typical model development lifecycle, which involves frequent experimentation, ablation studies, and hyperparameter tuning. To address this, we also present **JMTEB-lite**, a lightweight and practical version of the benchmark. JMTEB-lite is constructed by strategically reducing the corpus size in the most resource-intensive Retrieval and Reranking tasks, while carefully preserving the original query-document relevance pairs to maintain the integrity of the evaluation. As we will demonstrate empirically, JMTEB-lite significantly accelerates evaluation while maintaining a high rank correlation with the full benchmark, making it a reliable and efficient resource for agile development.

Our main contributions in this paper are threefold:

1. **A unified standard**: We introduce JMTEB, the first large-scale, unified benchmark for Japanese text embeddings, which addresses a long-standing fragmentation in community evaluation practices.
2. **A practical development resource**: We create and validate JMTEB-lite, a lightweight version that lowers the computational barrier to entry and promotes agile, iterative model development without compromising evaluation integrity.
3. **Extensive analysis of Japanese models**: We conduct a comprehensive empirical study of 40 Japanese text embedding models, establishing extensive baselines that map the current performance landscape and provide crucial insights for future research.

JMTEB[1] and JMTEB-lite[2] have been made publicly available at HuggingFace.

2. Related Work

Text embedding benchmarks have gained attention as text embedding models evolve. BEIR (**BE**nchmarking **IR**) (Thakur et al., 2021) is a heterogeneous, zero-shot benchmark focusing on retrieval tasks, enabling robust comparisons of dense and sparse approaches. In order to cover the possible applications of embedding models as much as possible, MTEB (**M**assive **T**ext **E**mbedding **B**enchmark) (Muennighoff et al., 2023) integrates a variety of different tasks, including retrieval, reranking, STS, classification, etc. MMTEB (**M**ultilingual MTEB) (Enevoldsen et al., 2025) considerably expands MTEB to make it the largest multilingual benchmark for embeddings by covering over 500 datasets and over 250 languages. MMTEB also adds more settings such as instruction following, long-document retrieval and code retrieval. In parallel, language- and domain-specific derivatives have emerged to adapt text embedding evaluation for more specialized scenarios. For example, C-MTEB (Xiao et al., 2023) specializes in Chinese text embeddings, and FinMTEB (Tang and Yang, 2025) focuses on the financial domain. These benchmarks indicate the needs for more comprehensive as well as specific evaluation of text embedding models.

3. Construction of JMTEB

3.1. Motivation

Text embeddings underpin retrieval, reranking, clustering, and semantic similarity across languages and domains. Yet there has not been a consolidated Japanese benchmark that rigorously evaluates Japanese embeddings under a consistent protocol. Japanese presents distinctive orthographic and morphological properties (e.g., lack of whitespace, script mixing) that warrant dedicated datasets and task setups rather than relying solely on translations or language-agnostic surrogates (Kurihara et al., 2022).

We introduce **JMTEB** (**J**apanese **M**assive Text **E**mbedding **B**enchmark), a Japanese-specific benchmark designed for comprehensive evaluation of Japanese text embedding models. As of Sept. 2025, JMTEB consists of 28 datasets across 5 task families—Retrieval, Reranking, Semantic Textual Similarity (STS), Classification and Clustering. Most datasets are sourced from public resources (and, where appropriate, adopted from prior benchmarks), while some subset are original to JMTEB. Each dataset was carefully reviewed for linguistic quality and suitability to its evaluation task by inspecting representative samples. In constructing JMTEB, we follow widely used design desiderata—diversity, simplicity, extensibility, and reproducibility—as articulated in prior benchmark work (Muennighoff et al., 2023). To ensure the benchmark's simplicity and reproducibility, in the current stage, we focus on datasets that are publicly accessible for automated downloading. Consequently, we have not yet included several high-quality datasets governed by strict access agreements (e.g., requiring individual user applications or manual approval processes). In what follows, we introduce the task composition and dataset choices, detail the diversity of domains and text formats, and report results for 40 publicly available text embedding models, including both Japanese-native and Japanese-capable multilingual systems.

[1] https://huggingface.co/datasets/sbintuitions/JMTEB

[2] https://huggingface.co/datasets/sbintuitions/JMTEB-lite

3.2. Tasks and Datasets

JMTEB follows the MTEB taxonomy and evaluates Japanese text embedding models across 5 tasks: Clustering, Classification, Semantic Textual Similarity (STS), Retrieval and Reranking. Notably, some tasks in MTEB/MMTEB are not included in JMTEB. Bitext mining is one such task: it matches two versions of the same sentence in different languages. Since this task is not suitable for evaluating monolingual Japanese text embeddings, it is not included in JMTEB.

1. **Clustering** aims to correctly assign texts with similar semantics/topic to the same cluster. The task aims to probe how well embeddings induce coherent topical structure without supervision. Specifically, we include `Livedoor News` dataset and the Japanese split of `MewsC-16` dataset (Nishikawa et al., 2022), as well as the Japanese split of `SIB-200` dataset (Adelani et al., 2023).
2. **Classification** aims to predict the correct category of the text only with its embedding. Given labeled texts, we convert each text to an embedding, and its embedding representation serves as its feature vector. These vectors will not be updated in the following training. Then we train a logistic regression classifier to test how well the vector represents the text. There are 7 datasets for the classification task, including `WRIME` (Suzuki et al., 2022), `Japanese Sentiment Classification` (Mollanorozy et al., 2023) and the Japanese split of `Amazon Review` (McAuley and Leskovec, 2013), `Amazon Counterfactual` (O'Neill et al., 2021), `Massive Intent`, `Massive Scenario` (FitzGerald et al., 2022), `SIB-200` (Adelani et al., 2023).
3. **Semantic Textual Similarity (STS)** predicts the semantic similarity between two sentences by computing the similarity between their embedding vectors, and correlations are computed between the prediction and the annotated similarity. `JSTS` (Kurihara et al., 2022) and `JSICK` (Yanaka and Mineshima, 2022) are two datasets for STS.
4. **Retrieval** aims to find the most relevant document with the query from the corpus, by computing embedding similarities. There are 11 datasets for the retrieval task, including the Japanese split of `Mr.TyDi` (Zhang et al., 2021), `MIRACL` (Zhang et al., 2023), `MLDR` (Chen et al., 2024) and `Mintaka` (Sen et al., 2022), as well as `JAQKET` (Suzuki et al., 2020), `JaCWIR` (Tateno, 2024a) and `JaGovFaqs-22k`. Inspired by the arXiv-related datasets in MTEB/MMTEB, we created 4 original datasets based on the Japanese NLP Journal LaTeX Corpus. The corpus consists of accepted articles in Japanese by the Journal of Natural Language Processing. We extracted titles, abstracts and introductions from the full articles. The goal is to find the corresponding abstract with the given title (`NLP Journal title-abs`), introduction with the given title (`NLP Journal title-intro`), introduction or full article with the given abstract (`NLP Journal abs-intro` and `NLP Journal abs-article`), through the similarities computed with text embeddings.
5. **Reranking** aims to rerank the retrieved documents according to their relevance to the query through computing embedding similarities. There are 5 datasets collected for reranking, including the Japanese split of `ESCI` (Reddy et al., 2022), `MIRACL` (Zhang et al., 2023) and `MLDR` (Chen et al., 2024), as well as `JaCWIR` (Tateno, 2024a) and `JQaRA` (Tateno, 2024b).

For each task, we set a main metric for unified comparisons: macro-F1 for Classification, v-measure (Rosenberg and Hirschberg, 2007) for Clustering, Spearman correlation for STS, and nDCG@10 for Retrieval and Reranking.

In the construction of JMTEB, we ensure that each dataset meets the basic criterion: linguistic fluency and suitability for an evaluation task. Above this, we emphasize the diversity that ensures the benchmark to be comprehensive, including domain coverage and text-format coverage. Regarding domains, the suite spans 8 types of domains: encyclopedic, news, FAQs, e-commerce, academia, blog and SNS posts, conversational commands, as well as image caption texts. Regarding text formats, it covers phrases, short queries/utterances, single sentences, paragraph-level passages, and long documents.

3.3. Evaluation of Embedding Models with JMTEB

3.3.1. Evaluation Settings

We collect 40 publicly available models, and evaluate them with JMTEB. These models include Japanese-native models that are publicly available, OpenAI's embedding APIs, as well as some Japanese-capable multilingual models that have the most downloads at HuggingFace. In practice, each model may require its own specific configuration such as prefixes and instructions. We document these configurations and adhere to them meticulously during the evaluation process to ensure that each model achieves its optimal performance. The results ranked by average score are shown in Tab. 1.

3.3.2. Evaluated Models

We analyze the models along two axes: model structure and training recipe, since they are empiri-

Model	Backbone	Training Recipe	#Params (M)	Avg.	Retrieval	STS	Classification	Reranking	Clustering
sbintuitions/sarashina-embedding-v2-1b	LLM	Weakly-sup	1,224	**76.38**	**76.48**	**84.22**	77.14	**86.28**	52.56
cl-nagoya/ruri-v3-310m	BERT	Weakly-sup	314	75.85	76.03	81.59	**77.65**	85.84	50.52
cl-nagoya/ruri-v3-130m	BERT	Weakly-sup	132	75.52	76.45	81.05	75.65	85.71	51.13
sbintuitions/sarashina-embedding-v1-1b	LLM	Weakly-sup	1,224	74.87	74.53	81.71	77.20	84.36	50.30
pfnet/plamo-embedding-1b	LLM	Weakly-sup	1,051	74.85	75.14	83.15	77.29	85.05	52.50
cl-nagoya/ruri-v3-70m	BERT	Weakly-sup	70	73.95	74.23	80.96	74.45	84.21	49.95
OpenAI/text-embedding-3-large	-	-	-	73.86	71.95	82.52	77.27	83.06	51.82
cl-nagoya/ruri-large-v2	BERT	Weakly-sup	337	73.63	71.87	83.18	76.10	83.89	50.88
cl-nagoya/ruri-v3-30m	BERT	Weakly-sup	37	72.95	72.84	81.78	73.35	82.93	49.90
BAAI/bge-m3	BERT	Weakly-sup	568	72.46	73.70	79.74	74.10	84.10	45.56
cl-nagoya/ruri-large	BERT	Weakly-sup	337	71.69	68.30	83.13	76.25	81.26	49.93
cl-nagoya/ruri-base-v2	BERT	Weakly-sup	111	71.66	68.96	83.03	75.59	82.46	46.84
cl-nagoya/ruri-small-v2	BERT	Weakly-sup	68	71.40	68.46	82.91	74.12	82.30	49.97
pkshatech/GLuCoSE-base-ja-v2	BERT	Weakly-sup	133	71.11	68.45	82.95	73.52	82.63	48.19
intfloat/multilingual-e5-large	BERT	Weakly-sup	560	70.67	67.65	80.86	72.30	83.01	50.58
google/embeddinggemma-300m	BERT	Weakly-sup	303	70.59	66.18	82.74	76.14	80.93	49.48
cl-nagoya/ruri-base	BERT	Weakly-sup	111	70.25	65.90	82.88	75.34	80.31	49.10
pkshatech/RoSEtta-base-ja	BERT	Weakly-sup	190	69.58	67.52	81.39	71.70	81.25	44.88
cl-nagoya/ruri-small	BERT	Weakly-sup	68	69.34	63.95	82.79	74.83	79.98	49.59
intfloat/multilingual-e5-base	BERT	Weakly-sup	278	68.06	64.48	80.46	69.70	79.46	50.12
intfloat/multilingual-e5-small	BERT	Weakly-sup	118	67.38	63.91	80.46	67.77	80.09	49.29
OpenAI/text-embedding-3-small	-	-	-	67.10	61.79	79.46	72.43	77.29	48.91
OpenAI/text-embedding-ada-002	-	-	-	65.13	59.58	79.02	69.39	75.63	48.78
hotchpotch/static-embedding-japanese	Non-cont.	Weakly-sup	-	63.80	60.51	80.16	66.73	77.09	35.91
pkshatech/GLuCoSE-base-ja	BERT	Weakly-sup	133	63.79	54.58	78.68	75.02	72.37	47.12
cl-nagoya/sup-simcse-ja-base	BERT	Sup.	111	59.91	45.00	82.05	72.72	70.36	**52.57**
MU-Kindai/Japanese-SimCSE-BERT-large-unsup	BERT	Unsup.	337	57.60	42.41	79.00	71.83	71.88	42.02
oshizo/sbert-jsnli-luke-japanese-base-lite	BERT	Sup.	133	56.75	38.08	76.56	74.53	69.81	48.75
cl-nagoya/sup-simcse-ja-large	BERT	Sup.	337	56.46	37.38	83.17	72.74	68.76	50.12
MU-Kindai/Japanese-SimCSE-BERT-base-unsup	BERT	Unsup.	111	55.78	39.85	77.96	71.46	69.92	39.27
MU-Kindai/Japanese-SimCSE-BERT-large-sup	BERT	Sup.	337	55.35	36.23	78.29	72.59	70.59	44.54
MU-Kindai/Japanese-MixCSE-BERT-base	BERT	Unsup.	111	54.65	36.24	77.75	71.81	68.58	43.45
cl-nagoya/unsup-simcse-ja-large	BERT	Unsup.	337	54.23	33.98	80.56	73.71	67.39	43.52
cl-nagoya/unsup-simcse-ja-base	BERT	Unsup.	111	53.86	35.34	78.74	72.41	66.20	41.29
MU-Kindai/Japanese-SimCSE-BERT-base-sup	BERT	Sup.	111	53.82	35.22	74.96	71.48	68.15	42.86
MU-Kindai/Japanese-DiffCSE-BERT-base	BERT	Unsup.	111	53.59	34.93	76.70	72.06	67.73	39.93
pkshatech/simcse-ja-bert-base-clcmlp	BERT	Sup.	110	53.48	32.80	76.81	70.67	68.02	49.45
sentence-transformers/LaBSE	BERT	Sup.	470	52.70	33.18	76.56	71.85	67.01	39.82
sentence-transformers/stsb-xlm-r-multilingual	BERT	Sup.	278	43.06	16.58	75.41	71.40	57.93	27.67
colorfulscoop/sbert-base-ja	BERT	Sup.	111	42.90	15.45	70.41	68.05	59.38	39.04

Table 1: Evaluation results of 40 public models, ranked by the Avg. score. The score of each task is the average of all datasets in this task, and the average score is the average of all 28 datasets. The evaluations were conducted on Sept. 30, 2025.

cally considered as the most intuitive and important factors that could influence model performance. In Tab. 1, we mark each model with corresponding notations.

According to model structure, we group the models to the following categories. (1) **LLM-based** models use pretrained *decoder-style* LLMs as backbones, ensuring larger parameter counts (often $\geq$1B) and longer context windows (Wang et al., 2023). (2) **BERT-based** build on bidirectional transformer encoders such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2019), LUKE (Yamada et al., 2020) and ModernBERT (Warner et al., 2024), and produce embeddings via CLS/mean/last-token pooling. (3) **Static/Non-contextual models** typically generate a sentence embedding by summing or averaging its constituent token embeddings.

According to training recipe, we group the models into the following categories. (1) **Supervised** models are trained primarily with labeled objectives such as natural language inference (NLI) and semantic textual similarity (STS) regression/classification. (2) **Unsupervised** contrastive recipes avoid human labels, including SimCSE-unsupervised (Gao et al., 2021), DiffCSE (Chuang et al., 2022) and MixCSE (Zhang et al., 2022). (3) **Multi-stage learning with a weakly supervised pretraining stage** begins with large-scale weakly supervised pretraining using *weakly labeled* text-pair corpora and a contrastive objective, and is typically followed by a supervised finetuning stage. This has become a de facto standard (Tsukagoshi and Sasano, 2024; Li et al., 2023; Xiao et al., 2023) in the training of strong text embedding models, because it significantly improves retrieval robustness and generalization (Wang et al., 2022).

3.3.3. Analysis on Benchmark Quality

Difficulty The average model scores, which range from 40 to 80 (Tab. 1), confirm that the benchmark is appropriately difficult. This range indicates a well-

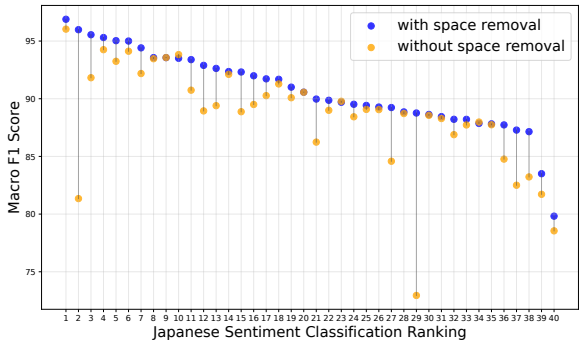

Figure 1: The scores of all 40 models on `Japanese Sentiment Classification` with/without space removal.

Query	ボルガライス、トルコライス、エスカロップといった料理に共通して登場する揚げ物は何でしょう? (What fried food commonly appears in dishes such as Volga Rice, Turkish Rice, and Escalope?)
Relevant	豚カツ（とんかつ）は、厚みのある豚のロースやヒレのスライス肉を、小麦粉・溶き卵・パン粉をまとわせて食用油で揚げた料理である...地方料理も多く、かつめし、エスカロップ、味噌カツ、トンカツラーメン、トルコライス、ボルガライスなど、また地元のブランド豚を使用するなど町おこしの一環としても利用されている。 (Tonkatsu is a dish made by coating thick slices of pork loin or fillet with flour, beaten egg, and breadcrumbs, then deep-frying them in cooking oil...There are also many local dishes, such as katsumeshi, escalope, miso katsu, tonkotsu ramen, Turkish rice, and Volga rice. In addition, local brand pork is used, and these dishes are also utilized as part of regional revitalization efforts.)
Pred. 1	カキフライとは、カキを材料とする日本の揚げ物料理（フライ）の一種... (Kaki Fry is a type of Japanese fried dish (fry) made with oysters....)
Pred. 2	フライとは、おもに魚貝類や野菜などの食材に卵白やパン粉をつけて、多量の食用油で揚げたもの... (Fry refers to food—mainly seafood or vegetables—coated with egg white and breadcrumbs, then deep-fried in a large amount of cooking oil...)
Pred. 3	バターライスは、米とバターを使った料理。米をバターで炒めてから炊く、米を炊く際にバターを一緒に炊き込む、炊き上がった飯をバターで炒める、飯にバターを混ぜ込むといった調理法がある... (Butter rice is a dish made with rice and butter. Cooking methods include sautéing the rice in butter before boiling it, cooking the rice together with butter, stir-frying the cooked rice with butter, or mixing butter into the cooked rice....)

Table 2: An example of `JAQKET` where even the top model in Tab. 1 failed to retrieve the document most relevant to the query.

calibrated challenge: the tasks are difficult enough that no model achieves a perfect score (avoiding ceiling effects), yet not so difficult that models cannot demonstrate meaningful performance (avoiding floor effects). Therefore, the benchmark is effective for distinguishing between model capabilities.

Linguistic fluency Some Japanese datasets in existing benchmarks in MTEB exhibit linguistic artifacts, such as unnatural phrasing or processing-induced noise. When constructing JMTEB, we addressed this by implementing a manual checking procedure during dataset selection. The impact of this procedure is evident in the `Japanese Sentiment Classification` dataset that originates from the `Multilingual Sentiment Classification` dataset in MTEB. The original MTEB version of this dataset contains artificial spaces within the Japanese texts, a common artifact of morphological preprocessing. We compare the performances of a panel of models on the MTEB's noisy version and JMTEB's version of this dataset. The results are shown in Fig. 1. In over half of the models, significant differences have been observed, as models tend to perform better on the JMTEB's version than on the MTEB's version, though they share the same texts. This performance gap indicates that the cleaned data better aligns with the distribution of natural Japanese texts that embedding models are intended to handle, while flawed texts cause evaluation distortion. This case demonstrates how our manual checking efforts in JMTEB facilitate a more reliable and valid evaluation of model performance on authentic Japanese.

Error analysis Tab. 2 presents an example from `JAQKET`. Even the top model in Tab. 1, `sbintuitions/sarashina-embedding-v2-1b` failed to retrieve the relevant document within its top-10 results (predictions beyond the top-3 omitted in Tab. 1). Real-world documents contain not only information relevant to a query but also a large amount of irrelevant information. This example shows that even state-of-the-art embedding models can fail to retrieve the relevant information, as they are misled by such irrelevant contents. Therefore, the benchmark remains challenging despite the rapid progress of recent text embedding technologies.

3.3.4. Analysis on Model Performance

Backbone model Within the same family (e.g., `ruri-v3` (Tsukagoshi and Sasano, 2024), `multilingual-e5` (Wang et al., 2024)), larger parameter sizes generally correlate with better performance. LLM-based models such as the `sarashina-embedding` series rank near the top, plausibly benefiting from scale. Notably, `ruri-v3-70m` (ModernBERT-based (Warner et al., 2024)) achieves an average score of 73.95, slightly surpassing `ruri-large-v2` (73.63) while being more than 4 times smaller in parameters. Although training data and recipes differ, this comparison highlights the efficiency gains ModernBERT brings.

Training recipe Nearly all models in the top half[3] of Tab. 1 adopt weakly supervised contrastive pretraining as a first stage, making it a de facto standard for high-performing embeddings. A time-series analysis of model performance reveals key trends in the evolution of Japanese text embeddings. Fig. 2 plots the performance of models from Tab. 1 according to their release dates. A striking

[3]OpenAI text embedding models are not considered here as their training details are disclosed.

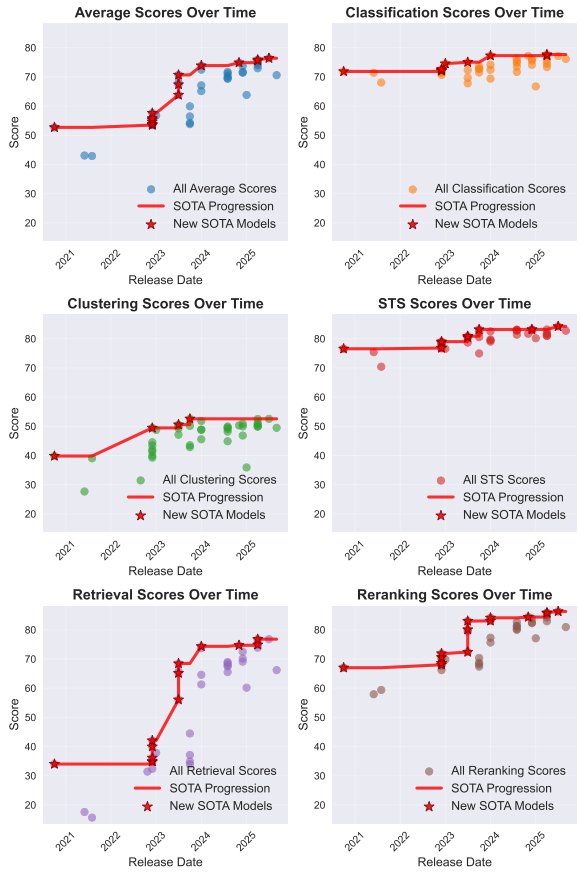

Figure 2: The evolution of state-of-the-art JMTEB average and each task's score.

trend is the dramatic improvement in Retrieval and Reranking tasks, with a particularly sharp increase in mid-2023, coinciding with industry demand for RAG. This improvement is strongly linked to the adoption of weakly-supervised pretraining on large text-pair corpora. The impact of this method is evident in the results in Tab. 1: nearly every model trained with weakly supervised learning achieved a Retrieval score above 60, whereas no model trained without it surpassed that threshold.

Alignment between training data and evaluation task In contrast to the significant improvements in Retrieval and Reranking tasks, the performance gains from the weakly-supervised training recipe did not uniformly transfer to other tasks like STS, Classification, or Clustering, which saw only modest progress, likely because models from both categories are finetuned on similar supervised datasets (e.g., NLI). This highlights a crucial finding: beyond the capability of the backbone model, the alignment between the training data and the characteristics of the evaluation task is a critical determinant of performance.

Furthermore, with the significant diminishing returns now evident from model scaling (as shown in Tab. 1, the top-performing 1.2B model offers only a minor average score improvement over a much more efficient 314M model), future breakthroughs are less likely to come from simply increasing parameter counts. Instead, progress in a specific task will likely depend on developing novel methodologies—such as new training objectives, curated datasets, or architectural refinements—that are specifically aligned with the distinct characteristics of that task. For practitioners, this signifies a shift in strategy: instead of defaulting to the largest model, one must select a model whose training and strengths align with the specific end-goal, such as semantic search or clustering. This application-aware approach also allows for the selection of smaller, more efficient models that offer competitive performance with lower resource consumption.

4. Construction of JMTEB-lite

While JMTEB provides a new gold standard for the definitive evaluation of Japanese embedding models, its scale introduces practical challenges. The computational resources, in terms of time, cost and energy, required to execute the full benchmark can create a significant bottleneck in the typical model development lifecycle. For example, evaluating `sbintuitions/sarashina-embedding-v1-1b` takes nearly 10 hours with a single NVIDIA A100 GPU. The cycle is characterized by frequent experimentation, and thus evaluation can be a big burden if the benchmark is computationally demanding. To address this scalability challenge, this section details the construction process of **JMTEB-lite**, a lightweight version of JMTEB, along with a validation of its speedup effect, fidelity and an additional analysis of robustness.

4.1. Construction Process

The primary cause of inefficiency within JMTEB is the massive corpus size of its retrieval and reranking tasks. For example, datasets such as `Mr.TyDi-ja` and `MIRACL-Retrieval`, contain approximately 7 million documents each, and `JAQKET` consists of over 110,000 documents of full length Wikipedia articles. The large scales of these datasets make frequent re-evaluation computationally restrictive for many researchers and organizations.

To address this, we focus on reducing the corpus size of these specific datasets by adopting the **hard negative pooling strategy**, a proven technique for benchmark optimization used in MMTEB (Enevoldsen et al., 2025). The principle behind this strategy is that the discriminative power of a benchmark comes not from the vast number of irrelevant documents (easy negatives), but from the relatively small subset of documents that are semantically related to the query but not correct answers (hard negatives). By selectively retaining these hard negatives while filtering out easy

Model	#Params	Time Usage by Dataset (seconds)						Total Time Usage (minutes)		
		Mr.TyDi-ja	MIRACL	JaCWIR-Ret.	JAQKET	JaCWIR-Rer.	JQaRA	JMTEB	JMTEB-lite	Speedup
sarashina-embedding-v2-1b	1.22B	13438 → 214 (62.7×)	14264 → 251 (56.8×)	1149 → 614 (1.9×)	2605 → 1488 (1.8×)	1339 → 428 (3.1×)	837 → 529 (1.6×)	597.7	96.0	6.2×
multilingual-e5-large	560M	5418 → 71 (76.3×)	5424 → 84 (64.3×)	456 → 218 (2.1×)	568 → 323 (1.8×)	651 → 167 (3.9×)	372 → 191 (2.0×)	234.5	37.2	6.3×
sup-simcse-ja-large	337M	7744 → 107 (72.2×)	7809 → 126 (61.8×)	677 → 346 (2.0×)	1584 → 1003 (1.6×)	872 → 391 (2.2×)	476 → 274 (1.7×)	352.2	70.2	5.0×
ruri-v3-130m	132M	5724 → 69 (83.3×)	5609 → 79 (70.6×)	485 → 229 (2.1×)	494 → 273 (1.8×)	678 → 275 (2.5×)	350 → 261 (1.3×)	250.8	48.2	5.2×
ruri-v3-70m	70M	4575 → 54 (84.7×)	4484 → 63 (71.5×)	403 → 182 (2.2×)	474 → 262 (1.8×)	600 → 143 (4.2×)	311 → 142 (2.2×)	202.2	35.5	5.7×
ruri-v3-30m	37M	3958 → 47 (83.6×)	3877 → 55 (70.3×)	359 → 160 (2.2×)	461 → 258 (1.8×)	549 → 126 (4.4×)	285 → 127 (2.3×)	176.1	30.9	5.7×
static-embedding-japanese	-	1849 → 20 (94.0×)	1797 → 23 (76.6×)	205 → 71 (2.9×)	450 → 249 (1.8×)	398 → 72 (5.5×)	207 → 72 (2.9×)	95.9	22.6	4.3×
Average of 41 models		(75.0×)	(63.6×)	(2.1×)	(1.7×)	(3.5×)	(1.8×)			5.1×

Table 3: Comparison of time usage for representative models on JMTEB-full and JMTEB-lite. The left side shows the time in seconds for individual datasets, with the format "Full → Lite (speedup)". The right side shows the total time in minutes and the overall speedup.

Task	Dataset	JMTEB Corpus	JMTEB-lite Corpus	Retain Rate
Retrieval	Mr.TyDi-ja	7,000,027	93,382	1.33%
	MIRACL-Retrieval	6,953,614	105,064	1.51%
	JaCWIR-Retrieval	513,107	302,638	58.98%
	JAQKET	114,229	65,802	57.60%
Reranking	JaCWIR-Reranking	513,107	188,033	36.64%
	JQaRA	250,609	172,897	68.99%

Table 4: Comparison of corpus sizes between the JMTEB and JMTEB-lite benchmarks for 6 reduced datasets.

ones, we can substantially reduce the corpus size without compromising the benchmark's ability to differentiate between high-performing models.

We adopt the following reduction process. First, the reduced corpus was initialized with all "gold" documents, which are the correct, relevant documents for every query. Then, using a set of high-performing "oracle" models, we searched the full document corpus for each query to retrieve the top-k most semantically similar documents. These retrieved documents, which act as challenging "hard negatives" or plausible distractors, were then added to the reduced corpus. This method effectively creates a smaller, more efficient corpus that retains the most difficult documents, which are crucial for distinguishing between high-performing models.

4.2. Construction Result

We construct **JMTEB-lite** with the following settings. We downsize 4 Retrieval and 2 Reranking datasets with the hard negative pooling strategy. Tab. 4 shows the result of corpus downsizing. The other datasets are kept exactly unchanged from the full version. For the hard negative pooling process, we set the retrieval depth $k = 50$ for retrieval and reranking tasks, and the oracle models are designated as the 5 top-performing models on the full JMTEB leaderboard.

We extend the empirical analysis (Enevoldsen et al., 2025) of the hard negative pooling strategy by conducting quantitative analyses on the speedup effect, fidelity and robustness.

4.3. Experimental Validations

4.3.1. Settings

To enlarge the range of models involved for more robust results, we conducted the validation process with 41 local models including some disclosed models we have. All experiments were conducted on a single NVIDIA A100 80GB GPU for the evaluation of each model.

We use Pearson correlation coefficient (r) (Gupta et al., 2024; Alam et al., 2025), Spearman's rank correlation coefficient (ρ) (Alam et al., 2025; Enevoldsen et al., 2025) and Kendall's rank correlation coefficient (τ) (Kendall, 1938) to quantify the fidelity of JMTEB-lite to the full JMTEB. We expect these coefficients to be high (~ 1), meaning good alignment between the full and lightweight benchmarks.

4.3.2. Results

Speedup As shown in Tab. 3, JMTEB-lite achieves a significant reduction in the computing time required for evaluation. On average, across all 41 models tested, JMTEB-lite provides a 5.1× speedup over the full JMTEB benchmark. The evaluation time for large models like sbintuitions/sarashina-embedding-v2-1b can be reduced from nearly 10 hours to just over 1.5 hours.

The source of this efficiency gain becomes clear when examining the performance on the 6 datasets targeted for reduction, as detailed in the left part of Tab. 3. The most substantial speedups are observed in the largest-scale retrieval tasks. For Mr.TyDi-ja and MIRACL-Retrieval, both of which feature massive corpora, our downsizing

Task	Dataset	Spearman	Pearson	Kendall
Retrieval	Mr.TyDi-ja	0.9946	0.9978	0.9634
	MIRACL-Retrieval	0.9913	0.9977	0.9439
	JaCWIR-Retrieval	0.9995	1.0000	0.9927
	JAQKET	0.9997	0.9999	0.9951
Reranking	JaCWIR-Reranking	0.9995	0.9997	0.9927
	JQaRA	0.9981	0.9991	0.9805

Table 5: Correlation coefficients between JMTEB and JMTEB-lite.

yields average speedups of 75.0× and 63.6×, respectively. This targeted optimization effectively removes the primary computational bottlenecks, transforming the benchmarking process from a prohibitive, multi-hour commitment into a manageable task suitable for rapid, iterative development.

Fidelity The primary goal of JMTEB-lite is to accelerate evaluation while maintaining high fidelity to the full JMTEB benchmark. To validate the fidelity of JMTEB-lite, we evaluate the 41 models with JMTEB-lite. The results of this analysis, presented in Tab. 5, show that the hard negative pooling strategy produces a lightweight benchmark with higher fidelity. The lite version created using the hard negative pooling strategy achieves near-perfect correlations with the full benchmark. Spearman's rank correlation (ρ) and Pearson correlation (r) are consistently above 0.99 for all datasets. While Kendall's rank correlation (τ) is slightly lower than ρ and r for a few tasks, it remains at high level ($\tau > 0.94$), confirming a strong preservation of model rankings.

We also visualize the absolute scores in JMTEB and JMTEB-lite in Fig. 3. We find that the absolute scores in JMTEB-lite are highly similar to those in JMTEB. It indicates that JMTEB-lite also faithfully preserves the scale of absolute scores and score gaps between models.

Robustness analysis A potential concern with the hard negative pooling strategy used in the creation of JMTEB-lite is its reliance on a fixed set of oracle models to identify challenging examples. As the state-of-the-art in text embeddings advances, the models used to construct JMTEB-lite may no longer represent the performance frontier. A document considered as hard negative for current models might not be a meaningful distractor for more powerful future models. To investigate this, we conducted a robustness analysis by varying the set of oracle models. In detail, we intentionally replaced the oracle models used for the construction of JMTEB-lite (see § 4.1) from top-5 models in the leaderboard to weaker models (models ranked 2-6, 3-7, ..., 37-41 in the leaderboard), and constructed corresponding lite versions. We conducted the same fidelity verifications on these versions. Experiments show that there is a slight decline in fidelity as the oracle models become weaker, yet the correlation scores remain high in all cases

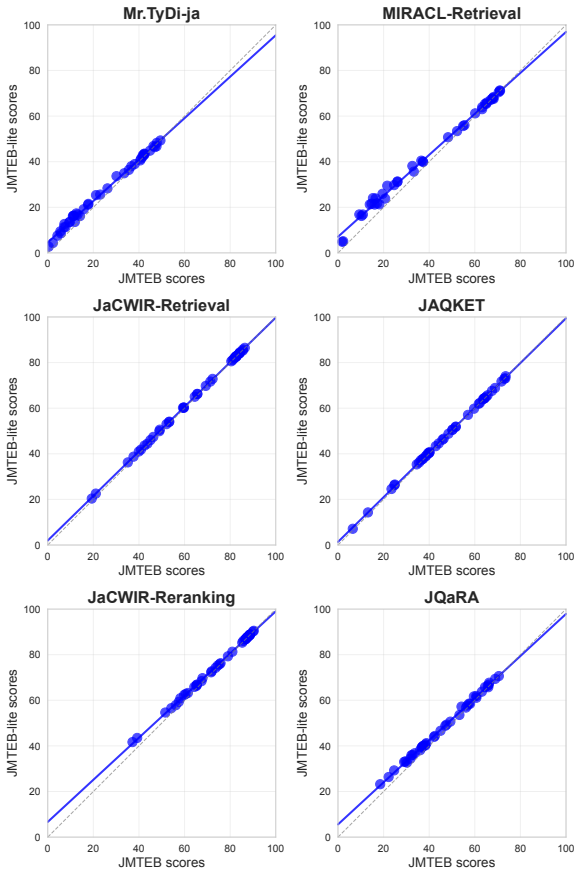

Figure 3: Score correlation between JMTEB-lite and JMTEB. Scatter plots comparing model scores on the full JMTEB (x-axis) against scores on JMTEB-lite (y-axis). The dashed identity line ($y = x$) indicates perfect score agreement. The JMTEB-lite scores show a tight linear correlation, confirming high fidelity.

($\rho > 0.982$, $r > 0.988$, $\tau > 0.900$). This analysis confirms that JMTEB-lite we constructed is not brittle. It effectively captures a diverse set of hard negatives regardless of the specific models used to mine them.

4.3.3. Discussion

Given JMTEB-lite's high fidelity to the full benchmark, the crucial question is why both versions are necessary. The answer lies in their distinct, complementary purposes. JMTEB-lite is optimized for agile development, enabling the rapid, iterative experimentation that is core to the model development lifecycle. In contrast, the full JMTEB is the gold standard for final evaluation. It is designed to simulate the challenge of real-world, web-scale retrieval, testing a model's robustness in a vast and noisy information space of millions of documents. Since strong performance on a smaller scale does not guarantee success in such a demanding setting, we maintain both versions to support both efficient development and comprehensive final assessment.

5. Conclusion

In this work, we addressed the comprehensive evaluation of Japanese text embeddings by introducing JMTEB, the first large-scale, unified benchmark for the Japanese language. Comprising 28 datasets across 5 task families, JMTEB provides a rigorous standard for assessing model capabilities. We conducted a comprehensive evaluation of 40 models on the benchmark to provide extensive baselines.

To facilitate agile development, we also created JMTEB-lite, a lightweight version that achieves an average 5.1× speedup while maintaining high fidelity to the full benchmark. By publicly releasing JMTEB, JMTEB-lite, and associated evaluation tools, we offer a lasting and reproducible foundation to standardize the evaluation of Japanese text embedding models, guide future research, and accelerate progress within the Japanese NLP community.

6. Limitations

As we look to the future, the rapid progress of text embedding technologies has also brought new challenges for the evaluation benchmark. We identify the following limitations and future directions:

- Rapid progress vs. benchmark difficulty: As illustrated in Fig. 2, state-of-the-art scores on the Retrieval task surged from a 35–50 point range to over 70 points in mid-2023 alone. This incredible progress shows that tasks once considered difficult are becoming rapidly solvable. To ensure the community can continue to differentiate top-tier models and drive the next wave of innovation, a critical direction for future work will be the continuous development of more challenging evaluation datasets that keep pace with these ever-improving model capabilities.
- Dataset contamination and memorization: To mitigate the possible influence of data leakage on the effectiveness of the evaluation—given that several Retrieval datasets (e.g., Wikipedia-based `JAQKET`, `MIRACL`, `Mr.TyDi-ja`) are likely present in many pretraining corpora—we currently request that developers declare possible data leakage when a model is added to our leaderboard. On the other hand, accurately distinguishing a model's understanding of language from mere memorization remains an ongoing challenge. In the future, development of novel, contamination-free datasets, as well as determining criteria or developing techniques (e.g., membership inference adapted for vector spaces) to automatically detect and quantify memorization within these benchmark datasets will be an important step to further ensure long-term benchmark integrity.
- Dataset accessibility: Due to strict access agreements (e.g., mandatory individual applications) of some datasets, we are currently unable to integrate certain high-quality resources, such as the NTCIR[4] datasets, into the benchmark's automated pipeline. In the future, we plan to explore compliant ways to incorporate these restricted-access datasets to further broaden the benchmark's coverage.
- Oracle model robustness in JMTEB-lite: While our robustness analysis shows that JMTEB-lite's fidelity is stable across different sets of current oracle models, a key long-term consideration is its validity in the face of major architectural breakthroughs. A future model with fundamentally superior capabilities might easily solve the "hard negatives" that challenge current models, thus reducing the benchmark's ability to differentiate top-tier performance. To ensure JMTEB-lite remains a challenging and relevant benchmark, we commit to its active maintenance, which will include periodically regenerating its corpus by mining new hard negatives with future state-of-the-art models.

7. Bibliographical 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.

Mahammad Parwez Alam, Gayathri Saranathan, Cong Xu, Javier Aula-Blasco, Martin Foltin, Tarun Kumar, Soon Yee Wong, and Suparna Bhattacharya. 2025. SubLIME*: Data Efficient Foundation Model Evaluation across Modalities, Languages and Benchmarks. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

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.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[4] https://research.nii.ac.jp/ntcir

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

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.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Vipul Gupta, Candace Ross, David Pantoja, Rebecca J Passonneau, Megan Ung, and Adina Williams. 2024. Improving model evaluation using smart filtering of benchmark datasets. *arXiv preprint arXiv:2410.20245*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with {P}ersian. 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, Loic 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 2014–2037.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-Aware Contrastive Learning of Sentence Embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.

James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I Wish I Would Have Loved This One, But I Didn't – A Multilingual Dataset for Counterfactual Detection in Product Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. *arXiv preprint arXiv:2206.06588*.

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.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A Complex, Natural, and Multilingual Dataset for End-to-End Question Answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2022. A Japanese dataset for subjective and objective sentiment polarity classification in micro blog domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7022–7028.

Masatoshi Suzuki, Jun Suzuki, Koji Matsuda, Kyosuke Nishida, and Naoya Inoue. 2020. JAQKET: Construction of a Quiz-based Japanese QA Dataset. In *Proceedings of the 26th Annual Meeting of the Association for Natural Language Processing (NLP2020)*. In Japanese.

Yixuan Tang and Yi Yang. 2025. Finmteb: Finance massive text embedding benchmark. *arXiv preprint arXiv:2502.10990*.

Yuichi Tateno. 2024a. JaCWIR: Japanese Casual Web IR - A small-scale dataset for Japanese IR consisting of casual web titles and outlines.

Yuichi Tateno. 2024b. JQaRA: Japanese Question Answering with Retrieval Augmentation - A Japanese Q&A Dataset for RAG Evaluation.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hayato Tsukagoshi and Ryohei Sasano. 2024. Ruri: Japanese general text embeddings. *arXiv preprint arXiv:2409.07737*.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2023. Japanese SimCSE technical report. *arXiv preprint arXiv:2310.19349*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional Evaluation on Japanese Textual Entailment

and Similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. *arXiv:2108.08787*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11730–11738.