

# Multilingual E5 Text Embeddings: A Technical Report

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei

Microsoft Corporation

{wangliang,nanya,fuwei}@microsoft.com

## Abstract

This technical report presents the training methodology and evaluation results of the open-source multilingual E5 text embedding models, released in mid-2023. Three embedding models of different sizes (small / base / large) are provided, offering a balance between the inference efficiency and embedding quality. The training procedure adheres to the English E5 model recipe, involving contrastive pre-training on 1 billion multilingual text pairs, followed by fine-tuning on a combination of labeled datasets. Additionally, we introduce a new instruction-tuned embedding model, whose performance is on par with state-of-the-art, English-only models of similar sizes. Information regarding the model release can be found at <https://github.com/microsoft/unilm/tree/master/e5>.

## 1 Introduction

Text embeddings serve as fundamental components in information retrieval systems and retrieval-augmented language models. Despite their significance, most existing embedding models are trained exclusively on English text (Reimers and Gurevych, 2019; Ni et al., 2022b,a), thereby limiting their applicability in multilingual contexts.

In this technical report, we present the multilingual E5 text embedding models (*mE5-{small / base / large}*), which extend the English E5 models (Wang et al., 2022). The training procedure adheres to the original two-stage methodology: weakly-supervised contrastive pre-training on billions of text pairs, followed by supervised fine-tuning on small quantity of high-quality labeled data. We also release an instruction-tuned embedding model<sup>1</sup> *mE5-large-instruct* by utilizing the synthetic data from Wang et al. (2023). Instructions can better inform embedding models about

the task at hand, thereby enhancing the quality of the embeddings.

For model evaluation, we first demonstrate that our multilingual embeddings exhibit competitive performance on the English portion of the MTEB benchmark (Muennighoff et al., 2023), and the instruction-tuned variant even surpasses strong English-only models of comparable sizes. To showcase the multilingual capability of our models, we also assess their performance on the MIRACL multilingual retrieval benchmark (Zhang et al., 2023) across 16 languages and on Bitext mining (Zweigenbaum et al., 2018; Artetxe and Schwenk, 2019) in over 100 languages.

## 2 Training Methodology

	# Sampled
Wikipedia	150M
mC4	160M
Multilingual CC News	160M
NLLB	160M
Reddit	160M
S2ORC	50M
Stackexchange	50M
xP3	80M
Misc. SBERT Data	10M
Total	~1B

Table 1: Data mixture for contrastive pre-training.

**Weakly-supervised Contrastive Pre-training** In the first stage, we continually pre-train our model on a diverse mixture of multilingual text pairs obtained from various sources as listed in Table 1. The models are trained with a large batch size  $32k$  for a total of  $30k$  steps, which approximately goes over  $\sim 1$  billion text pairs. We employ the standard InfoNCE contrastive loss with only in-batch negatives, while other hyperparameters remain consistent with the English E5 models (Wang et al., 2022).

<sup>1</sup>Here instructions refer to the natural language descriptions of the embedding tasks.

	# Sampled
MS-MARCO Passage	500k
MS-MARCO Document	70k
NQ, TriviaQA, SQuAD	220k
NLI	275k
ELI5	100k
NLLB	100k
DuReader Retrieval	86k
Fever	70k
HotpotQA	70k
Quora Duplicate Questions	15k
Mr. TyDi	50k
MIRACL	40k
Total	~1.6M

Table 2: Data mixture for supervised fine-tuning.

**Supervised Fine-tuning** In the second stage, we fine-tune the models from the previous stage on a combination of high-quality labeled datasets. In addition to in-batch negatives, we also incorporate mined hard negatives and knowledge distillation from a cross-encoder model to further enhance the embedding quality. For the *mE5*-{*small* / *base* / *large*} models released in mid-2023, we employ the data mixture shown in Table 2.

For the *mE5-large-instruct* model, we adopt the data mixture from Wang et al. (2023), which includes additional 500k synthetic data generated by GPT-3.5/4 (OpenAI, 2023). This new mixture encompasses 150k unique instructions and covers 93 languages. We re-use the instruction templates from Wang et al. (2023) for both the training and evaluation of this instruction-tuned model.

### 3 Experimental Results

**English Text Embedding Benchmark** Multilingual embedding models should be able to perform well on English tasks as well. In Table 3, we compare our models with other multilingual and English-only models on the MTEB benchmark (Muennighoff et al., 2023). Our best mE5 model surpasses the previous state-of-the-art multilingual model Cohere<sub>multilingual-v3</sub>, by 0.4 points and outperforms a strong English-only model, BGE<sub>large-en-v1.5</sub>, by 0.2 points. While smaller models demonstrate inferior performance, their faster inference and reduced storage costs render them advantageous for numerous applications.

**Multilingual Retrieval** We evaluate the multilingual retrieval capability of our models using

	MTEB (56 datasets)
LaBSE	45.2
Cohere <sub>multilingual-v3</sub>	64.0
BGE <sub>large-en-v1.5</sub>	64.2
mE5 <sub>small</sub>	57.9
mE5 <sub>base</sub>	59.5
mE5 <sub>large</sub>	61.5
mE5 <sub>large-instruct</sub>	<b>64.4</b>

Table 3: Results on the English portion of the MTEB benchmark. LaBSE (Feng et al., 2022) is exclusively trained on translation pairs. Limited information is available regarding the training data and model size are available for Cohere<sub>multilingual-v3</sub> (<https://txt.cohere.com/introducing-embed-v3/>). BGE<sub>large-en-v1.5</sub> (Xiao et al., 2023) is an English-only model. Full results are in Appendix Table 7.

	nDCG@10	R@100
BM25	39.3	78.7
mDPR	41.5	78.8
mE5 <sub>small</sub>	60.8	92.4
mE5 <sub>base</sub>	62.3	93.1
mE5 <sub>large</sub>	<b>66.5</b>	94.3
mE5 <sub>large-instruct</sub>	65.7	<b>94.6</b>

Table 4: Multilingual retrieval on the development set of the MIRACL benchmark. Numbers are averaged over 16 languages.

the MIRACL benchmark (Zhang et al., 2023). As shown in Table 4, mE5 models significantly outperform mDPR, which has been fine-tuned on the MIRACL training set, in both nDCG@10 and recall metrics. Detailed results on individual languages are provided in Appendix Table 6.

	BUCC 2018 4 langs	Tatoeba 112 langs
mContriever <sub>rmsmarco</sub>	93.7	37.7
LaBSE	98.8	81.1
mE5 <sub>small</sub>	93.2	64.2
mE5 <sub>base</sub>	98.1	68.1
mE5 <sub>large</sub>	98.6	75.7
mE5 <sub>large-instruct</sub>	<b>99.0</b>	<b>83.8</b>

Table 5: Bitext mining results. mContriever (Izacard et al., 2021) numbers are run by ourselves based on the released checkpoint.

**Bitext Mining** is a cross-lingual similarity search task that requires the matching of two sentences with little lexical overlap. As demonstrated in Table 5, mE5 models exhibit competitive performance across a broad range of languages, both

high-resource and low-resource. Notably, the mE5<sub>large-instruct</sub> model surpasses the performance of LaBSE, a model specifically designed for bitext mining, due to the expanded language coverage afforded by the synthetic data (Wang et al., 2023).

## 4 Conclusion

In this brief technical report, we introduce multilingual E5 text embedding models that are trained with a multi-stage pipeline. By making the model weights publicly available, practitioners can leverage these models for information retrieval, semantic similarity, and clustering tasks across a diverse range of languages.

## References

- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. *Ms marco: A human generated machine reading comprehension dataset*. *ArXiv preprint*, abs/1611.09268.
- 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*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. *Quora question pairs*.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. *ELIS: Long form question answering*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. *SimCSE: Simple contrastive learning of sentence embeddings*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. *Towards unsupervised dense information retrieval with contrastive learning*. *ArXiv preprint*, abs/2112.09118.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. *Dense passage retrieval for open-domain question answering*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.
- 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, Dubrovnik, Croatia. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.
- Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022a. *Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models*. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022b. *Large dual encoders are generalizable retrievers*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- OpenAI. 2023. *Gpt-4 technical report*. *ArXiv preprint*, abs/2303.08774.
- Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, QiaoQiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. *DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine*.

- In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- 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*, abs/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*.
- Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. Minilmv2: Multi-head self-attention relation distillation for compressing pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *ArXiv preprint*, abs/2309.07597.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.
- 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.
- Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. [Mr. TyDi: A multi-lingual benchmark for dense retrieval](#). In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xinyu Crystina 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.
- Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.

## A Implementation Details

**Contrastive Pre-training Text Pairs** In Table 1, to construct text pairs, we utilize (section title, section passage) for Wikipedia, (title, page content) for mC4 (Xue et al., 2021), (title, news content) for multilingual CCNews<sup>2</sup>, translation pairs for NLLB (Costa-jussà et al., 2022), (comment, response) for Reddit<sup>3</sup>, (title, abstract) and citation pairs for S2ORC (Lo et al., 2020), (question, answer) for Stackexchange<sup>4</sup>, (input prompt, response) for xP3 (Muennighoff et al., 2022). For the miscellaneous SBERT data<sup>5</sup>, we include the following datasets: SimpleWiki, WikiAnswers, AGNews, AltLex, AmazonQA, AmazonReview, CNN/DailyMail, CodeSearchNet, Flickr30k, GooAQ, NPR, SearchQA, SentenceCompression, Specter, WikiHow, XSum, and YahooAnswers.

**Data Mixture for Supervised Fine-tuning** It includes ELI5 (Fan et al., 2019)(sample at 20%), HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), MIRACL (Zhang et al., 2023), MSMARCO passage ranking and document ranking (sample at 20%) (Campos et al., 2016), NQ (Karpukhin et al., 2020), NLLB (sample at 100k) (Costa-jussà et al., 2022), NLI (Gao et al., 2021), SQuAD (Karpukhin et al., 2020), TriviaQA (Karpukhin et al., 2020), Quora Duplicate Questions (DataCanary et al.,

<sup>2</sup><https://commoncrawl.org/blog/news-dataset-available>

<sup>3</sup><https://www.reddit.com/>

<sup>4</sup><https://stackexchange.com/>

<sup>5</sup><https://huggingface.co/datasets/sentence-transformers/embedding-training-data>



	nDCG@10				R@100			
	mE5 <sub>small</sub>	mE5 <sub>base</sub>	mE5 <sub>large</sub>	E5 <sub>large-instruct</sub>	mE5 <sub>small</sub>	mE5 <sub>base</sub>	mE5 <sub>large</sub>	E5 <sub>large-instruct</sub>
ar	71.4	71.6	76.0	76.8	96.2	95.9	97.3	97.5
bn	68.2	70.2	75.9	73.9	97.4	96.6	98.2	98.2
en	48.0	51.2	52.9	51.5	85.3	86.4	87.6	88.2
es	51.2	51.5	52.9	53.7	87.6	88.6	89.1	89.3
fa	53.3	57.4	59.0	59.4	90.4	91.2	92.9	92.9
fi	73.3	74.4	77.8	77.3	96.3	96.9	98.1	97.9
fr	47.6	49.7	54.5	53.7	89.5	90.0	90.6	91.7
hi	55.2	58.4	62.0	60.3	91.0	92.6	93.9	94.1
id	50.7	51.1	52.9	52.1	86.2	87.4	87.9	88.4
ja	63.6	64.7	70.6	69.0	95.2	96.0	97.1	96.9
ko	61.2	62.2	66.5	65.3	92.0	91.6	93.4	93.0
ru	59.1	61.5	67.4	67.9	92.2	92.7	95.5	95.4
sw	68.4	71.1	74.9	72.5	94.7	95.6	96.7	97.2
te	81.3	75.2	84.6	83.4	97.6	98.0	99.2	99.0
th	75.0	75.2	80.2	78.6	98.2	98.0	98.9	98.7
zh	45.9	51.5	56.0	56.2	87.9	92.1	93.3	94.9
Avg	60.8	62.3	66.5	65.7	92.4	93.1	94.3	94.6

Table 6: nDCG@10 and R@100 on the development set of the MIRACL dataset.

2017)(sample at 10%), MrTyDi (Zhang et al., 2021), and DuReader (Qiu et al., 2022) datasets.

For the *mE5-large-instruct* model, we employ the new data mixture from Wang et al. (2023). The main difference is the inclusion of synthetic data from GPT-4.

**Training Hyperparameters** The mE5<sub>small</sub>, mE5<sub>base</sub> and mE5<sub>large</sub> are initialized from the multi-lingual MiniLM (Wang et al., 2021), *xlm-roberta-base* (Conneau et al., 2020), and *xlm-roberta-large* respectively. For contrastive pre-training, the learning rate is set to  $\{3, 2, 1\} \times 10^{-4}$  for the {small, base, large} models. For fine-tuning, we use batch size 512 and learning rate  $\{3, 2, 1\} \times 10^{-5}$  for the {small, base, large} models. All models are fine-tuned for 2 epochs. The *mE5-large-instruct* model adopts the same hyperparameters as the mE5<sub>large</sub> large, but is fine-tuned on the new data mixture by Wang et al. (2023).

Dataset	mE5 <sub>small</sub>	mE5 <sub>base</sub>	mE5 <sub>large</sub>	mE5 <sub>large-instruct</sub>
BIOSSES	82.3	85.1	82.5	87.0
SICK-R	77.5	78.5	80.2	81.7
STS12	76.6	76.7	80.0	82.6
STS13	77.0	78.0	81.5	87.2
STS14	75.5	76.6	77.7	85.0
STS15	87.1	88.2	89.3	91.0
STS16	83.6	84.3	85.8	87.3
STS17	86.4	87.8	88.1	90.0
STS22	60.9	61.8	63.1	67.6
STSBenchmark	84.0	85.6	87.3	88.4
SummEval	30.0	30.1	29.7	30.4
SprintDuplicateQuestions	92.2	93.0	93.1	91.2
TwitterSemEval2015	70.8	72.2	75.3	80.3
TwitterURLCorpus	84.8	85.5	85.8	87.1
AmazonCounterfactualClassification	73.8	79.0	79.1	76.2
AmazonPolarityClassification	88.7	90.6	93.5	96.3
AmazonReviewsClassification	44.7	44.5	47.6	56.7
Banking77Classification	79.4	82.7	84.7	85.7
EmotionClassification	42.5	45.2	46.5	51.5
ImdbClassification	80.8	85.5	90.2	94.6
MassiveIntentClassification	70.3	72.1	73.8	77.1
MassiveScenarioClassification	74.5	77.1	77.5	80.5
MTOPDomainClassification	91.1	93.1	93.7	93.9
MTOPIntentClassification	71.1	75.3	77.9	82.5
ToxicConversationsClassification	69.4	69.8	71.3	71.1
TweetSentimentExtractionClassification	62.6	61.3	62.0	64.6
AskUbuntuDupQuestions	57.9	58.2	60.3	63.9
MindSmallReranking	30.3	31.0	31.4	33.1
SciDocsRR	78.1	80.7	82.0	85.9
StackOverflowDupQuestions	49.2	49.4	49.7	51.5
ArxivClusteringP2P	39.2	40.3	44.3	46.4
ArxivClusteringS2S	30.8	35.4	38.4	40.5
BiorxivClusteringP2P	35.8	35.0	35.3	40.9
BiorxivClusteringS2S	27.1	29.5	33.5	36.3
MedrxivClusteringP2P	30.9	28.9	31.5	36.9
MedrxivClusteringS2S	27.3	28.4	29.7	35.5
RedditClustering	39.1	42.4	46.5	56.6
RedditClusteringP2P	59.0	55.2	63.2	64.3
StackExchangeClustering	53.5	55.3	57.5	66.8
StackExchangeClusteringP2P	32.1	30.5	32.7	42.5
TwentyNewsgroupsClustering	33.2	36.0	38.9	51.3
ArguAna	39.1	44.2	54.4	58.4
ClimateFEVER	22.6	23.9	25.7	29.9
CQADupstackAndroidRetrieval	36.1	38.5	39.7	42.7
DBPedia	37.8	40.4	41.3	38.4
FEVER	75.3	79.4	82.8	78.0
FiQA2018	33.3	38.2	43.8	47.7
HotpotQA	65.1	68.6	71.2	69.3
MSMARCO	41.0	42.3	43.7	40.4
NFCorpus	31.0	32.5	34.0	35.5
NQ	56.3	60.0	64.1	57.8
QuoraRetrieval	86.9	87.7	88.2	89.2
SCIDOCS	13.9	17.2	17.5	18.7
SciFact	67.7	69.3	70.4	71.8
Touche2020	21.2	21.4	23.4	27.2
TRECCOVID	72.6	69.8	71.3	82.0
Average	57.9	59.4	61.5	<b>64.4</b>

Table 7: Results for each dataset in the MTEB benchmark. The evaluation metrics are available in the original paper (Muennighoff et al., 2023).