HUME - Measuring the Human-Model Performance Gap in Text Embedding Tasks

Adnan El Assadi, Isaac Chung, Roman Solomatin, Niklas Muennighoff, Kenneth Enevoldsen

ICLR

BUILT ON **MTEB**

What is HUME?

OVERVIEW

- Human evaluation of embedding models is critical but rarely done, **HUME fills this gap** with an extensible, protocol-aligned framework.
- We measure human performance across **16 MTEB datasets** spanning reranking, classification, clustering, and semantic similarity in **5 languages**.
- Humans achieve **77.6% avg. performance** vs. 80.1% for the best model, with striking variation across tasks and languages.
- Models dominate on structured tasks; **humans lead on non-English and culturally nuanced** ones.
- We benchmark **9 LLMs as annotators,** they fall short of human quality (76.1% vs. 81.2%).

16 DATASETS — 26 task-language pairs, MTEB-aligned

5 LANGUAGES — EN · AR · RU · DA · NO

13 EMBEDDING MODELS — 22M – 7B params

9 LLM ANNOTATORS — GPT · Gemini · Mistral · Qwen

Human Evaluation Framework For Text Embeddings

01 Dataset Selection — 16 MTEB datasets across 4 task types, balanced for linguistic diversity and complexity. (EN, AR, RU, DA, NO)

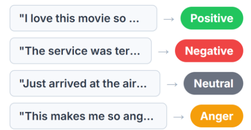
02 Annotation Interface — Task-specific Argilla interfaces matching MTEB protocols. Same labels, scales, and definitions. (Binary, Categorical, 0-5 Scale)

03 Human Annotators — NLP practitioners, native or near-native. 2 annotators per English task for IAA measurement. (Independent, Blind to GT)

04 Agreement Analysis — Task-appropriate IAA metrics to flag low-reliability datasets. (Fleiss' κ, ARI, Spearman ρ, Kendall τ)

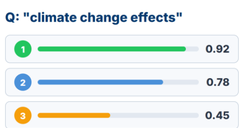
05 Model Comparison — 13 embedding models + 9 LLMs evaluated on identical instances with aligned metrics. (22M to 7B, Enc & Dec, Multilingual)

TASKS

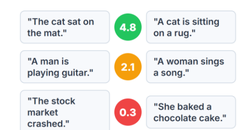


Classification
7 task-language pairs
40-48 SAMPLES · Accuracy METRIC

Clustering
7 task-language pairs
30 ITEMS · V-Measure METRIC



Reranking
6 task-language pairs
20-49 QUERIES · MAP METRIC



Semantic Similarity
6 task-language pairs
30-50 PAIRS · Spearman ρ METRIC

Table 1 : Human vs. 13 Embedding Models

Bold = highest (human or model) Underline = best model Green = human wins

	Classification				Clustering				Reranking			STS			Overall
MODEL	ARA	ENG	NOB	RUS	ARA	DAN	ENG	RUS	DAN	ENG	NOB	ARA	ENG	RUS	(26)
all-MiniLM-L6-v2	57.2	58.8	51.7	55.5	35.2	24.5	55.1	31.4	78.4	93.7	71.2	6.2	83.5	33.1	61.9
all-mpnet-base-v2	53.5	62.0	47.0	60.5	21.4	22.9	59.7	36.9	79.0	93.3	80.5	13.2	83.0	42.2	63.4
e5-mistral-7b-instruct	74.5	70.0	70.5	70.0	68.5	**76.0**	82.7	**77.7**	90.6	**96.4**	86.1	16.0	85.9	63.0	78.2
embedding-gemma-300m	71.0	58.6	54.0	73.5	19.2	43.7	65.1	36.9	74.3	86.9	71.4	36.7	69.9	66.2	64.2
gte-Qwen2-1.5B-instruct	75.2	76.5	70.8	74.5	73.7	67.1	75.9	72.2	84.3	95.3	87.8	28.8	84.0	54.2	77.5
jasper	63.5	**87.1**	70.5	79.8	64.3	54.7	83.2	43.7	90.1	95.8	90.0	40.9	88.1	**69.5**	**80.1**
multilingual-e5-base	75.8	64.7	73.8	77.2	35.9	40.6	45.6	36.2	92.2	94.4	87.5	31.0	85.2	62.7	68.2
multilingual-e5-large	77.0	64.9	75.0	80.0	34.6	31.0	52.5	46.9	**95.0**	95.3	**92.2**	33.8	86.3	68.8	70.4
multilingual-e5-small	72.2	62.2	69.2	81.2	35.5	38.0	51.7	59.1	88.6	94.2	88.3	28.8	85.2	60.3	69.0
mxbai-embed-large-v1	57.2	66.4	52.2	59.0	26.5	34.2	61.9	30.5	90.8	94.5	82.0	12.7	87.6	43.7	66.6
Qwen3-Embedding-0.6B	77.2	74.7	59.8	74.8	**78.8**	58.5	68.4	68.3	90.0	95.5	83.6	38.0	**88.5**	60.3	76.8
SFR-Embedding-Mistral	77.5	69.8	68.8	72.5	73.1	71.2	**85.1**	68.9	89.2	96.3	86.1	15.3	86.4	64.0	78.3
stella-en-1.5B-v5	65.8	84.0	67.0	79.2	36.8	42.6	78.6	46.7	91.7	96.0	88.6	37.2	86.7	62.1	76.9
Human	95.0	70.3	85.0	92.5	76.0	62.7	67.4	68.0	91.4	87.2	89.8	67.5	83.1	58.7	77.6

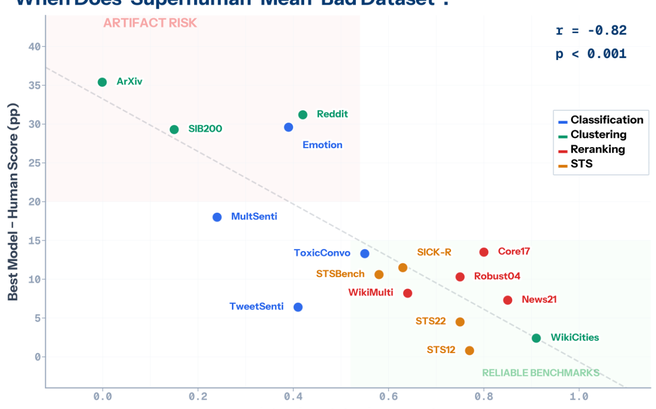
Table 2 : LLM-as-Annotator vs. Human

Bold = best LLM Clustering excluded (hard to elicit from generative models) All scores in %

Task	Human	GPT-5		GPT-4.1		Gemini	Mistral	Qwen3			Best Emb.
		Full	Mini	Full	Mini	2.5 Flash	Small-24B	30B	32B	Coder	Score (model)
Classification	**79.1**	**78.9**	77.2	76.6	76.1	77.6	73.8	74.2	73.0	76.3	80.3 (jasper)
Reranking	**88.3**	75.1	75.5	75.7	77.2	76.2	**78.0**	75.6	74.8	73.8	94.8 (e5-large)
STS	**76.5**	73.0	69.0	74.9	74.9	69.3	**75.0**	67.1	68.6	71.3	77.1 (jasper)
Average	**81.2**	**75.8**	**74.1**	**75.8**	**76.1**	**74.5**	**75.5**	**72.4**	**72.2**	**73.9**	--

- **Models achieve their largest advantages precisely where human annotators disagree most (r = -0.82).**
- **"Superhuman" scores on low-agreement tasks reflect artifact fitting, not genuine understanding.**

When Does 'Superhuman' Mean 'Bad Dataset'?



FINDINGS

1 Humans Rank 4th Overall
Score **77.6 avg**, outperforming 10 of 13 models, but trailing three large models.
Rank 4 / 14

2 "Superhuman" is Often Illusory
Models beat humans **where agreement is lowest**. High scores reflect artifact fitting, not understanding.
k=0.39 → 190%

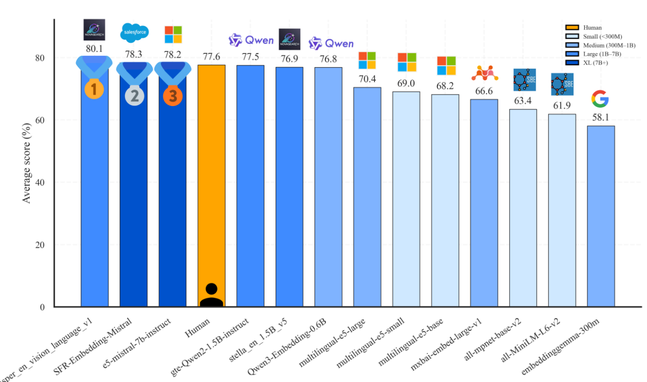
3 Humans Win Non-English Tasks
Arabic, Russian, Norwegian sentiment: humans dominate. **Arabic STS: +26.6 pts** over best model.
+26.6 pts gap

4 LLMs Can't Replace Humans Yet
Best LLM scores **76.1%** vs. human **81.2%**. On reranking, LLMs fall **10 pts short**.
81.2 vs 76.1

5 Task Quality = Benchmark Quality
Low-agreement tasks should be **deprecated from leaderboards**. Report IAA alongside scores.
Report IAA

6 "Human-Level" is Not One Number
Performance ranges from **45.8%** to **97.6%** across tasks. Task-level human baselines are essential.
45.8 – 97.6%

Human vs. Top Embedding Models



Paper arxiv.org/abs/2510.10062

Leaderboard huggingface.co/spaces/mteb/leaderboard