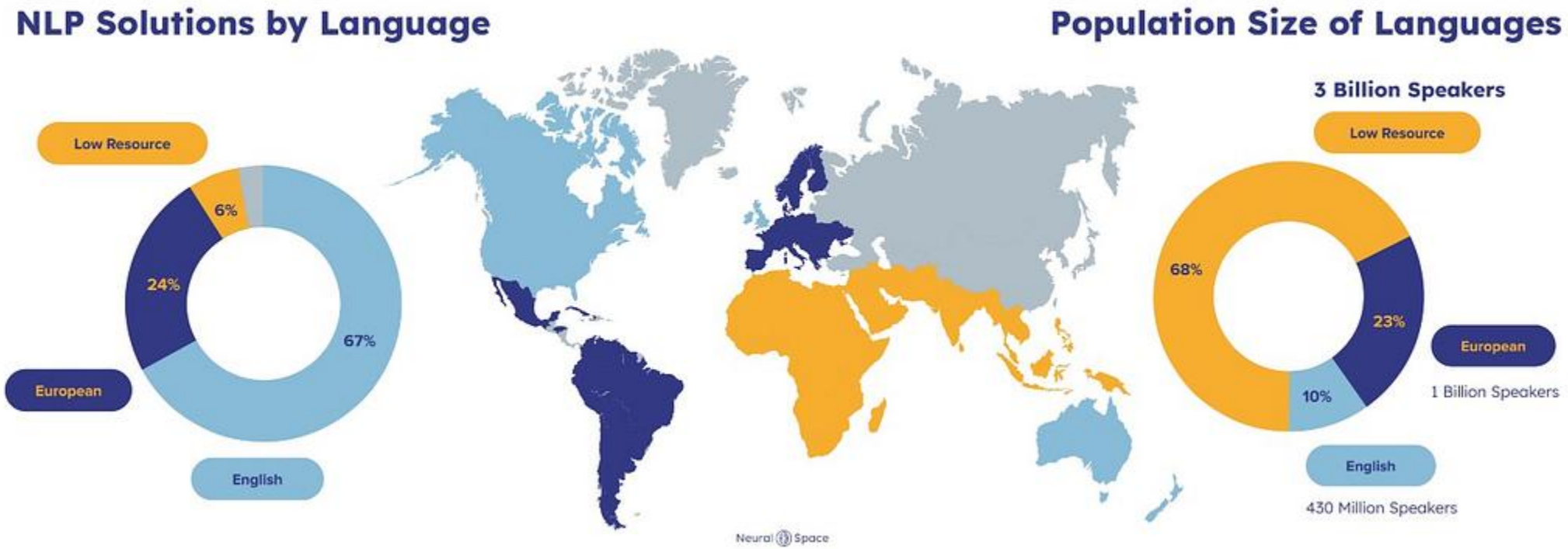
MMTEB: MASSIVE MULTILINGUAL TEXT EMBEDDING BENCHMARK

Kenneth Enevoldsen[1,†,‡], Isaac Chung[2,‡], Imene Kerboua[3,4,‡], Márton Kardos[1,‡], Ashwin Mathur[2], David Stap[5], Jay Gala[6], Wissam Siblini[2], Dominik Krzemiński[2], Genta Indra Winata[2], Saba Sturua[7], Saiteja Utpala[8], Mathieu Ciancone[9], Marion Schaeffer[9], Gabriel Sequeira[2], Diganta Misra[57,58], Shreeya Dhakal[2], Jonathan Rystrøm[11], Roman Solomatin[12,‡], Ömer Çağatan[13], Akash Kundu[14,15], Martin Bernstorff[1], Shitao Xiao[16], Akshita Sukhlecha[2], Bhavish Pahwa[8], Rafał Poświata[17], Kranthi Kiran GV[18], Shawon Ashraf[19], Daniel Auras[19], Björn Plüster[19], Jan Philipp Harries[19], Loïc Magne[2], Isabelle Mohr[7], Mariya Hendriksen[5], Dawei Zhu[20], Hippolyte Gisserot-Boukhlef[21,22], Tom Aarsen[23,‡], Jan Kostkan[1], Konrad Wojtasik[24], Taemin Lee[25], Marek Šuppa[27,28], Crystina Zhang[29], Roberta Rocca[1], Mohammed Hamdy[30], Andrianos Michail[31], John Yang[32], Manuel Faysse[21,26], Aleksei Vatolin[33], Nandan Thakur[29], Manan Dey[34], Dipam Vasani[2], Saksham Thakur[2], Pranjal Chitale[35], Simone Tedeschi[36,37], Nguyen Tai[38], Artem Snegirev[39], Michael Günther[7], Mengzhou Xia[40], Weijia Shi[41], Xing Han Lù[10], Jordan Clive[42], Gayatri Krishnakumar[43], Anna Maksimova[39], Silvan Wehrli[44], Maria Tikhonova[39,45], Henil Panchal[46], Aleksandr Abramov[39], Malte Ostendorff[47], Zheng Liu[16], Simon Clematide[31], Lester James Miranda[48], Alena Fenogenova[39], Guangyu Song[49], Ruqiya Bin Safi[50], Wen-Ding Li[51], Alessia Borghini[37], Federico Cassano[52], Hongjin Su[53], Jimmy Lin[29], Howard Yen[40], Lasse Hansen[1], Sara Hooker[30], Chenghao Xiao[54,‡], Vaibhav Adlakha[10, 55,‡], Orion Weller[56,‡], Siva Reddy[10,55,‡], Niklas Muennighoff[32,48,59,‡]

Presenter: Riya Bhadani

Background

- Text embeddings form the bedrock of LLM understanding
- Enable semantic understanding across languages
- Critical in improving the performance of models on low-to mid-resource languages
  - Low- to mid- resource language: limited high-quality, annotated corpus
  - High-resource languages: many high-quality copra, consists of English, Mandarin, Japanese



Laumann, 2022

Prior Works

Text Embeddings

- MNET: Massive Text Embedding Benchmark (Muennighoff et al., 2023): 8 embeddings tasks covering 58 datasets and 112 languages
  - Key Finding**:** no single benchmark dominated across tasks, indicating that as of now, there is no single optimal embedding model
- MIRACL (Zhang et al., 2022): examines monolingual retrieval across 18 languages
- MINERS (Winata et al., 2024): evaluate multilingual LMs’ abilities in semantic retrieval tasks

Motivation

01

Existing embedding benchmarks are limited by **language**, domains, and tasks

02

Computational expensive to run, limiting their accessibility

Contributions

01

MMTEB construction spans 1000+ languages, 550 tasks, and 17 domains

02

Accessible to low-compute communities

03

Provide starter code to community to encourage the creation of more benchmarks

Tasks

Classification

Pair classification: predict labels pairs of embedded documents

Bitext mining: matching pairs of sentences

Clustering and hierarchical clustering

Retrieval: retrieval of documents within then copra based on basic query and keywords

Multi-label classification

Instruction retrieval: retrieval tasks based on detailed instructions

Reranking: rank documents in terms of relevant based on a query

Semantic text similarity

Data

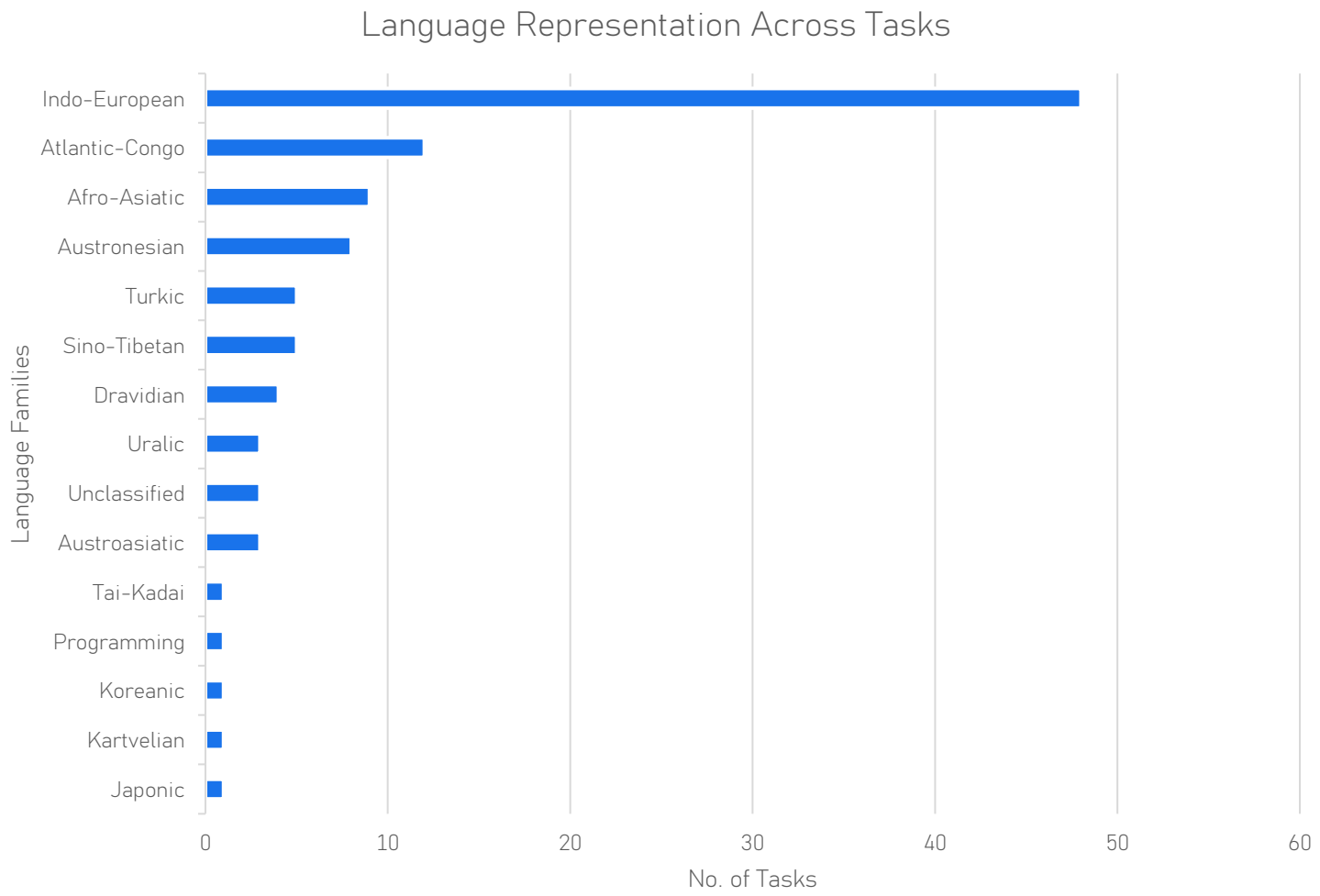
- Largely from publicly available datasets for practical purposes
- Sources:

Academic sources	Religious texts	News	Web	Nonfiction	Fiction
Social media communications	Reviews	Subtitles	Poetry	Official government documents	Programming

Data Selection and Quality

- Encouraged participated such industry experts, academia, and individuals from low-resource communities to submit datasets
- Quality was ensured by testing tasks on two multilingual models
  - Multilingual-e5-small (trained on 100 languages) and MiniLM-L12 (trained on 50+ languages)
- Tasks examined for flawed implementation and poor data quality
- Tasks were not excluded if it meant removing a language from the corpus

Language Coverage Across Tasks



Language Coverage Across Tasks

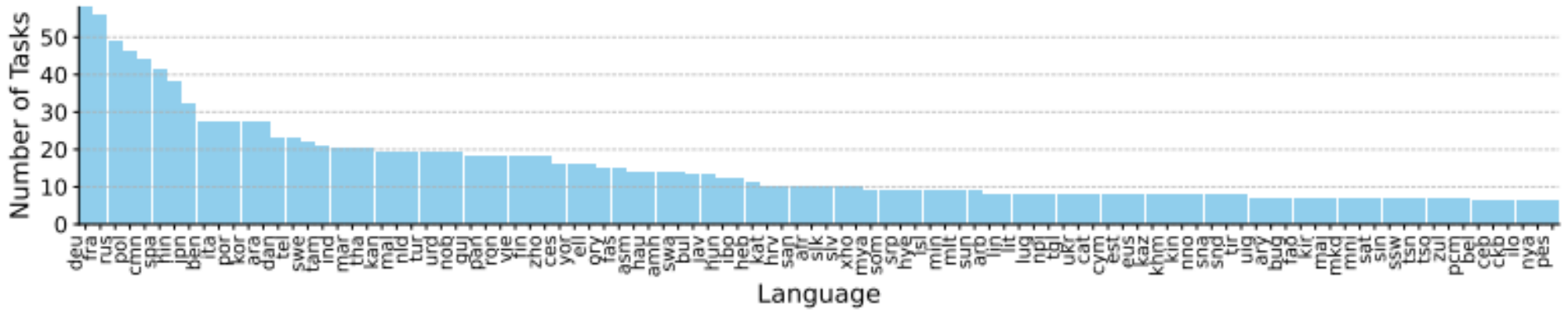


Figure 6: Number of tasks per language. For readability, we remove English (290 tasks) and only plot the 100 languages with the most tasks.

Benchmark Optimizations

1. Downsampling datasets: minimize number of documents to be encoded
   - Clustering: k-means clustering of embeddings to construct representative sample
   - Retrieval: aggregate scores from multiple models to construct subset
   - Balance between robustness and speed up: downsample to 4% of the dataset
2. Caching embeddings
   - Bitext Mining: handle duplicate datapoints across languages
3. Encouraged smaller dataset submissions

Effects of Downsampling on Clustering Tasks

Task	Spearman	Speedup
Biorxiv P2P	0.9505	31.50x
Biorxiv S2S	0.9890	14.31x
Medrxiv P2P	0.9615	21.48x
Medrxiv S2S	0.9560	8.39x
Reddit S2S	0.9670	11.72x
Reddit P2P	0.9670	22.77x
StackExchange S2S	0.9121	9.55x
StackExchange P2P	0.9670	20.20x
TwentyNewsgroups	1.0000	5.02x
Average	0.9634	16.11x

Table 6: Agreement on model rankings on a selection of English clustering tasks using Spearman's correlation across the scores of 13 models of various sizes.

Model Ranking Preserved with Downsampling

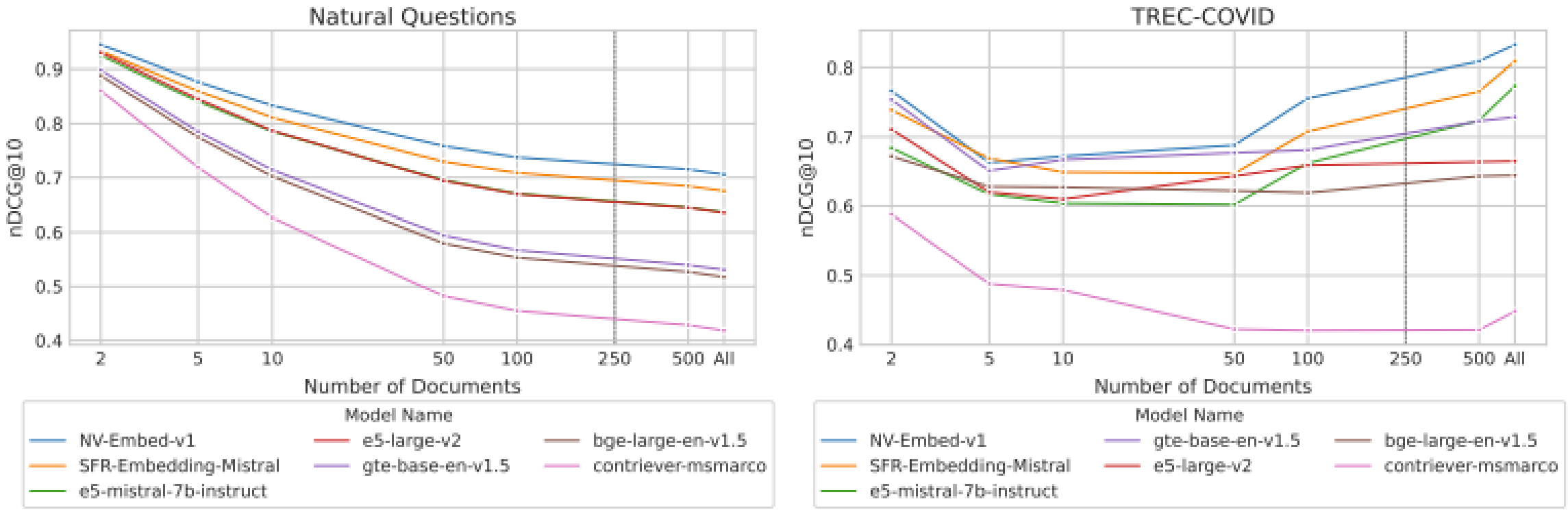


Figure 10: Absolute scores of different models on subsampled versions of the datasets using hard negatives. NQ has 1 relevant document per query while TREC-COVID has 500+ relevant documents per query which is why we see NQ scores gradually increasing whereas TREC-COVID scores vary.

Benchmarks

- MTEB(Multilingual)
- MTEB(Europe) and MTEB(Indic)
- MTEB(eng, v2) = Faster version of MTEB(eng, v1)

Benchmark	Initial Scope	Refined Scope	Task Selection and Review
MTEB(Multilingual)	>500	343	132
MTEB(Europe)	420	228	74
MTEB(Indic)	55	44	23
MTEB(eng, v2)	56	54	41

Results Across Models: MTEB(Multi)

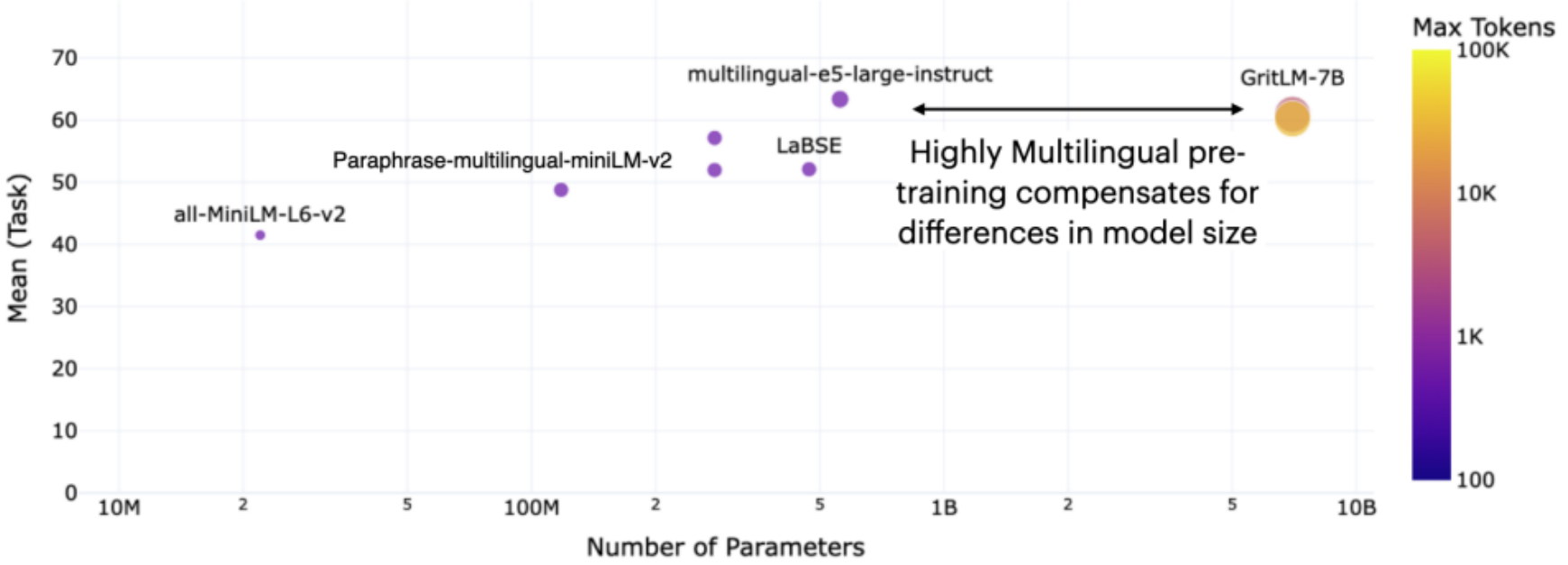


Figure 2: Mean performance across tasks on MTEB(Multilingual) according to the number of parameters. The circle size denotes the embedding size, while the color denotes the maximum sequence length of the model. To improve readability, only certain labels are shown. We refer to the public leaderboard for interactive visualization. We see that the notably smaller model obtains comparable performance to Mistral 7B and GritLM-7B, note that these overlap in the figure due to the similarity of the two models.

Results Across Tasks: MTEB(Multi)

Model (↓)	Rank (↓)	Average Across		Average per Category							
	Borda Count	All	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 Across Tasks: MTEB(Europe) vs. MTEB(Indic)

Model (↓)	Rank (↓) Borda Count	Average Across All	Average Across Category	Average per Category Btxt Pr Clf	Clf	STS	Rtrvl	M. Clf	Clus	Rrnk
MTEB(Europe)										
Number of datasets (→)	(74)	(74)	(74)	(7)	(6)	(21)	(9)	(15)	(2)	(6)	(3)
GritLM-7B	1 (737)	**63.0**	**62.7**	**90.4**	89.9	**64.7**	76.1	**57.1**	**17.6**	45.3	**60.3**
multilingual-e5-large-instruct	2 (732)	62.2	62.3	90.4	90.0	63.2	**77.4**	54.8	17.3	**46.9**	58.4
e5-mistral-7b-instruct	3 (725)	61.7	61.9	89.6	**91.2**	62.9	76.5	53.6	15.5	46.5	59.8
multilingual-e5-large	4 (586)	58.5	58.7	84.5	88.8	60.4	75.8	50.8	15.0	38.2	55.9
multilingual-e5-base	5 (499)	57.2	57.5	84.1	87.4	57.9	73.7	50.2	14.9	38.2	53.9
multilingual-mpnet-base	6 (463)	54.4	54.7	79.5	90.7	56.6	74.3	41.2	6.9	35.8	52.3
multilingual-e5-small	7 (399)	55.0	55.7	80.9	86.4	56.1	71.6	46.1	14.0	36.5	54.1
LaBSE	8 (358)	51.8	53.5	88.8	85.2	55.1	65.7	34.4	16.3	34.3	48.7
MTEB(Indic)										
Number of datasets (→)	(23)	(23)	(23)	(4)	(1)	(13)	(1)	(2)	(0)	(1)	(1)
multilingual-e5-large-instruct	1 (209)	**70.2**	**71.6**	**80.4**	76.3	**67.0**	**53.7**	**84.9**		**51.7**	**87.5**
multilingual-e5-large	2 (188)	66.4	65.1	77.7	75.1	64.7	43.9	82.6		25.6	86.0
multilingual-e5-base	3 (173)	64.6	62.6	74.2	72.8	63.8	41.1	77.8		24.6	83.8
multilingual-e5-small	4 (164)	64.7	63.2	73.7	73.8	63.8	40.8	76.8		29.1	84.4
GritLM-7B	5 (151)	60.2	58.0	58.4	67.8	60.0	27.2	79.5		28.0	84.7
e5-mistral-7b-instruct	6 (144)	60.0	58.4	59.1	73.0	59.6	23.0	77.3		32.7	84.4
LaBSE	7 (139)	61.9	59.7	74.1	64.6	61.9	52.8	64.3		21.1	79.0
multilingual-mpnet-base	8 (137)	58.5	55.2	44.2	**82.0**	61.9	34.1	57.9		32.1	74.3

Analysis

1. **Instruction-tuned-models outperformed non-instruction-tuned models across all task categories**
   - For many tasks, prompts are generic
   - No model-specific tuning of prompts was performed
2. **Differences in performance are likely due to pre-training and not finetuning datasets**
   - Multilingual-e5-large(-instruct) (560M) outperformed the much larger e5-mistral-7b-instruct (7.42B) and GritLM-7B (7.2B)
   - All three models were fined-tuned on similar multilingual datasets
   - XLM-R pretraining dataset contained 100+ languages
   - Mistral models are predominately pre-trained in English
3. **Larger models are not necessarily better at multi-lingual tasks**
   - XLM-Roberta-based multilingual-e5-large-instruct typically outperforms the larger GritLM-7B, e5-mistral-7b-instruct, multilingual-mpnet-base models on low-to mid-resource languages

Limitations

- **English or translated-text bias:** Some tasks contain datasets that were human-translated from English
- **Limited language representation:** Low-resource languages remain underrepresented across datasets

Future Directions

- **Creating high quality low-and mid-resource language datasets**
  - Specifically, those that contain nuanced and cultural information
  - LLMs can’t always capture the nuances of culture in the languages which takes away from the language understanding [c paper]
- **More tasks focused on the quality of the embeddings themselves**
  - Most tasks are correlation tasks which is further underscore by vast use of Spearman and Person correlation metrics
  - No tests as to how well the embeddings are even encoding the information