MMTEB — Massive Multilingual Text Embedding Benchmark

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**, Niklas **Muennighoff**

OVERVIEW & RATIONALE

Rationale

Text embeddings are **often evaluated on a limited set of tasks**, constrained by language, domain, and task diversity

Text embeddings play an **essential role in LLM inference**, including RAG and few-shot classification, data curation, and more

Contributions

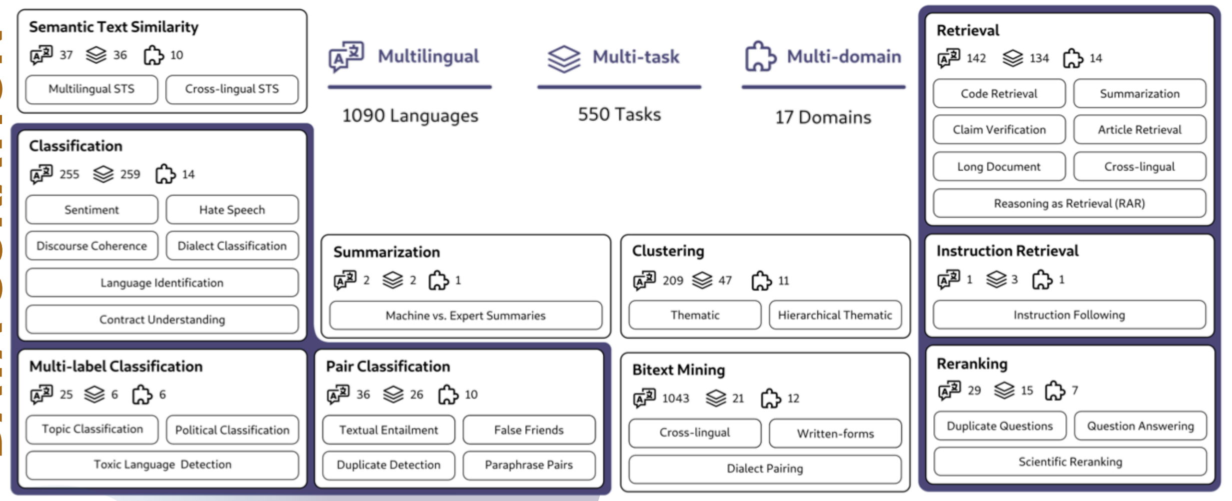
MMTEB covers >500 quality-controlled evaluation tasks across >250 languages, making it the **largest multilingual benchmark**

MMTEB **introduces a diverse set of tasks**, including instruction following, long document retrieval, code retrieval, and more

We **reveal a notable performance gap** appearing already among mid-resource languages such as German and Polish

Significantly speed up evaluation using only 2% of previous benchmark documents for comparable benchmarks

DATA CURATION



Task Curation

All task in MMTEB was collected through an **open-science effort** using a point-system to determine co-authorship.

Each task has extensive metadata on annotation source, dataset source, license, dialect, citation information, etc.

And was **reviewed along various axes**, including checking for performance ceilings, implementation bugs, and the ability to discriminate between models.

>500 tasks
>250 Languages

OPTIMIZATIONS

Speed Optimizations

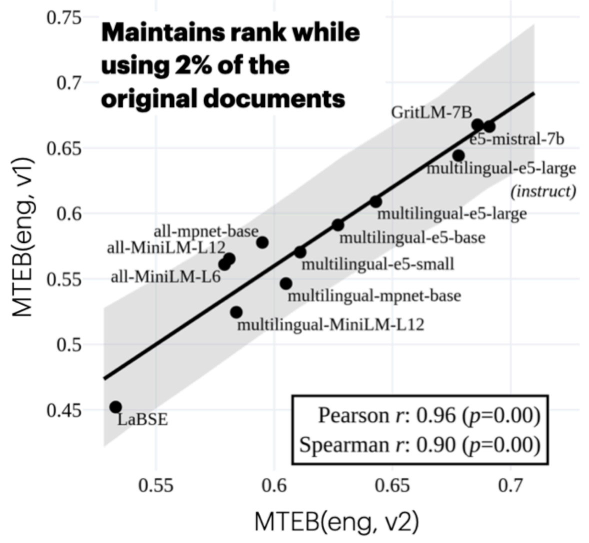
To keep the benchmark accessible for low-resource communities, we optimize by:

Encouraging smaller dataset submission, typically ~2048 samples is enough to differentiate between models

For clustering tasks, we used bootstrapping-based **downsampling to reduce the number of documents by ~16x**

For retrieval, we use **hard-negative mining** across diverse models, keeping the top 250 ranked documents pr. query

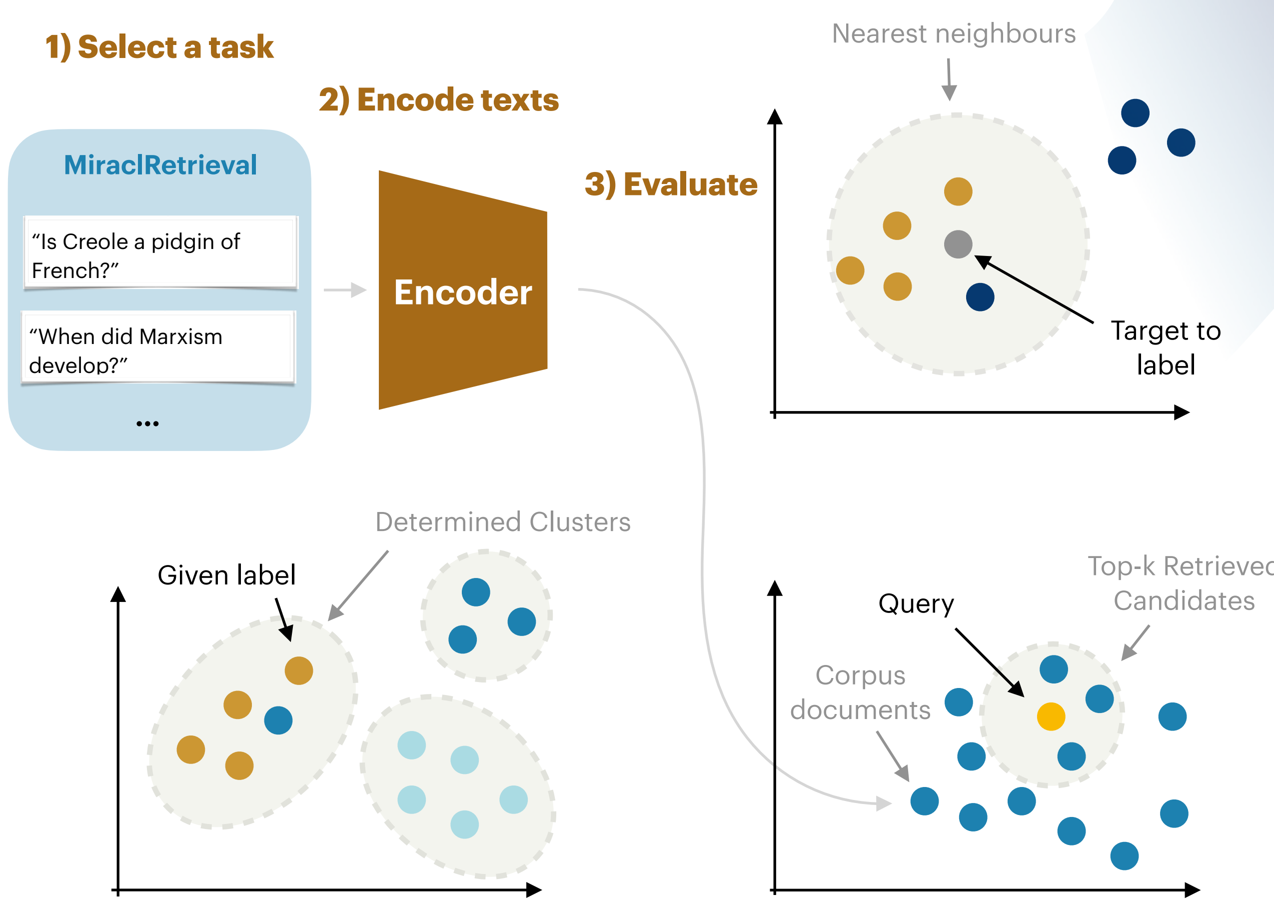
We perform **task downsampling** to remove highly correlated tasks while maintaining benchmark sensitivity



Same or better **performance estimate using 50x fewer documents**

EVALUATING EMBEDDINGS

Primer on Evaluation Embeddings



Largest Multilingual Benchmark for Embeddings

Want know more? Check out the paper!

Want a similar benchmark for image and text embeddings? Great news! We will soon release an upcoming work on images. See the results

MULTILINGUAL EVALUATION

Model (↓)	Rank (↓) Borda Count	Average Across All	Average Across Category	Average per Category Btxt	Pr Clf	Clf	STS	Rtrvl	M. Clf	Clust	Rrnk
MTEB(Multilingual)											
Number of datasets (→)	(132)	(132)	(132)	(13)	(11)	(43)	(16)	(18)	(5)	(17)	(6)
multilingual-e5-large-instruct	1 (1375)	**63.2**	**62.1**	**80.1**	80.9	**64.9**	**76.8**	57.1	**22.9**	**51.5**	62.6
GritLM-7B	2 (1258)	60.9	60.1	70.5	79.9	61.8	73.3	**58.3**	22.8	50.5	**63.8**
e5-mistral-7b-instruct	3 (1233)	60.3	59.9	70.6	81.1	60.3	74.0	55.8	22.2	51.4	**63.8**
multilingual-e5-large	4 (1109)	58.6	58.2	71.7	79.0	59.9	73.5	54.1	21.3	42.9	**62.8**
multilingual-e5-base	5 (944)	57.0	56.5	69.4	77.2	58.2	71.4	52.7	20.2	42.7	60.2
multilingual-mpnet-base	6 (830)	52.0	51.1	52.1	**81.2**	55.1	69.7	39.8	16.4	41.1	53.4
multilingual-e5-small	7 (784)	55.5	55.2	67.5	76.3	56.5	70.4	49.3	19.1	41.7	60.4
LaBSE	8 (719)	52.1	51.9	76.4	76.0	54.6	65.3	33.2	20.1	39.2	50.2
multilingual-MiniLM-L12	9 (603)	48.8	48.0	44.6	79.0	51.7	66.6	36.6	14.9	39.3	51.0
all-mpnet-base	10 (526)	42.5	41.1	21.2	70.9	47.0	57.6	32.8	16.3	40.8	42.2
all-MiniLM-L12	11 (490)	42.2	40.9	22.9	71.7	46.8	57.2	32.5	14.6	36.8	44.3
all-MiniLM-L6	12 (418)	41.4	39.9	20.1	71.2	46.2	56.1	32.5	15.1	38.0	40.3

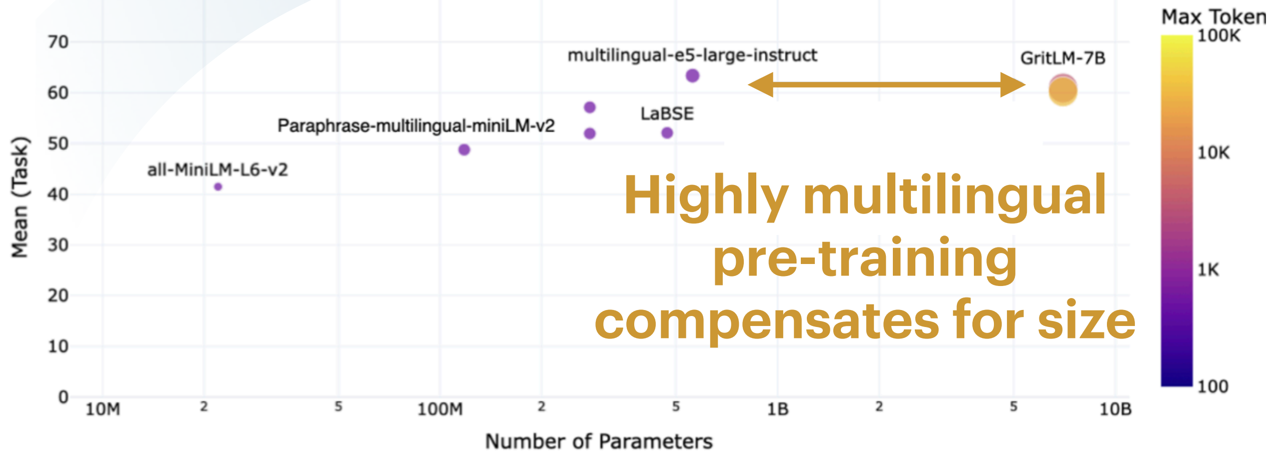
Results and Findings

Models trained with instruction-tuning perform significantly better compared to those without. The two large multilingual e5 models are a clear example of this.

Discrepancies in multilingual benchmarks stem from differences in the pre-training. This suggests that multilingual pre-training of the base models will likely lead to performance gains.

Want to see more performance scores? Check out the public leaderboard with >200 models

LANGUAGE GAP



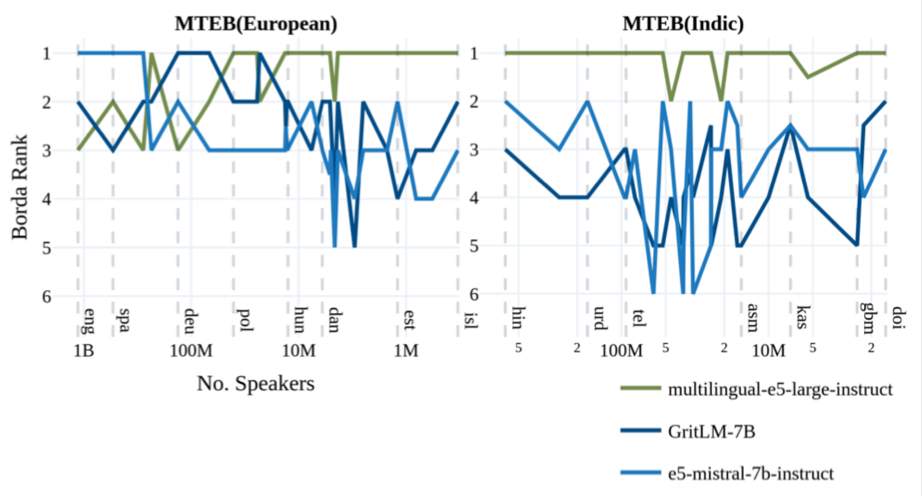
Highly multilingual pre-training compensates for size

Language Gap

Multilingual 7b models outperformed by notably smaller models (560M) in low-resource settings. This is likely due to pre-training of the base model.

In truly low-resource settings, the smaller XLM-R-based multilingual-e5-large-instruct consistently outperforms larger models.

This **suggests the need for better multilingual base models** as XLM-R-based models still outperform Mistral or Llama-based encoders.



CENTER FOR HUMANITIES COMPUTING