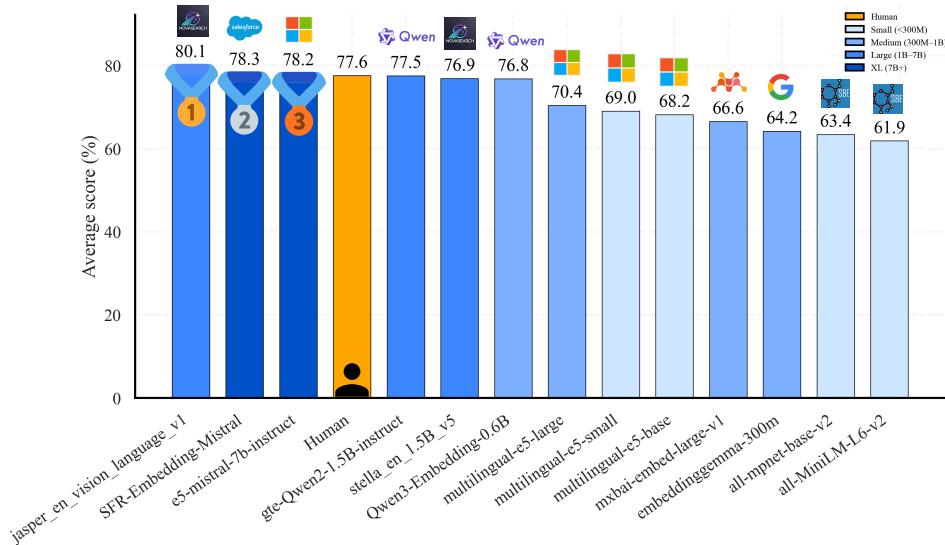

HUME: MEASURING THE HUMAN-MODEL PERFORMANCE GAP IN TEXT EMBEDDING TASKS

000
001
002
003
004
005 **Anonymous authors**
006 Paper under double-blind review
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029

ABSTRACT

030
031
032 Comparing human and model performance offers a valuable perspective for understanding the strengths and limitations of embedding models, highlighting where they succeed and where they fail to capture meaning and nuance. However, such comparisons are rarely made, as human performance on embedding tasks is difficult to measure. To fill this gap, we introduce HUME: **H**uman **E**valuation **F**ramework for **T**ext **E**mbeddings. While frameworks like MTEB provide broad model evaluation, they lack reliable estimates of human performance, limiting the interpretability of model scores. We measure human performance across 16 MTEB datasets spanning reranking, classification, clustering, and semantic textual similarity across linguistically diverse high- and low-resource languages. Humans achieve an average performance of 77.6% compared to 80.1% for the best embedding model, though with substantial variation: models reach high performance on some datasets while struggling on notably low-resource languages. Our human annotation also reveals multiple notable dataset issues. We also benchmark nine LLMs as annotators, finding they fall short of human performance (76.1% vs. 81.2%) despite offering scalability advantages. We provide human performance baselines, insights into task difficulty patterns, and an extensible evaluation framework that enables a more meaningful interpretation of the model and informs the development of both models and benchmarks. Our code, dataset, and leaderboard are publicly available at [anonymized_url](#).
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049



050 Figure 1: Overall ranking of human performance versus 13 embedding models across 16 tasks.
051 Human annotators achieve 4th place with a score of 77.6, demonstrating competitive but not dominant
052 performance. The ranking reveals significant variation in model performance across different
053 parameter scales and architectures. Darker shades of blue means a larger model.

054 1 INTRODUCTION
055
056

057 Embedding models are central to modern NLP systems, powering applications such as search, recom-
058 mendation, semantic analysis, and information retrieval. Many benchmarks test the performance of
059 embedding models, with the most comprehensive offering a diverse suite of tasks that test their gener-
060 ality and robustness (Muennighoff et al., 2022; Xiao et al., 2025; Enevoldsen et al., 2025). Despite
061 these advances, the interpretation and quality of these scores are often unclear as it is absent of human
062 performance references. Current metrics define performance in terms of theoretical maxima (e.g.,
063 $MAP = 1.0$) that assume perfect consensus on task outcomes. However, many NLP tasks inherently
064 involve ambiguity and disagreement (Plank, 2022), making a model’s score difficult to meaningfully
065 interpret without reasonable references. This interpretability gap has serious consequences. When
066 benchmarks reward models for fitting noisy labels, labels where even human annotators disagree, the
067 field risks “blind optimization”: expending R&D resources to replicate annotation artifacts rather
068 than achieve semantic progress. For instance, if a model achieves 0.85 MAP in reranking, it is unclear
069 whether this should be considered strong, mediocre, or beyond what annotators typically achieve.
070 This disconnect highlights the need for human-centered evaluation that contextualizes benchmark
071 results. Importantly, human performance should not be treated as an upper bound but as a diagnostic
072 signal: a way to understand where tasks are inherently noisy, where models may surpass typical
073 annotator agreement, and where model behaviour diverges from human judgment.

074 To address this, we introduce **HUME**: a **H**uman **E**valuation framework for text embedding tasks.
075 HUME evaluates annotator performance across four task categories: reranking, classification, clus-
076 tering, and semantic textual similarity (STS), using 16 diverse datasets from the Massive Text
077 Embedding Benchmark (MTEB) (Muennighoff et al., 2022), adapted for human annotation feasibility.
078 Through multi-annotator experiments, we analyze task difficulty, quantify variation across humans,
079 and compare results directly against state-of-the-art embedding models.

080 Our contributions are threefold: (1) a generalizable framework for human evaluation of embedding
081 tasks, (2) empirical evidence of how humans perform across diverse datasets and task types, and
082 (3) comparative analysis of models and humans that highlights strengths, weaknesses, and ambiguities
083 in both benchmarks and models that yield actionable insights. Together, these contributions establish
084 a foundation for human-aligned evaluation of embedding models and guide future benchmark design.

085 Beyond human evaluation, we also investigate whether Large Language Models (LLMs) can serve as
086 scalable proxies for human judgment. The promise of LLM-as-annotator approaches is compelling: if
087 LLMs can reliably replicate human judgments, they could enable large-scale benchmark development
088 at lower cost. However, this assumes LLMs capture the same semantic distinctions humans make,
089 rather than exhibiting systematic biases. We evaluate nine state-of-the-art LLMs (GPT-5, GPT-4.1,
090 Gemini, Mistral, and Qwen3 variants) on identical annotation tasks to assess their viability as human
091 proxies and identify task-specific limitations.

092 2 RELATED WORK
093

094 2.1 TEXT EMBEDDING MODELS AND MTEB

095 Text embedding models map natural language into dense vectors that capture semantic information.
096 They have progressed from static embeddings (Mikolov et al., 2013; Pennington et al., 2014) to
097 contextual encoders (Devlin et al., 2019; Liu et al., 2019) and more recently to models optimized for
098 embeddings like Sentence-BERT (Reimers & Gurevych, 2019), E5 (Wang et al., 2022), and GTE (Li
099 et al., 2023), powering applications such as search, classification, and clustering.

100 To evaluate these models across its diverse use-cases, MTEB (Muennighoff et al., 2022; Enevoldsen
101 et al., 2025) has risen as the de facto benchmark framework for embeddings and consolidates
102 evaluation across diverse tasks and datasets. Despite its breadth and community-driven extensions –
103 spanning multilingual, multimodal, and domain-specific variants (Xiao et al., 2025; Kasmaee et al.,
104 2025; Tang & Yang, 2025; Xiao et al., 2024; Ciancone et al., 2024; Enevoldsen et al., 2024; Zinvandi
105 et al., 2025; Wehrli et al., 2024; Poświatka et al., 2024; Snegirev et al., 2024), MTEB lacks human
106 performance baselines, making it difficult to contextualize model achievements.

108 2.2 HUMAN EVALUATION IN NLP
109

110 Human evaluation is well established in NLP, especially for generative tasks like machine translation
111 (Graham et al., 2013), summarization (Fabbri et al., 2021), and dialogue (Gupta et al., 2019). In
112 contrast, embedding-based tasks have relied almost exclusively on automated metrics, with little
113 attention to human baselines.

114 In information retrieval, initiatives such as TREC (Voorhees & Tice, 2000) collect human relevance
115 judgments, but these serve as gold standards rather than benchmarks of human performance under
116 model metrics (e.g., nDCG, MRR). Similarly, GLUE, SuperGLUE (Wang et al., 2018; 2019), and
117 MERA (Fenogenova et al., 2024) – a Russian GLUE-like benchmark – report human baselines, but
118 mainly for classification and reasoning tasks. For embeddings, works like STS (Cer et al., 2017)
119 report inter-annotator agreement, yet these are not converted into model-comparable performance
120 scores. This leaves a gap: human performance on embedding benchmarks such as MTEB remains
121 largely unquantified.

122
123 3 METHODOLOGY
124

125 3.1 FRAMEWORK DESIGN
126

127 Our framework builds on MTEB by establishing reproducible human evaluation protocols that align
128 directly with model evaluation. It consists of task-specific annotation interfaces, principled dataset
129 sampling, a standardized results format, and the use of aligned metrics. This design reveals where
130 evaluation practices introduce ambiguity or inconsistency.

131 3.2 TASKS, DATASETS, AND METRICS
132

133 Our selection criteria ensures comprehensive coverage across multiple dimensions: (1) **linguistic**
134 **diversity**: including both high-resource languages (English, Arabic, Russian) and lower-resource
135 languages (Norwegian Bokmål, Danish)¹ to test cross-lingual generalization, (2) **domain variety**:
136 spanning news, social media, encyclopedic content, scientific literature, and forum discussions
137 to capture real-world application diversity, (3) **construction methods**: including both curated
138 human annotations and synthetic dataset creation to understand how dataset origin affects human-
139 model alignment, (4) **task relevance**: using tasks from established benchmarks widely adopted
140 in the embedding evaluation community, and (5) **task complexity variation**: ranging from binary
141 classification to fine-grained similarity judgments. This systematic selection ensures our findings
142 generalize across the diverse landscape of embedding applications while maintaining direct relevance
143 to existing evaluation frameworks.

144 Each task category uses a primary evaluation metric to enable consistent human–model comparisons.
145 We summarize the datasets, their domains, and the primary metrics applied in Appendix A. Detailed
146 task examples are provided in Appendix C.

147 **Retrieval Proxy via Reranking** We use Reranking as a human-evaluable proxy for Information
148 Retrieval. Direct human evaluation of large-scale retrieval is methodologically infeasible—requiring
149 annotators to evaluate thousands of candidate documents per query. Reranking preserves the core
150 semantic challenge of discriminating query-candidate relevance while remaining tractable: humans
151 evaluate only the top-k candidates, establishing a baseline conceptually equivalent to embedding-
152 space behavior and ensuring human-model comparability.

154 3.3 INSTRUCTIONS
155

156 Human instructions are designed to match the task definitions exactly (e.g., identical label sets for
157 classification, same 1-5 scale for STS) to ensure valid comparisons. However, formal, detailed
158 annotation protocols are not publicly available for many MTEB datasets, which limits our ability
159 to verify alignment. To mitigate this, we designed instructions based on the original dataset papers’

160 ¹With 0 being "The Left-Behinds" and 5 being "The Winners", we cover eng: 5, ara: 5, rus: 4, dan: 3, nob: 1
according to the 0-5 scale by (Joshi et al., 2021).

162 task descriptions, ensuring annotators understood the semantic distinctions required for each task.
163 Instructions were piloted with a small subset before full annotation to identify and resolve ambiguities.
164

165 **3.4 ANNOTATION PROCEDURE**
166

167 Our annotations process follows a trend similar to recent embedding benchmarks Enevoldsen et al.
168 (2025); Xiao et al. (2025) focusing on a diverse set of tasks with fewer samples rather than large
169 singular tasks. We choose this approach as it allow us to better cover the broad scopes of current
170 benchmark. Annotators were recruited with a focus on cultural and language diversity.

171 All annotations are conducted in Argilla (Argilla Project Contributors, 2025) using task-specific inter-
172 faces: binary relevance for reranking, categorical labels for classification, free cluster ID assignment
173 for clustering, and 0–5 similarity scores for STS. Sample sizes balance task complexity: reranking
174 (20–49 queries), classification (40–48 examples), clustering (30 items), and STS (30–50 pairs).

175 All annotators were male, aged 20–35, from culturally diverse backgrounds, and experienced NLP
176 practitioners with native or near-native proficiency in the evaluated languages. They followed
177 structured guidelines and completed all annotations independently, without access to ground truth or
178 model predictions. The downsampled task subsets used for comparisons are included in the MTEB
179 package, with detailed task examples provided in Appendix C.

180 English tasks were annotated by two annotators to enable agreement analysis. Multilingual tasks were
181 annotated by a single annotator with corresponding language expertise. Inter-annotator agreement
182 was assessed with task-appropriate metrics: Fleiss’ kappa (Fleiss, 1971) for classification, pairwise
183 ARI (Strehl & Ghosh, 2003) for clustering, pairwise Spearman correlation (Agirre et al., 2012a) for
184 STS, and mean Spearman/Kendall’s tau (Manning et al., 2008) for reranking. A detailed agreement
185 analysis is provided in Appendix F. This controlled evaluation setup minimizes potential confounds
186 from dataset variation and enables direct performance comparisons on identical evaluation instances.
187

188 **3.5 MODEL SELECTION AND EVALUATION**
189

190 We evaluate 13 embedding models chosen to cover multiple dimensions: (1) **parameter scale**
191 (22M–7B), (2) **architecture** (encoder- and decoder-based), (3) **instruction tuning** (instruction-tuned
192 and standard), and (4) **multilingual capability** (English and multilingual). This selection spans
193 diverse computational budgets and training paradigms, capturing the current embedding landscape.
194 All evaluated models are provided in Appendix H.

195 All models are evaluated on the downsampled instances annotated by humans, using identical
196 metrics, protocols, and computational settings. Human performance is computed using the metrics
197 in Appendix A, mirroring MTEB protocols. For primary analyses, we report MAP for reranking,
198 Accuracy for classification, V-Measure for clustering, and Spearman correlation for STS.

199 To account for sample size constraints, we determine statistical significance using 95% confidence
200 intervals computed via Wilson Score Intervals (accuracy) and Fisher z -transformations (correlation),
201 as detailed in Appendix E.

202 **3.6 LLM-AS-ANNOTATOR EVALUATION**
203

204 To assess whether automated evaluation can proxy human judgment, we evaluate nine state-of-the-art
205 Large Language Models (LLMs) as annotators on the exact same tasks. We employ GPT-5 (full
206 and mini), GPT-4.1 (full and mini), Gemini 2.5 Flash, Mistral Small-24B-Instruct, and three Qwen3
207 variants (30B, 32B, Coder-30B), prompting them with identical instructions provided to human
208 annotators (see Appendix C).

209 LLMs receive the same task instances, evaluation metrics, and scoring protocols as human annotators,
210 enabling direct performance comparisons. For classification and STS tasks, LLMs provide categorical
211 labels or numerical similarity scores. For reranking, LLMs rank candidate documents by relevance to
212 the query. Clustering tasks were excluded from LLM evaluation due to fundamental difficulties in
213 eliciting consistent, structured cluster assignments from generative models.

214 This controlled setup determines whether LLMs can serve as scalable, low-cost proxies for human
215 evaluation or whether they exhibit systematic biases that limit their utility for benchmark development.

216 By evaluating LLMs on the same instances as humans and embedding models, we can directly
 217 compare their annotation quality and identify task-specific strengths and limitations.
 218

219 4 RESULTS AND ANALYSIS

220 Figure 1 provides an overview of human performance relative to 13 state-of-the-art embedding
 221 models across 16 tasks. Human annotators rank 4th overall with a score of 0.776, trailing 3 large
 222 models but outperforming 10 others. However, a raw ranking obscures the nuance of task difficulty
 223 and data reliability. As shown in Table 1, humans neither represent a uniform performance ceiling nor
 224 a lower bound, but rather occupy a middle ground that varies significantly by task category, language,
 225 and dataset quality.

226 We computed 95% confidence intervals for human performance using metric-appropriate methods
 227 (Wilson Score Intervals for classification accuracy, Fisher z -transformation for correlation-based
 228 metrics, and empirical annotator ranges for clustering and reranking). Models perform outside
 229 human CIs in 14 of 26 tasks ($p < 0.05$), often on datasets with low inter-annotator agreement where
 230 “superhuman” performance may reflect artifact fitting rather than genuine capability (see Appendix E
 231 for methodology and complete results). Below we analyze performance patterns by task category and
 232 language, with full per-task results in Appendix B.

Model	Classification				Clustering				Reranking				STS			Overall
	ara	eng	nob	rus	ara	dan	eng	rus	dan	eng	nob	ara	eng	rus		
Number of datasets	(1)	(4)	(1)	(1)	(1)	(1)	(4)	(1)	(1)	(4)	(1)	(1)	(4)	(1)	(1)	(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-mppnet-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	
embeddinggemma-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_en_vision_language_v1	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	99.5	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	

251 Table 1: Human performance compared to 13 embedding models across task categories and languages.
 252 **Bold** indicates highest performance (human or model), underline indicates best model performance.
 253 Humans achieve top performance in 5 of 14 *aggregated* task-language pairs, particularly excelling in
 254 non-English sentiment analysis and Arabic semantic similarity. Overall results are aggregated over
 255 the 26 task-language pairs.

256 4.1 PERFORMANCE PATTERNS BY TASK CATEGORY

257 Figure 2 shows human performance relative to model performance across all 26 task-language pairs.
 258 Each task shows human performance (point) positioned within the full spectrum from worst to best
 259 model performance (range bars). Humans consistently perform in the upper portion of model ranges,
 260 typically exceeding median model performance (61.5% of tasks) while rarely matching the best
 261 models (15.4% of tasks). Classification tasks show the strongest human performance, with humans
 262 outperforming all models in 3 of 7 tasks, while clustering and reranking reveal consistent gaps where
 263 humans fall short of top-performing models. Detailed gap analysis can be found in Appendix G.

264 **Classification:** Human performance averages 70.3, ranging from 45.8 on emotion classification
 265 ($\kappa = 0.39$, fair agreement) to 95.0 on Arabic sentiment analysis. Models generally exceed human
 266 performance (best: 87.1), but humans outperform models on non-English sentiment analysis, particu-
 267 larly in Arabic (95.0 vs. 77.5) and Russian (92.5 vs. 81.2), likely benefiting from native cultural and
 268 linguistic understanding that current models fail to capture.

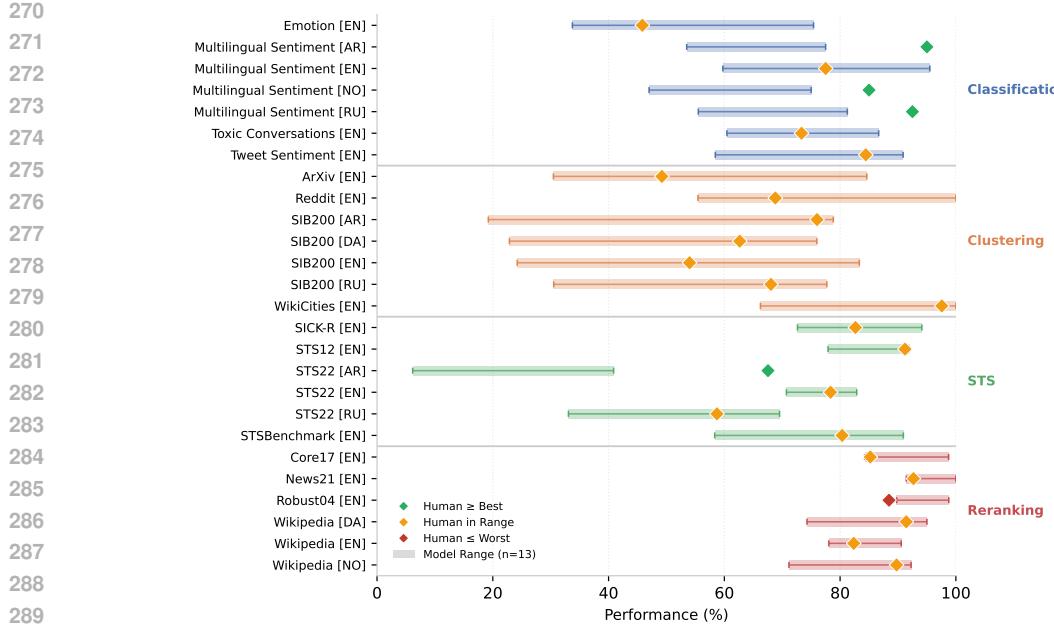


Figure 2: Comprehensive view of human performance relative to all model performance ranges across 16 tasks by language.

Clustering: Humans average 67.4 V-measure with extreme variation. Near-perfect performance on WikiCities (97.6, ARI = 0.91) contrasts sharply with poor performance on ArXiv papers (49.2, ARI = -0.001). Models consistently outperform humans (best: 85.1%). The poor inter-annotator agreement on ArXiv indicates fundamental task ambiguity rather than human limitation.

Reranking: Humans achieve strong performance (87.2 average MAP) with high inter-annotator agreement ($\rho = 0.64$ -0.85), demonstrating intuitive relevance understanding. Models exceed human performance (best: 96.4), but the high human agreement suggests these tasks align well with human judgment and provide reliable evaluation targets.

STS: Humans average 83.2 Spearman correlation, with notable variation: STS12 achieves 91.2 while STS22-Russian drops to 58.7, likely reflecting dataset quality issues discussed in §4.3. Models achieve comparable performance (best: 88.5), with moderate inter-annotator agreement ($\rho = 0.58$ -0.77).

These results reveal a critical insight that challenges conventional evaluation paradigms: human performance variation often reflects task quality rather than human limitations. Tasks with high human performance and agreement (reranking, toxicity classification) represent well-specified problems with clear ground truth, while tasks with low human agreement (emotion classification, academic clustering) may suffer from ambiguous annotation guidelines or inherently subjective judgments. Models achieve “superhuman” performance by reproducing consistent label patterns from training data, but this consistency may mask fundamental issues with task specification. Rather than treating low human performance as a ceiling to surpass, our findings suggest that it often signals the need for improved task design and clearer annotation frameworks.

4.2 CROSS-LINGUAL PERFORMANCE ANALYSIS

Human win rates vary substantially by language and comparison baseline (see Appendix G for detailed breakdown). Against the best models, humans win only 15% of tasks overall, but this rises to 62% against median models. The advantage is strongest for non-English tasks: humans achieve a 29% win rate on multilingual tasks versus 0% on English-only tasks against best models.

Arabic exhibits the strongest human advantage: a 67% win rate against the best models and 100% against the mean, with the largest gap in semantic similarity (67.5% vs. 40.9%, a 26.6-point margin).

324 Russian and Norwegian also show consistent human superiority in sentiment analysis, where humans
325 achieve 92.5% and 85.0%, respectively, substantially outperforming the best models. These advan-
326 tages likely stem from cultural and contextual knowledge that models fail to capture, especially in
327 lower-resource languages.

328 English tasks are more balanced, with models generally matching or exceeding humans, reflecting
329 the language’s dominance in training data. Danish shows mixed outcomes, possibly due to stronger
330 multilingual coverage for Germanic languages.
331

332 4.3 DATASET QUALITY AND EVALUATION CHALLENGES

333 Our analysis reveals systematic quality issues in several MTEB datasets that fundamentally com-
334 promise their reliability as evaluation benchmarks. Human performance variation often correlates
335 with underlying ambiguity rather than genuine human limitations, providing a diagnostic tool for
336 identifying problematic evaluation targets. See Appendix D for further qualitative analysis of these
337 failure modes; below, we highlight two such examples.:
338

339 **Emotion Classification Ambiguity:** The emotion classification dataset exemplifies inherent labeling
340 ambiguity, achieving only fair inter-annotator agreement ($\kappa = 0.39$) with 52.1% consensus. Real
341 examples demonstrate the fundamental challenges: “I feel like such a noob when the customers make
342 really dull and stupid jokes that im supposed to find funny” could reasonably be labeled as sadness (0),
343 anger (3), or even surprise (5) depending on interpretation. Similarly, “I am feeling very indecisive
344 and spontaneous” contains mixed emotional states that resist single-label categorization. Sarcastic
345 expressions like “I got paid too much because I get so many deliveries at work Im feeling a bit shamed”
346 present surface emotions that differ from intended meaning. When human experts fundamentally
347 disagree on correct answers for such inherently ambiguous cases, the apparent “superhuman” model
348 performance (87.1% vs. 45.8% human) likely reflects consistent reproduction of arbitrary majority
349 label patterns rather than superior emotional understanding.
350

351 **ArXiv Clustering Breakdown:** Academic paper clustering shows complete breakdown of human
352 agreement ($ARI = -0.001$), indicating fundamental disagreement about how to categorize academic
353 papers. Real examples illustrate the core ambiguity: papers like “Self-Supervised Audio-Visual
354 Representation Learning with Relaxed Cross-Modal Synchronicity” could legitimately cluster with
355 computer vision, machine learning, or audio processing groups depending on the annotator’s perspec-
356 tive on primary methodology versus application domain. “The architecture of innovation: Tracking
357 face-to-face interactions with ubicomp technologies” spans social science, computer science, and
358 architecture domains. Such interdisciplinary papers create fundamental disagreement about correct
359 clustering approaches, with no objectively correct answer. The task uses derived labels from ArXiv
360 categories, but the core issue is that academic papers often span multiple domains, making any
361 single clustering scheme inherently ambiguous. The high model performance (84.6% vs. 49.2%
362 human) suggests that the models are reproducing consistent labeling patterns rather than solving the
363 fundamental categorization challenge.
364

365 **High-Quality Benchmark Identification:** Conversely, tasks with high human agreement provide
366 reliable evaluation targets. Reranking tasks achieve strong inter-annotator agreement ($\rho = 0.64 - 0.85$)
367 with clear performance targets, while toxicity classification shows moderate agreement ($\kappa = 0.55$)
368 with 77.8% annotator consensus. These represent genuine evaluation challenges where model
369 improvements likely reflect meaningful progress rather than pattern matching to flawed labels.
370

371 These patterns suggest that apparent “superhuman” model performance often occurs precisely where
372 human agreement is lowest, indicating that models excel not through superior understanding but
373 through consistent reproduction of systematic labeling patterns. This raises concerns about the label
374 quality in embeddings benchmarks, and we encourage future benchmark developers to critically
375 examine the datasets before including them in a benchmark, potentially using human annotations
376 framework like HUME. Detailed analysis of specific quality issues is provided in Appendix D.
377

378 4.4 CAN LLMs REPLACE HUMAN ANNOTATORS?

379 We benchmark nine LLMs as annotators to assess whether they can serve as reliable proxies for
380 human judgment. As shown in Table 2, the best-performing LLM (GPT-4.1-mini) achieves 76.1%

Task	Human	GPT-5		GPT-4.1		Gemini	Mistral	Qwen3			Best Emb. Score
		Full	Mini	Full	Mini	2.5 Flash	Small-24B-I	30B	32B	Coder	
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)
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	–

Table 2: LLM-as-annotator performance compared to human annotators and best embedding models per task category. Human and LLM performance is computed over 19 task-language pairs (clustering tasks excluded due to difficulty eliciting cluster assignments). Best embedding model per task category shown with abbreviated name: jasper (jasper_en_vision_language_v1), SFR (SFR-Embedding-Mistral), e5 (multilingual-e5-large). **Bold** indicates best LLM performance (humans and embedding models consistently outperform LLMs and are not bolded).

average accuracy, falling short of human performance (81.2%). We exclude clustering tasks in this comparison due to the difficulty of eliciting cluster assignments from generative models.

Task-specific patterns reveal important limitations. On classification, LLMs approach human performance (GPT-5: 78.9% vs. Human: 79.1%). However, a substantial gap emerges in reranking, where humans achieve 88.3% compared to the best LLM at 78.0% (Mistral-Small): a 10-point deficit on tasks where humans show high agreement ($\rho = 0.64\text{--}0.85$). For STS, humans (76.5%) outperform all LLMs (best: 75.0%, Mistral-Small). Embedding models achieve the highest scores across all categories (Classification: 80.3%, Clustering: 79.1%, Reranking: 94.8%, STS: 77.1%). Detailed per-task LLM performance is provided in Appendix I.

To assess whether humans and LLMs face similar challenges, we computed Spearman rank correlations between human and LLM performance across the 19 task-language pairs. The moderate positive correlation ($\rho = 0.52$, $p < 0.05$, $n = 19$) indicates that tasks where humans perform well also tend to be tasks where LLMs perform well, suggesting partially shared difficulty patterns. However, the correlation is moderate rather than strong, indicating that humans and LLMs do not face identical challenges across all tasks (see § I.1 for detailed correlation analysis).

5 DISCUSSION

5.1 TASK QUALITY AND EMBEDDING EVALUATION RELIABILITY

Our analysis reveals a striking pattern: models achieve their highest relative performance precisely where human experts show the least agreement. This confirms that on low-quality datasets, current metrics do not measure semantic understanding but rather the model’s ability to reproduce consistent annotation artifacts.

This finding reframes the role of human evaluation in benchmark design. Rather than serving merely as a performance benchmark, human consensus establishes a *validity threshold* for evaluation tasks. When models significantly exceed this bound on low-agreement tasks, it signals that the benchmark itself has lost its descriptive power: the task may be measuring annotation artifacts rather than the capability it claims to assess. HUME provides the empirical mechanism to identify and deprecate these invalid evaluation targets, ensuring that leaderboards measure genuine capability rather than overfitting to noise.

Tasks with high human agreement, such as reranking and toxicity classification, provide reliable evaluation targets. Conversely, low-agreement tasks (e.g., ArXiv clustering, Emotion classification) suffer from ambiguous guidelines or subjective judgments. Cultural factors add another dimension to evaluation reliability. Humans retain substantial advantages in Arabic semantic similarity and multilingual sentiment analysis, revealing genuine model limitations in cross-cultural understanding.

These findings suggest reliable evaluation depends as much on task quality as model capability. Rather than treating high model performance as automatic progress, we recommend prioritizing high-agreement tasks for development, addressing cultural competence gaps, and critically examining whether apparent model superiority on ambiguous tasks reflects genuine capability or evaluation

432 bias. Novel benchmarks should report human agreement alongside model scores: 85% accuracy on
433 emotion classification (185% of human performance, $\kappa = 0.39$) represents a fundamentally different
434 achievement than 85% on reranking (97% of human performance, $\rho = 0.75$).
435

436 **5.2 LLM-BASED EVALUATION**
437

438 Our LLM annotation experiments reveal important limitations for using LLMs as proxies for human
439 judgment. While LLMs achieve competitive performance on classification tasks, a substantial
440 gap emerges on reranking, where humans significantly outperform even the best LLMs. Notably,
441 reranking tasks show strong inter-annotator agreement ($\rho = 0.64\text{--}0.85$), suggesting LLMs struggle
442 with precisely the nuanced relevance judgments where human consensus is highest. This pattern
443 contrasts with classification, where lower human agreement ($\kappa = 0.24\text{--}0.55$) coincides with near-
444 parity between humans and LLMs.

445 This has important implications for benchmark development. The reranking gap suggests that LLMs
446 may not reliably capture the semantic distinctions humans make on well-defined tasks, even as they
447 approach human performance on more ambiguous ones. Using LLMs for large-scale annotation
448 may therefore introduce systematic biases, particularly for tasks requiring fine-grained semantic
449 judgments. The architectural mismatch between generative LLMs and discriminative evaluation tasks
450 further limits their utility, as evidenced by our inability to evaluate clustering tasks.

451 While LLMs offer scalability advantages, these limitations suggest they should augment rather
452 than replace human annotation, particularly for benchmark development where task quality directly
453 impacts model development priorities. Future work should explore hybrid approaches that leverage
454 LLM efficiency for initial annotation while reserving human judgment for high-agreement tasks and
455 uncertain cases.

456 **5.3 IMPLICATIONS FOR MODEL DEVELOPMENT AND EVALUATION PRACTICES**
457

458 Our findings reveal concrete directions for both embedding model development and evaluation
459 methodology that address the fundamental quality issues we've identified.

460 **Prioritize High-Agreement Tasks for Development:** Development efforts should focus on tasks
461 where humans demonstrate both high performance and agreement, as these provide the most reliable
462 benchmarks for measuring genuine progress. Reranking tasks, with their clear performance targets
463 and strong agreement ($\rho = 0.64\text{--}0.85$), offer dependable evaluation where the persistent model-
464 human gap (96.4% vs. 87.2%) represents meaningful challenges requiring better modeling of
465 relevance relationships. Toxicity classification, despite moderate agreement ($\kappa = 0.55$), provides
466 another reliable target with 77.8% human consensus. In contrast, optimizing for tasks with poor
467 human agreement (emotion classification $\kappa = 0.39$, ArXiv clustering $ARI = -0.001$) may lead
468 models to excel at reproducing arbitrary labeling patterns rather than developing genuine semantic
469 capabilities.
470

471 **Address Cultural and Linguistic Competence Gaps:** The substantial human advantages in non-
472 English tasks reveal critical model limitations that scaling training data alone cannot address. Arabic
473 semantic similarity shows the largest human advantage (67.5% vs. 40.9% best model), while multilin-
474 gual sentiment demonstrates consistent human superiority in non-English languages (95.0% Arabic,
475 92.5% Russian). These gaps suggest that current models lack the cultural and contextual knowledge
476 necessary for cross-lingual understanding, requiring architectural innovations or training approaches
477 that go beyond simple data scaling to capture cultural nuances and contextual understanding.

478 **Replace Problematic Benchmark Datasets:** Our analysis identifies specific datasets that compro-
479 mise benchmark reliability and should be replaced in future MTEB iterations: emotion classification,
480 ArXiv clustering, and STS22-Russian (systematic parsing artifacts). These tasks provide unreliable
481 evaluation targets that may mislead model development efforts by rewarding pattern matching to
482 flawed gold standards. Replacement datasets should demonstrate reasonable human agreement and
483 clear task specifications, validated through human evaluation before inclusion in benchmark suites.

484 **Report Dataset Quality Measures:** Model performance should be interpreted in light of dataset
485 quality indicators to provide proper context for evaluation results. We propose that benchmark leader-
boards report human agreement metrics alongside model scores. A model achieving 85% accuracy

486 on emotion classification (185% of human performance, $\kappa = 0.39$) represents a fundamentally
487 different achievement than 85% on reranking (97% of human performance, $\rho = 0.75$). High model
488 performance on low-agreement tasks should be viewed skeptically as potential artifacts of flawed
489 evaluation targets rather than genuine capability improvements. For tasks where human agreement
490 falls below established thresholds ($\kappa < 0.4$ or $\rho < 0.6$), we recommend either improving task
491 specifications or removing the dataset from benchmark suites entirely. However, it is important to
492 recognize that some degree of human disagreement reflects natural variability in judgment rather than
493 dataset flaws. Future benchmarks could incorporate evaluation frameworks that preserve and leverage
494 this variability rather than collapsing it to single gold labels (Plank, 2022; Basile et al., 2021).

495

496 5.4 LIMITATIONS

497 Our study has several limitations. First, our prioritization of breadth over depth—covering 16 diverse
498 tasks—resulted in smaller sample sizes per task (20–50 instances). While we provide significance
499 analyses to validate our statistical conclusions, larger samples would better capture the full complexity
500 of human performance variation and provide more robust estimates of human judgment distributions.

501 Second, our multi-task design constrained task-specific training. Annotators were average or above-
502 average raters without specialized training; experts would likely perform better, particularly on
503 technical tasks. This was compounded by sparse annotation guidelines in original datasets, making
504 alignment with original procedures difficult—though this reflects realistic annotation scenarios where
505 perfect replication is often infeasible.

506 Third, while three annotators participated overall, only two annotations were collected for most tasks,
507 limiting our ability to fully characterize agreement patterns. Additionally, while we ensured cultural
508 and linguistic diversity among annotators, they were all male and aged 20–35, which does not fully
509 represent human judgment distributions across broader demographic groups.

510 Fourth, While our study quantifies where models diverge from human performance, it does not
511 fully explain why these gaps arise. Identifying the underlying factors - such as gaps in training data
512 coverage, domain or cultural biases, and linguistic variability, particularly in low-resource settings -
513 remains a critical direction for future research. However, detailed information about model training
514 corpora is often unavailable, limiting such analysis.

515 Finally, while our study evaluates human performance across diverse tasks, we did not systematically
516 investigate how task design features—such as specification clarity versus meaningful chal-
517 lenge—affect human agreement and model discrimination. The field needs rigorous research on these
518 design principles to avoid both ambiguity, which depresses human agreement, and oversimplification,
519 which diminishes discriminative power.

520

521 6 CONCLUSION

522 We introduce HUME, a comprehensive human evaluation framework for MTEB, addressing a critical
523 gap in understanding empirical performance bounds for embedding models. By measuring human
524 performance across 16 datasets spanning reranking, classification, clustering, and STS, we establish
525 statistically robust baselines that reframe how model achievements should be interpreted.

526 Our findings show that human performance varies substantially by task categories. Tasks with high
527 agreement provide reliable benchmarks, while low-agreement tasks often reveals design issues in the
528 task formulation.

529 Finally, our benchmarking of nine LLM-as-annotator systems demonstrates that while they offer
530 scalability, they cannot yet replace human judgment entirely. The best LLM (GPT-4.1-mini, 76.1%)
531 falls short of human performance (81.2%), particularly on reranking. This suggests that future
532 benchmarks should leverage LLMs to augment, but not replace, human evaluation.

533

534 6.1 ETHICAL CONSIDERATIONS

535 Annotators were co-authors who consented to the study. There were no external crowd workers
536 involved in any part of the annotation process.

540 REFERENCES

541

542 David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao,
543 Haonan Gao, and Annie En-Shiun Lee. Sib-200: A simple, inclusive, and big evaluation dataset
544 for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*, 2023.

545 Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot
546 on semantic textual similarity. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar,
547 Yuval Marton, and Deniz Yuret (eds.), **SEM 2012: The First Joint Conference on Lexical and
548 Computational Semantics – Volume 1: Proceedings of the main conference and the shared task,
549 and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval
550 2012)*, pp. 385–393, Montréal, Canada, 7–8 June 2012a. Association for Computational Linguistics.
551 URL <https://aclanthology.org/S12-1051/>.

552 Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: a pilot
553 on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and
554 Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task,
555 and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval
556 '12, pp. 385–393, USA, 2012b. Association for Computational Linguistics.

557 Argilla Project Contributors. Argilla: A collaboration tool for building high-quality ai datasets.
558 <https://argilla.io/>, 2025. Version as of Sep. 2025; open-source data labeling / curation
559 platform.

560 arXiv.org submitters. arxiv dataset, 2024. URL <https://www.kaggle.com/dsv/7548853>.

561 Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. XLM-T: Multilingual language
562 models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language
563 Resources and Evaluation Conference*, pp. 258–266, Marseille, France, June 2022. European
564 Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.27>.

565 Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo
566 Poesio, and Alexandra Uma. We need to consider disagreement in evaluation. In Kenneth Church,
567 Mark Liberman, and Valia Kordon (eds.), *Proceedings of the 1st Workshop on Benchmarking: Past,
568 Present and Future*, pp. 15–21, Online, August 2021. Association for Computational Linguistics.
569 doi: 10.18653/v1/2021.bppf-1.3. URL <https://aclanthology.org/2021.bppf-1.3>.

570 Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1:
571 Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard,
572 Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens
573 (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*,
574 pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi:
575 10.18653/v1/S17-2001. URL <https://aclanthology.org/S17-2001/>.

576 Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz,
577 Scott Hale, David Jurgens, and Mattia Samorøy. SemEval-2022 task 8: Multilingual news article
578 similarity. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer,
579 Nathan Schneider, Siddharth Singh, and Shyam Ratan (eds.), *Proceedings of the 16th International
580 Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1094–1106, Seattle, United States, July
581 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.155. URL
582 <https://aclanthology.org/2022.semeval-1.155>.

583 Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. Mteb-french: Resources
584 for french sentence embedding evaluation and analysis, 2024. URL <https://arxiv.org/abs/2405.20468>.

585 cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum.
586 Jigsaw unintended bias in toxicity classification, 2019. URL <https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification>.

594 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep
595 bidirectional transformers for language understanding. In *North American Chapter of the Association
596 for Computational Linguistics*, 2019. URL <https://api.semanticscholar.org/CorpusID:52967399>.

598 Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Nielbo. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems*, 2024. URL <https://nips.cc/virtual/2024/poster/97869>.

603 Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap,
604 Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala,
605 Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan
606 Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita
607 Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras,
608 Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu,
609 Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek
610 Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang,
611 Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale,
612 Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi,
613 Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria
614 Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide,
615 Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia
616 Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker,
617 Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. MMTEB:
618 Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi:
619 10.48550/arXiv.2502.13595. URL <https://arxiv.org/abs/2502.13595>.

620 Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and
621 Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the
622 Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL
623 <https://aclanthology.org/2021.tacl-1.24/>.

624 Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova,
625 Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana
626 Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin,
627 Polina Mikhailova, Denis Dimitrov, Alexander Panchenko, and Sergei Markov. Mera: A compre-
628 hensive llm evaluation in russian, 2024. URL <https://arxiv.org/abs/2401.04531>.

629 R. A. Fisher. Frequency distribution of the values of the correlation coefficients in samples from
630 an indefinitely large population. *Biometrika*, 10(4):507–521, 05 1915. ISSN 0006-3444. doi:
631 10.1093/biomet/10.4.507. URL <https://doi.org/10.1093/biomet/10.4.507>.

632 Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76
633 (5):378–382, 1971. doi: 10.1037/h0031619.

634 Wikimedia Foundation. Wikimedia downloads. URL <https://dumps.wikimedia.org>.

635 Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. Tweac: Transformer with
636 extendable qa agent classifiers. *arXiv preprint*, abs/2104.07081, 2021. URL <https://arxiv.org/abs/2104.07081>.

637 Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement
638 scales in human evaluation of machine translation. In *LAW@ACL*, 2013. URL <https://api.semanticscholar.org/CorpusID:1128384>.

639 Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P. Bigham.
640 Investigating evaluation of open-domain dialogue systems with human generated multiple refer-
641 ences, 2019. URL <https://arxiv.org/abs/1907.10568>.

642 Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218,
643 Dec 1985. doi: 10.1007/BF01908075. URL <https://doi.org/10.1007/BF01908075>.

648 Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM*
649 *Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418.
650 URL <https://doi.org/10.1145/582415.582418>.

651 Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and
652 fate of linguistic diversity and inclusion in the nlp world, 2021. URL <https://arxiv.org/abs/2004.09095>.

653 Ali Shiraei Kasmee, Mohammad Khodadad, Mohammad Arshi Saloot, Nicholas Sherck, Stephen
654 Dokas, Hamidreza Mahyar, and Soheila Samiee. Chemteb: Chemical text embedding benchmark,
655 an overview of embedding models performance & efficiency on a specific domain, 2025. URL
656 <https://arxiv.org/abs/2412.00532>.

657 Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread -
658 new fluffy embeddings model, 2024. URL <https://www.mixedbread.ai/blog/mxbai-embed-large-v1>.

659 Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

660 Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards
661 general text embeddings with multi-stage contrastive learning, 2023. URL <https://arxiv.org/abs/2308.03281>.

662 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,
663 Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach.
664 *ArXiv*, abs/1907.11692, 2019. URL <https://api.semanticscholar.org/CorpusID:198953378>.

665 Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information
666 Retrieval*. Cambridge University Press, 2008.

667 Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto
668 Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models.
669 In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard,
670 Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the
671 Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–
672 223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL
673 http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

674 Philip May. Machine translated multilingual sts benchmark dataset. 2021. URL <https://github.com/PhilipMay/stsb-multi-mt>.

675 Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-
676 embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog,
677 2024. URL <https://www.salesforce.com/blog/sfr-embedding/>.

678 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representa-
679 tions of words and phrases and their compositionality. In *Proceedings of the 27th International
680 Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119, Red
681 Hook, NY, USA, 2013. Curran Associates Inc.

682 Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. Cross-lingual transfer learning with {P}ersian.
683 In Lisa Beinborn, Koustava Goswami, Saliha Murado
684 uglu, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Edoardo M. Ponti, Ryan Cot-
685 terell, and Ekaterina Vylomova (eds.), *Proceedings of the 5th Workshop on Research in Com-
686 putational Linguistic Typology and Multilingual NLP*, pp. 89–95, Dubrovnik, Croatia, May
687 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigtyp-1.9. URL
688 <https://aclanthology.org/2023.sigtyp-1.9>.

689 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding
690 benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

691 Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the
692 Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

702 Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word
703 representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the*
704 *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543,
705 Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.
706 URL <https://aclanthology.org/D14-1162/>.

707 Barbara Plank. The “problem” of human label variation: On ground truth in data, modeling and
708 evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the*
709 *2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682,
710 Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguis-
711 tics. doi: 10.18653/v1/2022.emnlp-main.731. URL <https://aclanthology.org/2022.emnlp-main.731/>.

712 Rafał Poświata, Sławomir Dadas, and Michał Perelkiewicz. Pl-mteb: Polish massive text embedding
713 benchmark. *arXiv preprint arXiv:2405.10138*, 2024.

714 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.
715 In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL <https://api.semanticscholar.org/CorpusID:201646309>.

716 Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster
717 evaluation measure. In Jason Eisner (ed.), *Proceedings of the 2007 Joint Conference on Empir-
718 ical Methods in Natural Language Processing and Computational Natural Language Learning
719 (EMNLP-CoNLL)*, pp. 410–420, Prague, Czech Republic, June 2007. Association for Compu-
720 tational Linguistics. URL <https://aclanthology.org/D07-1043/>.

721 Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER:
722 Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang,
723 Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical
724 Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, 2018. Association
725 for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL <https://aclanthology.org/D18-1404>.

726 Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov.
727 The russian-focused embedders’ exploration: rumteb benchmark and russian embedding model
728 design, 2024. URL <https://arxiv.org/abs/2408.12503>.

729 Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classifi-
730 cation tasks. *Information Processing & Management*, 45(4):427–437, 2009. ISSN 0306-4573.
731 doi: <https://doi.org/10.1016/j.ipm.2009.03.002>. URL <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.

732 C Spearman. The proof and measurement of association between two things. *International Journal
733 of Epidemiology*, 39(5):1137–1150, 10 2010. ISSN 0300-5771. doi: 10.1093/ije/dyq191. URL
734 <https://doi.org/10.1093/ije/dyq191>.

735 Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for
736 combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617, March 2003. ISSN
737 1532-4435. doi: 10.1162/153244303321897735. URL <https://doi.org/10.1162/153244303321897735>.

738 Yixuan Tang and Yi Yang. Finmteb: Finance massive text embedding benchmark, 2025. URL
739 <https://arxiv.org/abs/2502.10990>.

740 Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram
741 Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi,
742 Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen,
743 Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao
744 Zheng, Jyotinder Singh, Abheesht Sharma, Divya Sreepat, Aashi Jain, Adham Elarabawy, AJ Co,
745 Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca,
746 Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábreo, Hesen Zhang, Hui
747 Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul
748

756 Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie
757 Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram
758 Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong,
759 Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon,
760 Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts,
761 Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba
762 Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations, 2025. URL
763 <https://arxiv.org/abs/2509.20354>.

764 Nguyen Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings
765 comparison: Variants, properties, normalization and correction for chance. *Journal of Machine
766 Learning Research*, 11:2837–2854, 10 2010.

767 Ellen Voorhees and D Tice. Building a question answering test collection. (34), 2000-07-01 2000.

768 Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE:
769 A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen,
770 Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop Black-
771 boxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium,
772 November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL
773 <https://aclanthology.org/W18-5446/>.

774 Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer
775 Levy, and Samuel R. Bowman. *SuperGLUE: a stickier benchmark for general-purpose language
776 understanding systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

777 Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Dixin Jiang, Rangan
778 Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training.
779 *ArXiv*, abs/2212.03533, 2022. URL <https://api.semanticscholar.org/CorpusID:254366618>.

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

782 Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. German text embedding clustering benchmark,
783 2024. URL <https://arxiv.org/abs/2401.02709>.

784 Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme,
785 Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models
786 to follow instructions, 2024.

787 Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Statist.
788 Assoc.*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

789 Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman
790 Solomatín, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive
791 image embedding benchmark, 2025. URL <https://arxiv.org/abs/2504.10471>.

792 Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack:
793 Packed resources for general chinese embeddings, 2024. URL <https://arxiv.org/abs/2309.07597>.

794 Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota
795 embedding models, 2025a. URL <https://arxiv.org/abs/2412.19048>.

796 Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun
797 Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embed-
798 ding: Advancing text embedding and reranking through foundation models, 2025b. URL
799 <https://arxiv.org/abs/2506.05176>.

800 Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi,
801 and Arash Amini. Famteb: Massive text embedding benchmark in persian language. *arXiv preprint
802 arXiv:2502.11571*, 2025.

810 A DATASET SPECIFICATIONS

812 Table 3 details the 16 datasets selected for this study, including their domains, descriptions, and the
 813 primary metrics used for evaluation.

	Datasets	Description	Metrics
816 Ranking	Core17, News21, Robust04 (Weller et al., 2024)	Information retrieval benchmarks (news, documents)	MAP, MRR@10, nDCG@10
817	WikiMulti (Enevoldsen et al., 2025)	Wikipedia article reranking (eng, dan, nob)	nDCG@10
818			
819 Classification	Emotion (Saravia et al., 2018) Tweet Senti (Barbieri et al., 2022)	Emotion classification from social media text Sentiment analysis of tweets	Accuracy, F1, Weighted F1
820			
821 Toxicity (cjadams et al., 2019)	Toxic content detection		
822 Multilingual Sentiment	Sentiment classification (ara, eng, nob, rus)		
823 (Mollanorozy et al., 2023)			
824			
825 Clustering	WikiCities (Foundation) ArXiv (arXiv.org submitters, 2024)	Entity clustering from Wikipedia Academic paper topic clustering (derived labels)	V-Measure, ARI, AMI
826			
827 Reddit (Geigle et al., 2021)	Forum discussion topic clustering		
828 SIB200 (Adelani et al., 2023)	Multilingual sentence clustering (ara, dan, eng, rus)		
829			
830 STS	STSBenchmark (May, 2021) SICK-R (Marelli et al., 2014)	General semantic similarity benchmark Semantic relatedness and entailment	
831			
832 STS12 (Agirre et al., 2012b)	Shared task semantic similarity		Spearman, Pearson
833 STS22 (Chen et al., 2022)	Multilingual semantic similarity (ara, eng, rus)		

834 Table 3: Complete list of 16 datasets and evaluation metrics used for human annotation. For the
 835 metrics we use MAP (Manning et al., 2008), MRR (Manning et al., 2008), nDCG (Järvelin &
 836 Kekäläinen, 2002), Accuracy/F1 (Sokolova & Lapalme, 2009), V-Measure (Rosenberg & Hirschberg,
 837 2007), ARI (Hubert & Arabie, 1985), AMI (Vinh et al., 2010), Spearman/Pearson (Spearman, 2010)
 838 (Pearson, 1895), following the MTEB implementations.

841 B DETAILED RESULTS BY TASK CATEGORY

843 This section provides comprehensive results for all tasks, organized by category. Each table includes
 844 human performance alongside all 13 evaluated models, with inter-annotator agreement metrics where
 845 available.

846 Table 4 presents full results of the clustering tasks. Table 5 presents full results of the classification
 847 tasks. Table 6 presents full results of the reranking tasks. Table 7 presents full results of the STS
 848 tasks.

850 C TASK EXAMPLES

853 This section provides screenshots of the actual Argilla annotation interfaces used in our study,
 854 illustrating the annotation challenges and interface design that human annotators encountered.



861 Figure 3: Emotion Classification annotation interface showing the 6-category emotion labeling task.
 862 This task achieved fair inter-annotator agreement ($\kappa = 0.39$) due to ambiguous emotional states and
 863 mixed emotions in social media text. Human performance: 45.8%, Best model: 87.1%.

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880

Figure 4: Tweet Sentiment Classification annotation interface demonstrating sentiment polarity annotation. This task achieved moderate inter-annotator agreement ($\kappa = 0.48$) with reasonable consensus on positive/negative sentiment. Human performance: 84.4%, Best model: 90.9%.

881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

Figure 5: ArXiv Clustering annotation interface showing academic papers that caused complete annotator disagreement ($ARI = -0.001$) due to interdisciplinary research overlap. Papers could be categorized by methodology, application domain, or research community, leading to fundamental disagreement. Human performance: 49.2%, Best model: 84.6%.

Figure 6: Reddit Clustering annotation interface demonstrating thematic grouping of discussion topics. This task achieved fair agreement ($ARI = 0.34$) due to overlapping themes across different discussion topics. Human performance: 68.8%, Best model: 100%.

Figure 7: SIB200 Clustering annotation interface showing multilingual sentence clustering task. This task achieved moderate inter-annotator agreement ($ARI = 0.42$) with variation across languages depending on cultural context and sentence complexity.

Figure 8: Robust04 Reranking annotation interface showing document relevance assessment for information retrieval queries. This task achieved strong inter-annotator agreement ($\rho = 0.72$). Human performance: 88.5%, Best model: 98.8%.

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933

Query
What are the critical reviews and ratings for the film Existenz?

Document 1
Conversely, James Berardinelli gave the film a two out of four star rating in his review. He cites that the film had a "disjointed feel", and called it a "missed opportunity" that suffered from being released near *The Matrix* and *Open Your Eyes*, which he states did similar things that were accomplished better in those films.

Document 2
The film received generally positive reviews, with a 74% approval rating at Rotten Tomatoes based on 74 reviews, with an average rating of 6.60/10. The site's critical consensus reads, "Ooopy, slimy, grotesque fun." Metacritic assigned a score of 68 out of 100, based on 29 critics, indicating "generally favorable reviews".

Document 3
Roger Ebert gave the film three stars out of four in his review of the film, noting its release after fellow science-fiction film *The Matrix*. He compared the two films, stating that while both have special effects, Cronenberg's film was stranger along with having his best effects involve "gooey, indescribable organic things".

Document 4
Nominated, Best Sound Editing in a Foreign Feature: David Evans, Wayne Griffin, Mark Gingras, John Laing, Tom Bjelic, and Paul Shikata

Document 1 Relevance * Relevant Not Relevant
Document 2 Relevance * Relevant Not Relevant
Document 3 Relevance * Relevant Not Relevant
Document 4 Relevance * Relevant Not Relevant
Document 5 Relevance * Relevant Not Relevant
Document 6 Relevance * Relevant Not Relevant
Document 7 Relevance * Relevant Not Relevant

Discard Save as draft Submit

934 Figure 9: Wikipedia Multilingual Reranking annotation interface demonstrating cross-lingual relevance judgment. This task achieved moderate agreement ($\rho = 0.64$) due to cross-lingual complexity.
935

936
937
938
939
940
941
942
943
944
945
946
947
948

Sentence 1
twice the quantity or amount of something

Sentence 2
a quantity that is twice as great as another.

Similarity Score * 3
0 1 2 3 4 5

Discard Save as draft Submit

949 Figure 10: STS12 annotation interface showing semantic similarity assessment using a 0-5 scale.
950 This well-curated dataset achieved strong inter-annotator agreement ($\rho = 0.77$). Human performance:
951 91.2%, Best model: 92.0%.
952

953
954
955
956
957
958
959
960
961
962
963
964
965

Sentence 1
A woman is putting on eyeshadow

Sentence 2
The woman is removing make-up

Similarity Score * 3
1 2 3 4 5

Discard Save as draft Submit

966 Figure 11: SICK-R annotation interface showing semantic relatedness and entailment task. This task
967 achieved moderate agreement ($\rho = 0.68$) due to task complexity. Human performance: 82.6%, Best
968 model: 94.1%.
969

970
971

Model	Arxiv	Reddit	SIB200	WikiCities
all-MiniLM-L6-v2	61.8	56.5	33.2	73.8
all-mpnet-base-v2	56.0	63.0	33.2	86.7
e5-mistral-7b-instruct	77.2	74.9	73.8	100.0
embeddinggemma-300m	73.0	72.8	36.0	70.3
gte-Qwen2-1.5B-instruct	70.8	88.4	72.3	73.6
jasper_en_vision_language_v1	83.9	95.1	59.0	95.1
multilingual-e5-base	31.9	60.0	34.4	66.2
multilingual-e5-large	51.1	55.4	37.8	69.2
multilingual-e5-small	30.5	76.9	38.7	72.6
mxbai-embed-large-v1	48.3	72.1	33.0	90.8
Qwen3-Embedding-0.6B	69.1	76.3	65.5	74.8
SFR-Embedding-Mistral	75.5	81.4	74.8	100.0
stella_en_1.5B_v5	84.6	100.0	44.2	86.4
Human	49.2	68.8	65.2	97.6

Table 4: Full clustering results.

Model	Emotion	MultilSenti	ToxicConvo	TweetSenti
all-MiniLM-L6-v2	42.1	57.9	66.4	59.6
all-mpnet-base-v2	44.0	61.5	60.4	58.4
e5-mistral-7b-instruct	47.3	75.9	68.0	75.8
embeddinggemma-300m	40.4	64.6	66.4	67.8
gte-Qwen2-1.5B-instruct	56.2	78.1	78.7	79.3
jasper_en_vision_language_v1	75.4	77.3	86.7	90.9
multilingual-e5-base	36.2	78.6	66.0	69.1
multilingual-e5-large	38.5	81.1	63.3	65.3
multilingual-e5-small	33.8	75.8	65.1	69.6
mxbai-embed-large-v1	42.1	64.4	68.7	65.8
Qwen3-Embedding-0.6B	51.9	74.1	74.4	87.8
SFR-Embedding-Mistral	46.7	77.1	67.3	75.6
stella_en_1.5B_v5	71.9	76.1	82.9	89.1
Human	45.8	87.5	73.3	84.4

Table 5: Full Classification results.

D TASK QUALITY ANALYSIS

D.1 DATASET QUALITY ISSUES

Our analysis revealed quality issues across multiple datasets that significantly impact human-model performance comparisons. These issues fall into several categories that help explain performance patterns observed in our study.

D.1.1 STS22-RUSSIAN

The Russian subset of “STS22” dataset shows patterns that may help explain the comparatively low human agreement we observed.

1026	Model	Core17	News21	Robust04	Wikipedia
1027	all-MiniLM-L6-v2	95.6	98.8	96.3	77.8
1028	all-mpnet-base-v2	98.6	98.6	97.8	79.2
1029	e5-mistral-7b-instruct	98.8	99.5	98.8	88.4
1030	embeddinggemma-300m	84.2	91.4	89.8	76.0
1031	gte-Qwen2-1.5B-instruct	97.5	99.2	98.5	86.1
1032	jasper_en_vision_language_v1	98.2	100.0	98.7	88.8
1033	multilingual-e5-base	96.2	98.6	96.9	88.5
1034	multilingual-e5-large	95.7	97.8	97.2	92.6
1035	multilingual-e5-small	95.6	98.1	97.5	87.6
1036	mxbai-embed-large-v1	97.2	98.0	98.6	85.6
1037	Qwen3-Embedding-0.6B	97.0	100.0	98.5	86.8
1038	SFR-Embedding-Mistral	97.9	99.7	98.8	87.9
1039	stella_en_1.5B_v5	98.6	100.0	98.3	89.2
1040	Human	85.2	92.7	88.5	87.9

1043 Table 6: Full Reranking results.
 1044

1045	Model	SICK-R	STS12	STS22	STSB
1046	all-MiniLM-L6-v2	91.5	85.7	48.4	81.8
1047	all-mpnet-base-v2	89.8	83.7	54.3	78.4
1048	e5-mistral-7b-instruct	93.2	89.1	58.5	85.9
1049	embeddinggemma-300m	72.6	77.9	57.9	58.3
1050	gte-Qwen2-1.5B-instruct	93.4	86.9	60.3	80.0
1051	jasper_en_vision_language_v1	93.8	92.0	67.2	88.7
1052	multilingual-e5-base	91.5	86.5	63.3	82.9
1053	multilingual-e5-large	89.4	89.9	65.9	83.6
1054	multilingual-e5-small	88.6	87.6	63.3	81.9
1055	mxbai-embed-large-v1	93.4	91.1	53.9	88.9
1056	Qwen3-Embedding-0.6B	93.3	91.8	63.6	90.9
1057	SFR-Embedding-Mistral	94.1	89.2	59.3	86.4
1058	stella_en_1.5B_v5	92.3	89.1	64.3	87.1
1059	Human	82.6	91.2	68.2	80.4

1060 Table 7: Full STS results.
 1061

1062 **Context Expansion Issues:**

1063

- 1064 Sentence pairs labeled as “4” (identical meaning) where one sentence contains basic information and the paired sentence includes additional backstory or context
- 1065 Translated example pattern: “Company reports earnings” vs. “Company reports earnings of \$X million, exceeding expectations due to strong performance in sector Y”
- 1066 Human annotators correctly identify these as semantically different (similarity 2-3), while gold labels mark them as identical

- This explains the low human performance on STS22-Russian (58.5%) compared to models (69.5%)

Parsing and Processing Errors:

- Incomplete sentence parsing affecting semantic interpretation
- Parsing artifacts from web pages (e.g., page numbers, lists of automatically generated related news, ads)

D.1.2 MULTILINGUAL SENTIMENT CLASSIFICATION-RUSSIAN

The Russian subset of “MultilingualSentimentClassification” consists of news articles from different news sites. The task is to classify each text as “positive” or “negative”. However, the dataset presents several challenges:

Neutral and Ambiguous Content:

- Many samples are based on press releases from companies or government departments, which are often neutral in tone and difficult to categorize as positive or negative.
- Translated example: “The total amount of pension savings accumulated in JSC ‘Unified Accumulative Pension Fund’ (UAPF) as of September 1, 2016, amounted to about 6.41 trillion tenge, the press center of the pension fund said, KazTAG reports. ...”
- Such sentences are more informative than sentiment-bearing.

Parsing and Processing Errors:

- Similar to the issues described in § D.1.1, the dataset contains parsing artifacts from web pages.

D.1.3. EMOTION CLASSIFICATION

The emotion classification dataset suffers from inherent label ambiguity that explains the low human agreement ($\kappa = 0.39$):

Mixed Emotional States:

- Texts expressing multiple emotions simultaneously: “i am feeling very indecisive and spontaneous” (labeled as fear but could be surprise)
- “i was feeling pretty anxious all day but my first day at work was a very good day and that helped a lot” (contains both fear and joy)
- “i am feeling crampy and cranky” (physical discomfort mixed with anger)

Sarcastic and Ironic Expressions:

- “i got paid too much because i get so many deliveries at work im feeling a bit shamed so will curb the spending for a bit” (sarcasm about being “overpaid”)
- “i feel like such a noob when the customers make really dull and stupid jokes that im supposed to find funny” (surface sadness but underlying anger/frustration)
- “i feel very cheated since i am supporting the family and doing all the other stuff while he spends hours a day gaming” (labeled as joy but clearly expressing anger)

Contextual and Subjective Interpretation:

- “i feel shame in a strange way” (ambiguous emotional context, labeled as surprise)
- “i feel all funny sometimes” (vague emotional description that could be multiple categories)
- “i feel underappreciated and under valued” (could be sadness, anger, or fear depending on interpretation)

1134
1135

D.1.4 ARXIV CLUSTERING CHALLENGES

1136
1137
1138

Academic paper clustering presents fundamental categorization difficulties that explain the complete breakdown of human agreement ($ARI = -0.001$). This task uses derived labels from ArXiv paper categories:

1139
1140

Interdisciplinary Research Papers:

1141
1142
1143

- “Self-Supervised Audio-Visual Representation Learning with Relaxed Cross-Modal Synchronicity” (could cluster with computer vision, audio processing, or self-supervised learning)
- “The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies” (spans social science, computer science, and architecture)
- “PIINET: A 360-degree Panoramic Image Inpainting Network Using a Cube Map” (computer vision, graphics, or deep learning focus)

1144
1145
1146
1147

Methodological vs. Application Domain Confusion:

1148
1149

- “Convergent Actor-Critic Algorithms Under Off-Policy Training and Function Approximation” (reinforcement learning methodology vs. control theory application)
- “Learning-Based Adaptive IRS Control with Limited Feedback Codebooks” (machine learning method vs. wireless communications application)
- “Structure-preserving numerical methods for stochastic Poisson systems” (numerical methods vs. mathematical physics)

1150
1151
1152
1153
1154
1155
1156

Emerging and Cross-Domain Research:

1157
1158
1159
1160
1161
1162
1163
1164

- “The modularity of action and perception revisited using control theory and active inference” (cognitive science, control theory, or neuroscience)
- “Food-chain competition influences gene’s size” (evolutionary biology, computational biology, or mathematical modeling)
- “Wavelet Analysis of Dengue Incidence and its Correlation with Weather and Vegetation Variables in Costa Rica” (epidemiology, signal processing, or environmental science)

1165
1166

D.2 IMPACT ON EVALUATION

1167
1168

These quality issues have several critical implications for embedding evaluation:

1169
1170
1171
1172

1. **Artificial Model Advantages:** Models may achieve “superhuman” performance by consistently reproducing systematic labeling patterns rather than demonstrating superior semantic understanding. This is particularly evident in tasks with low human agreement where models can exploit consistent but incorrect labeling patterns.
2. **Misleading Benchmarks:** Tasks with fundamental quality issues provide unreliable targets for model development. High model performance on such tasks may not indicate genuine capability improvements but rather successful pattern matching to flawed gold standards.
3. **Cultural and Linguistic Bias:** Quality issues disproportionately affect non-English tasks, potentially masking genuine model limitations in cross-cultural understanding while artificially inflating performance on problematic English datasets.
4. **Evaluation Validity:** The validity of using these datasets as benchmarks is questionable when human experts cannot agree on correct answers, suggesting fundamental issues with task specification rather than human limitations.

1182
1183
1184

E STATISTICAL ROBUSTNESS ANALYSIS

1185
1186
1187

E.1 CONFIDENCE INTERVAL METHODOLOGY

Given sample size constraints ($N = 20 - 50$), we computed 95% Confidence Intervals (CIs) for human performance using metric-specific analytical methods rather than generic approximations:

- **Classification (Accuracy):** Wilson Score Interval (Wilson, 1927), which is robust for binomial proportions with small sample sizes and avoids the “zero-width” errors of normal approximations.
- **STS (Correlation):** Fisher z -transformation (Fisher, 1915) to compute CIs for Spearman correlations, ensuring valid bounds within $[-1, 1]$.
- **Clustering & Reranking:** Empirical range between annotators as a conservative bound given $N_{\text{annotators}} = 2$.

Statistical significance (*) is determined by a non-parametric overlap test: a model is considered significantly different if its score falls outside the human 95% CI (corresponding to $p < 0.05$).

E.2 DETAILED RESULTS WITH CONFIDENCE INTERVALS

Table 8 presents human performance with 95% confidence intervals for all 26 task-language pairs. Models perform outside the human confidence interval in 14 of 26 tasks, indicating statistically significant differences. Crucially, this separation often occurs in tasks with low to moderate inter-annotator agreement. For example, in EmotionClassification, the best model (75.4) significantly exceeds human performance (45.8, CI:[32.6, 59.7]), but the low agreement ($\kappa = 0.39$) suggests this “superhuman” performance may reflect fitting to annotation artifacts rather than genuine semantic superiority. Conversely, in MultilingualSentiment (Arabic) and STS22 (Arabic), humans significantly outperform models, highlighting genuine cultural gaps that current models fail to bridge.

Task	Lang	N	K	Tot	Human (95% CI)	IAA	Best Model
<i>Classification</i>							
EmotionClassification	eng	48	2	96	45.8 [32.6, 59.7]	$\kappa = 0.39$	75.4*
MultilingualSentiment	ara	40	1	40	95.0 [83.5, 98.6]	N/A	77.5*
MultilingualSentiment	eng	40	2	80	77.5 [62.5, 87.7]	$\kappa = 0.24$	95.5*
MultilingualSentiment	nob	40	1	40	85.0 [70.9, 92.9]	N/A	75.0
MultilingualSentiment	rus	40	1	40	92.5 [80.1, 97.4]	N/A	81.3
ToxicConversations	eng	45	2	90	73.3 [59.0, 84.0]	$\kappa = 0.55$	86.7*
TweetSentimentExtraction	eng	45	2	90	84.4 [71.2, 92.3]	$\kappa = 0.41$	90.9
<i>Clustering</i>							
ArxivClusteringP2P	eng	30	2	60	49.2 [35.3, 63.2]	ARI=-0.00	84.6*
RedditClusteringP2P	eng	30	2	60	68.8 [63.2, 74.4]	ARI=0.42	100.0*
SIB200ClusteringS2S	ara	30	1	30	76.0 [58.4, 87.8]	N/A	78.8
SIB200ClusteringS2S	dan	30	1	30	62.7 [44.9, 77.6]	N/A	76.0
SIB200ClusteringS2S	eng	30	2	60	54.0 [41.8, 66.3]	ARI=0.15	83.3*
SIB200ClusteringS2S	rus	30	1	30	68.1 [50.2, 81.9]	N/A	77.7
WikiCitiesClustering	eng	30	2	60	97.6 [95.2, 100.0]	ARI=0.91	100.0
<i>Reranking</i>							
Core17Instruction	eng	20	2	40	85.2 [83.6, 86.8]	$\rho = 0.80$	98.8*
News21Instruction	eng	31	2	62	92.7 [91.3, 94.1]	$\rho = 0.85$	100.0*
Robust04Instruction	eng	49	2	98	88.5 [82.2, 94.8]	$\rho = 0.75$	98.8*
WikipediaMultilingual	dan	30	1	30	91.4 [76.2, 97.3]	N/A	95.0
WikipediaMultilingual	eng	30	2	60	82.4 [75.6, 89.1]	$\rho = 0.64$	90.6*
WikipediaMultilingual	nob	30	1	30	89.8 [74.1, 96.4]	N/A	92.3
<i>STS</i>							
SICK-R	eng	40	2	80	82.7 [69.4, 90.5]	$\rho = 0.63$	94.1*
STS12	eng	50	2	100	91.2 [84.9, 94.9]	$\rho = 0.77$	92.0
STS22	ara	30	1	30	67.6 [41.7, 83.3]	N/A	40.9*
STS22	eng	30	2	60	78.4 [59.0, 89.2]	$\rho = 0.75$	82.9
STS22	rus	30	1	30	58.7 [28.8, 78.2]	N/A	69.5
STSBenchmark	eng	50	2	100	80.4 [67.7, 88.4]	$\rho = 0.58$	90.9*

Table 8: Human performance with 95% confidence intervals. N = number of samples; K = number of annotators; Tot = total annotations (N \times K). CIs computed via Wilson Score Interval (Classification), Fisher’s z -transformation (STS), and Annotator Range (Clustering/Reranking). * indicates model score outside human 95% CI ($p < 0.05$). IAA = Inter-Annotator Agreement.

1242 F INTER-ANNOTATOR AGREEMENT ANALYSIS
1243

1244 F.1 AGREEMENT METRICS BY TASK CATEGORY
1245

1246 This section provides detailed inter-annotator agreement analysis using task-appropriate metrics.
1247 Agreement levels follow standard guidelines: $\kappa > 0.8$ (excellent), $0.6 < \kappa \leq 0.8$ (substantial),
1248 $0.4 < \kappa \leq 0.6$ (moderate), $0.2 < \kappa \leq 0.4$ (fair), $\kappa \leq 0.2$ (poor). For correlations: $\rho > 0.7$ (strong),
1249 $0.4 < \rho \leq 0.7$ (moderate), $\rho \leq 0.4$ (weak).
1250

1251 F.1.1 CLASSIFICATION TASKS
1252

- **Emotion Classification:** $\kappa = 0.39$ (fair agreement)
 - 2 annotators, 48 items, 96 total annotations
 - Mean percentage agreement: 52.1%
 - Performance: Human 45.8%, Best model 87.1%
- **Toxicity Classification:** $\kappa = 0.55$ (moderate agreement)
 - 2 annotators, 45 items, 90 total annotations
 - Mean percentage agreement: 77.8%
 - Performance: Human 73.3%, Best model 86.7%
- **Tweet Sentiment Classification:** $\kappa = 0.41$ (moderate agreement)
 - 2 annotators, 45 items, 90 total annotations
 - Mean percentage agreement: 62.2%
 - Performance: Human 84.4%, Best model 90.9%
- **Multilingual Sentiment Classification:** Agreement only for English
 - English: $\kappa = 0.24$ (fair agreement), 2 annotators, 40 items, 62.5% agreement
 - Arabic, Norwegian, Russian: Single annotator (no agreement metrics)
 - Performance: Human advantages in non-English variants

1274 F.1.2 CLUSTERING TASKS
1275

- **ArXiv Clustering:** ARI = -0.001 (no agreement)
 - 2 annotators, 30 items, 60 total annotations
 - Complete breakdown of consensus on academic paper categories
 - Performance: Human 49.2%, Best model 84.6%
- **Reddit Clustering:** ARI = 0.42 (moderate agreement)
 - 2 annotators, 30 items, 60 total annotations
 - Moderate consensus on discussion topic groupings
 - Performance: Human 68.8%, Best model 100%
- **WikiCities Clustering:** ARI = 0.91 (excellent agreement)
 - 2 annotators, 30 items, 60 total annotations
 - High consensus on geographical entity groupings
 - Performance: Human 97.6%, Best model 100%
- **SIB200 Clustering:** Agreement only for English
 - English: ARI = 0.15 (weak agreement), 2 annotators, 30 items
 - Arabic, Danish, Russian: Single annotator (no agreement metrics)
 - Performance varies significantly across languages

1296 F.1.3 RERANKING TASKS
1297
1298 • **News21**: $\rho = 0.85$ (strong agreement)
1299 – 2 annotators, 31 items, 62 total annotations
1300 – Mean Kendall tau: 0.85, Binary kappa: 0.83
1301 – Performance: Human 92.7%, Best model 100%
1302
1303 • **Core17**: $\rho = 0.80$ (strong agreement)
1304 – 2 annotators, 20 items, 40 total annotations
1305 – Mean Kendall tau: 0.80, Binary kappa: 0.78
1306 – Performance: Human 85.2%, Best model 98.8%
1307
1308 • **Robust04**: $\rho = 0.75$ (strong agreement)
1309 – 2 annotators, 49 items, 98 total annotations
1310 – Mean Kendall tau: 0.75, Binary kappa: 0.72
1311 – Performance: Human 88.5%, Best model 98.8%
1312
1313 • **Wikipedia Multilingual Reranking**: Agreement only for English
1314 – English: $\rho = 0.64$ (moderate agreement), 2 annotators, 30 items
1315 – Mean Kendall tau: 0.64, Binary kappa: 0.60
1316 – Danish, Norwegian: Single annotator (no agreement metrics)
1317 – Performance varies across languages

1318 F.1.4 STS TASKS
1319
1320 • **STS12**: $\rho = 0.77$ (strong agreement)
1321 – 2 annotators, 50 items, 100 total annotations
1322 – Performance: Human 91.2%, Best model 92.0%
1323
1324 • **STSBenchmark**: $\rho = 0.58$ (moderate agreement)
1325 – 2 annotators, 50 items, 100 total annotations
1326 – Performance: Human 80.4%, Best model 90.9%
1327
1328 • **SICK-R**: $\rho = 0.63$ (moderate agreement)
1329 – 2 annotators, 40 items, 80 total annotations
1330 – Performance: Human 82.6%, Best model 94.1%
1331
1332 • **STS22**: Agreement only for English
1333 – English: $\rho = 0.75$ (strong agreement), 2 annotators, 30 items
1334 – Arabic, Russian: Single annotator (no agreement metrics)
1335 – Performance varies significantly by language

1336 F.2 AGREEMENT PATTERNS AND TASK RELIABILITY
1337

1338 F.2.1 HIGH-AGREEMENT TASKS (RELIABLE BENCHMARKS)

1339 Tasks with high human agreement ($\kappa > 0.6$ or $\rho > 0.7$) consistently demonstrate:

1340
1341 • Clear, objective task specifications with minimal ambiguity
1342 • Adequate context for making informed judgments
1343 • Cultural and linguistic familiarity for annotators
1344
1345 • Well-defined evaluation criteria with concrete examples
1346 • Minimal dataset quality issues or processing artifacts
1347
1348 • Consistent performance patterns across annotators
1349

Examples: WikiCities clustering, News21/Core17/Robust04 reranking, STS12, STSBenchmark

1350 F.2.2 LOW-AGREEMENT TASKS (PROBLEMATIC BENCHMARKS)

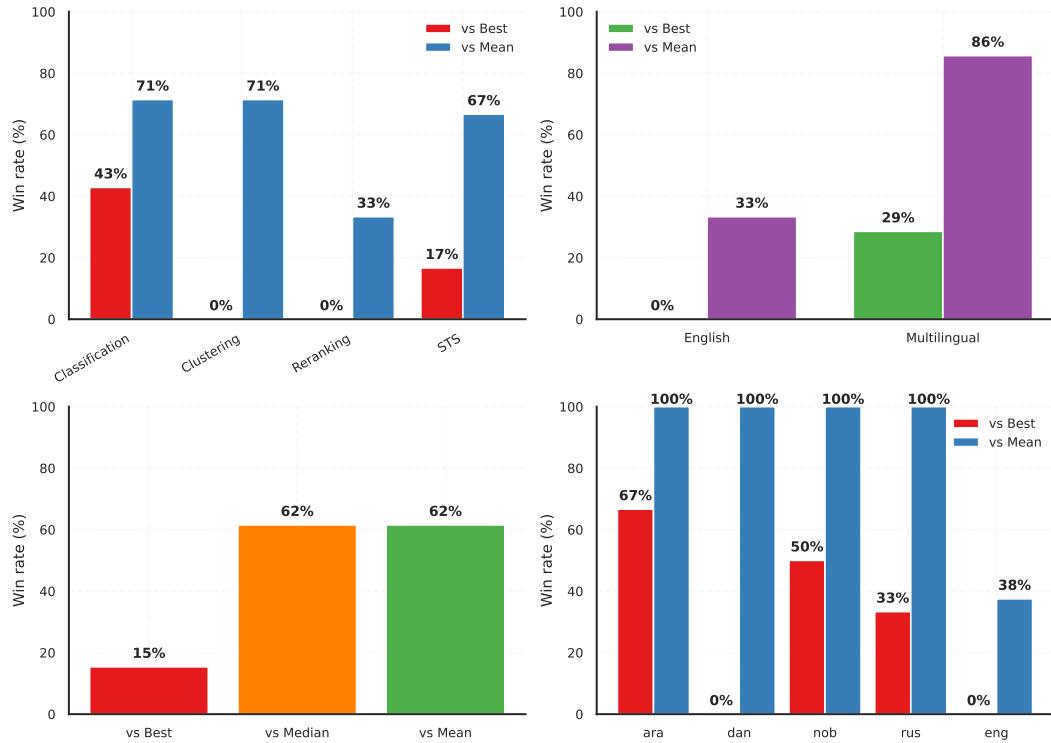
1351 Tasks with low agreement ($\kappa < 0.4$ or $\rho < 0.6$) often exhibit:

1352

- 1353 • Ambiguous annotation guidelines or subjective judgment requirements
- 1354 • Cross-cultural interpretation challenges
- 1355 • Insufficient context for accurate assessment
- 1356 • Systematic dataset quality issues or processing artifacts
- 1357 • Inherently subjective or multi-faceted concepts
- 1358 • Inconsistent or contradictory gold standard labels

1359 **Examples:** Emotion classification, ArXiv clustering, STS22-Russian

1360 G ADDITIONAL HUMAN VS MODEL ANALYSIS



1393 Figure 12: Human win rates across task categories and languages. Top left: By task category
1394 shows humans perform moderately in classification but struggle in clustering, reranking, and STS
1395 against best models. Top right: English-only vs multilingual tasks reveals humans perform better on
1396 multilingual tasks (29% vs 0% against best models). Bottom left: Performance varies dramatically
1397 by baseline comparison (15% vs best, 62% vs mean models). Bottom right: Language-specific
1398 breakdown shows varying performance across different language codes.

1399
1400 This section contains additional analysis on human vs model. Figure 14 shows the human performance
1401 gaps versus median-performing models over all tasks by language. Figure 15 shows the task
1402 difficulty categorization based on human performance levels. Figure 16 shows the model consistency
1403 analysis showing performance ranges across tasks. Figure 2 shows a comprehensive view of human
1404 performance relative to all model performance ranges across 16 tasks by language.

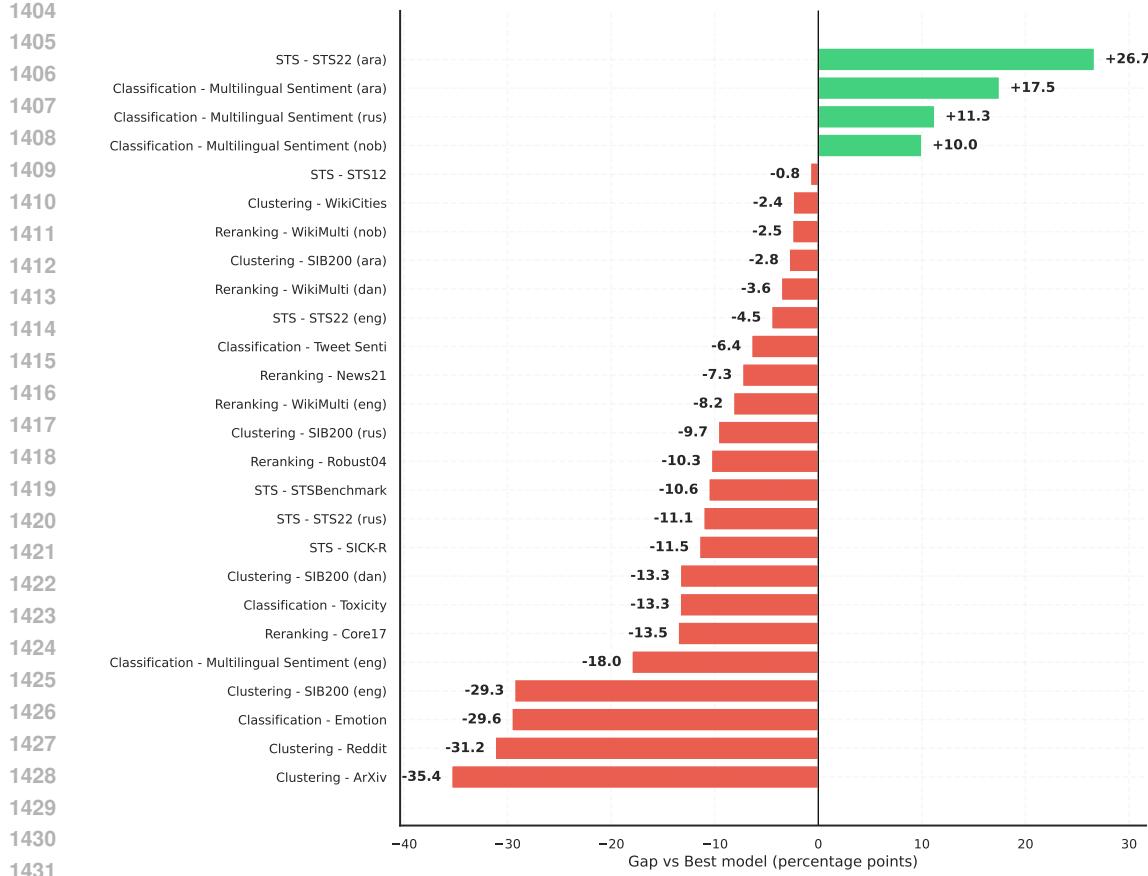


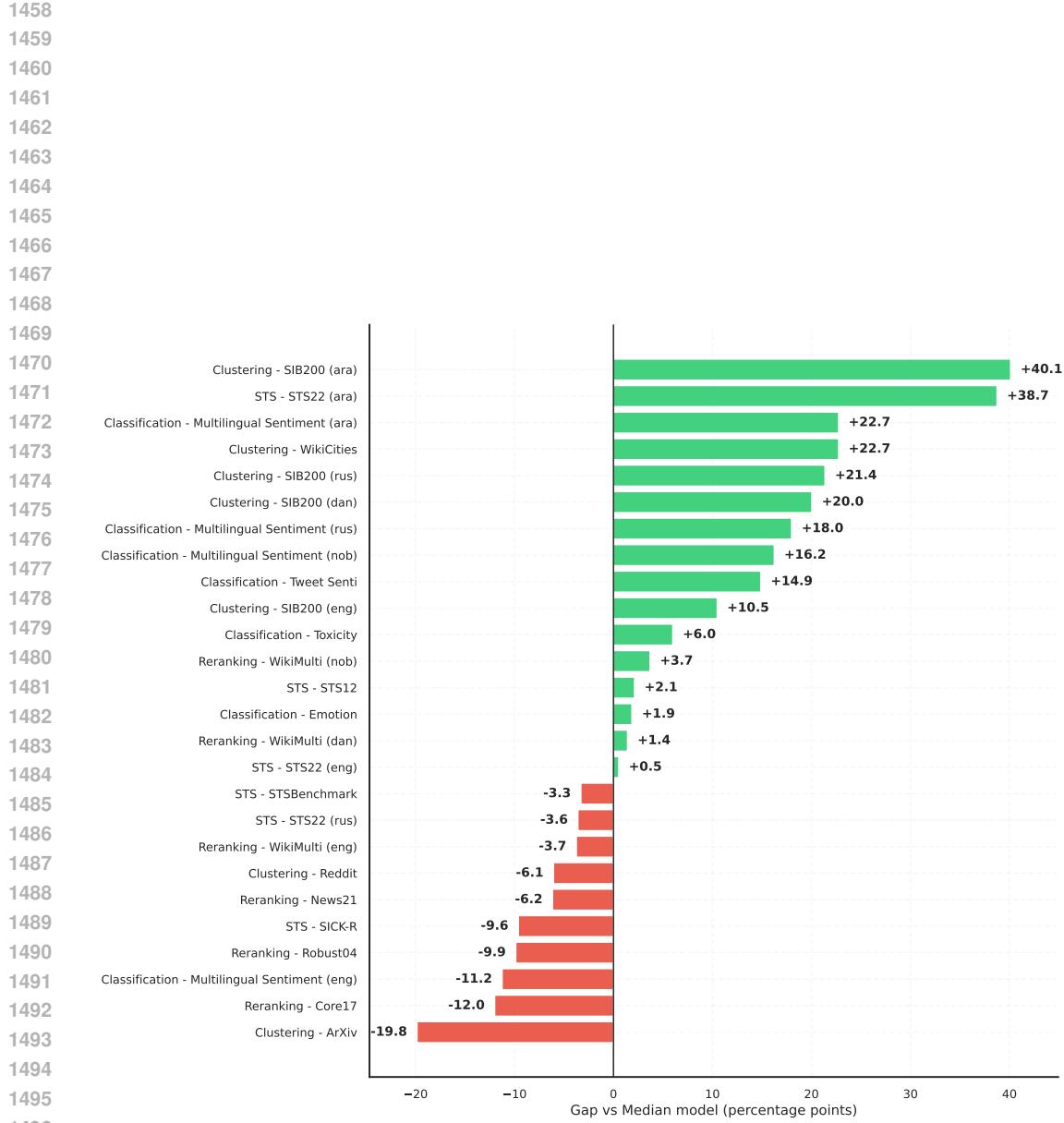
Figure 13: Human performance gaps versus best-performing models across 26 task-language pairs. Humans outperform the best models on only 4 tasks (15.4%), with largest advantages in Arabic semantic similarity and sentiment analysis. The analysis reveals systematic model advantages in technical domains (clustering, reranking) versus human advantages in culturally-informed tasks.

H MODELS EVALUATED

Table 9 shows information about each evaluated model.

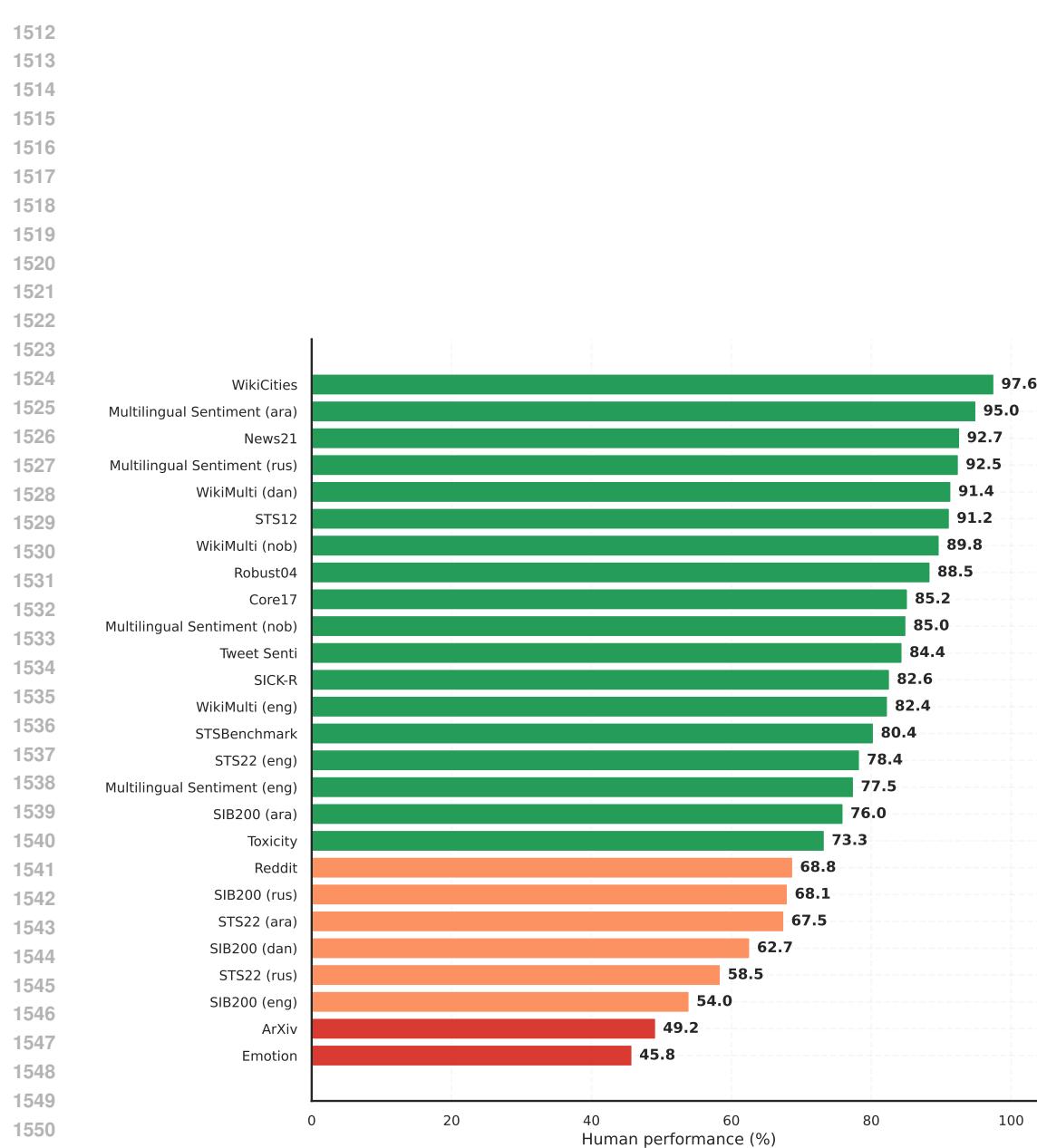
Model	Parameters (Millions)
Alibaba-NLP/gte-Qwen2-1.5B-instruct Li et al. (2023)	1780
google/embeddinggemma-300m Vera et al. (2025)	300
intfloat/e5-mistral-7b-instruct Wang et al. (2023; 2022)	7111
intfloat/multilingual-e5-large Wang et al. (2022)	560
intfloat/multilingual-e5-base Wang et al. (2022)	278
intfloat/multilingual-e5-small Wang et al. (2022)	118
mixedbread-ai/mxbai-embed-large-v1 Lee et al. (2024); Li & Li (2023)	335
NovaSearch/jasper_en_vision_language_v1 Zhang et al. (2025a)	1999
Qwen/Qwen3-Embedding-0.6B Zhang et al. (2025b)	596
Salesforce/SFR-Embedding-Mistral (Meng et al., 2024)	7110
sentence-transformers/all-MiniLM-L6-v2 Reimers & Gurevych (2019)	22.7
sentence-transformers/all-mpnet-base-v2 Reimers & Gurevych (2019)	109

Table 9: List of all evaluated models. Model sizes are in millions of parameters



1497 Figure 14: Human performance gaps versus median-performing models across 26 tasks by language.
 1498 Humans achieve 61.5% win rate against median models, demonstrating competitive performance
 1499 when compared to typical rather than best-performing models. This analysis reveals that human
 1500 performance is much more competitive when compared against representative model performance
 1501 rather than cherry-picked best results.

1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511



1550
1551
1552 Figure 15: Task difficulty categorization based on human performance levels. The majority of tasks
1553 (69%) fall into the “easy” category (human performance ≥ 0.7), shown in green. Only two tasks fall
1554 below 0.5 (shown in red), both with notably low inter-annotator agreement, suggesting fundamental
1555 task ambiguity rather than limitations of human ability.

1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

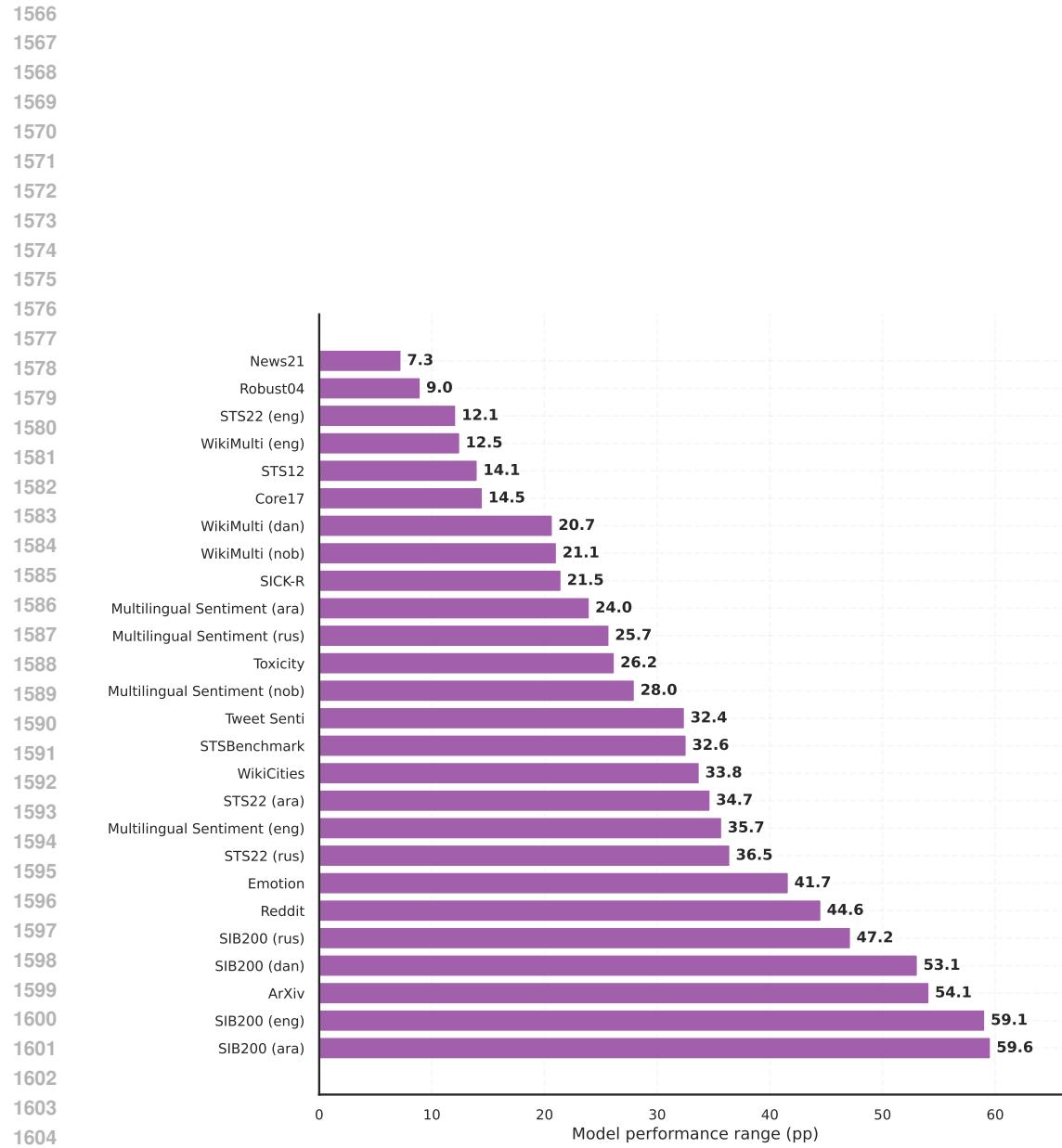


Figure 16: Model consistency analysis showing performance ranges across tasks. **Higher value indicates greater variability across models** (lower consistency). Tasks with small ranges (high consistency) often align with high human agreement, whereas tasks with large ranges (low consistency) typically correspond to tasks where humans also struggle. This pattern suggests that both human and model performance reflect underlying task quality and clarity of task specification.

1620 I DETAILED LLM-AS-ANNOTATOR RESULTS

1621
 1622 This section provides detailed LLM-as-annotator performance for each task-language pair. LLMs
 1623 were not evaluated on clustering tasks due to the difficulty of eliciting cluster assignments from
 1624 generative models.

1626 1627 1628 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673 Dataset	Human	GPT-5		GPT-4.1		Gemini 2.5 Flash	Mistral		Qwen3		
		Full	Mini	Full	Mini		Small-24B-I	30B	32B	Coder	
Emotion [eng]	45.8	37.5	33.3	41.7	29.2	35.4	43.1	37.5	38.3	42.1	
Multilingual Sentiment [ara]	95.0	92.5	95.0	90.0	90.0	92.5	90.5	92.8	85.5	91.2	
Multilingual Sentiment [eng]	77.5	100.0	100.0	97.5	95.0	97.5	97.2	99.0	93.0	96.5	
Multilingual Sentiment [nob]	85.0	92.5	85.0	87.5	82.5	85.0	86.8	86.8	77.2	88.5	
Multilingual Sentiment [rus]	92.5	87.5	85.0	75.0	85.0	92.5	64.0	74.2	75.8	78.0	
Toxic Conversations [eng]	73.3	73.3	66.7	73.3	75.6	71.1	69.6	60.0	68.2	66.9	
Tweet Sentiment [eng]	84.4	68.9	75.6	71.1	75.6	68.9	65.6	68.9	72.9	70.7	
Average	79.1	78.9	77.2	76.6	76.1	77.6	73.8	74.2	73.0	76.3	

Table 10: Detailed LLM-as-annotator results for classification tasks. **Bold** indicates best performance per row.

Dataset	Human	GPT-5		GPT-4.1		Gemini 2.5 Flash	Mistral		Qwen3		
		Full	Mini	Full	Mini		Small-24B-I	30B	32B	Coder	
Core17 [eng]	85.2	76.6	83.2	72.7	78.0	76.8	74.4	89.8	74.2	72.2	
News21 [eng]	92.7	72.9	78.7	75.1	78.2	75.0	77.4	80.5	77.9	75.2	
Robust04 [eng]	88.5	75.7	84.0	79.5	79.3	79.7	77.2	84.6	79.2	73.6	
Wikipedia [dan]	91.4	84.7	78.3	85.8	83.6	82.8	84.0	76.6	79.8	79.9	
Wikipedia [eng]	82.4	69.6	61.0	70.8	68.6	74.2	75.1	59.8	73.4	69.5	
Wikipedia [nob]	89.8	71.5	68.1	70.6	75.8	68.6	79.6	62.6	64.2	72.3	
Average	88.3	75.1	75.5	75.7	77.2	76.2	78.0	75.6	74.8	73.8	

Table 11: Detailed LLM-as-annotator results for reranking tasks. **Bold** indicates best performance per row.

Dataset	Human	GPT-5		GPT-4.1		Gemini 2.5 Flash	Mistral		Qwen3		
		Full	Mini	Full	Mini		Small-24B-I	30B	32B	Coder	
SICK-R [eng]	82.6	68.5	67.5	66.9	72.9	59.0	66.7	59.4	57.3	68.8	
STS12 [eng]	91.2	83.2	82.7	87.3	87.8	83.8	83.9	83.8	81.7	84.5	
STS22 [ara]	67.5	57.1	45.5	56.9	55.8	43.1	56.9	35.3	50.4	47.9	
STS22 [eng]	78.4	67.8	80.3	81.6	79.7	81.3	78.2	74.1	76.2	76.4	
STS22 [rus]	58.7	71.9	51.3	67.2	66.6	64.9	78.6	65.9	62.4	67.0	
STSBenchmark [eng]	80.4	89.3	86.7	89.7	86.3	83.8	85.4	84.3	84.0	82.9	
Average	76.5	73.0	69.0	74.9	74.9	69.3	75.0	67.1	68.6	71.3	

Table 12: Detailed LLM-as-annotator results for sts tasks. **Bold** indicates best performance per row.

I.1 HUMAN-LLM TASK DIFFICULTY CORRELATION

To assess whether humans and LLMs face similar challenges, we computed Spearman rank correlations between human and LLM performance across the 19 task-language pairs (excluding clustering). A positive correlation indicates that tasks where humans perform well also tend to be tasks where LLMs perform well.

The overall correlation is moderate and statistically significant ($\rho = 0.52$, $p = 0.023$), suggesting partially shared difficulty patterns.

LLM Model	Spearman ρ	<i>p</i> -value
GPT-5 Full	0.47	0.042
GPT-5 Mini	0.52	0.023
GPT-4.1 Full	0.43	0.066
GPT-4.1 Mini	0.57	0.012
Gemini 2.5 Flash	0.50	0.029
Mistral Small	0.30	0.215
Qwen3-30B	0.50	0.031
Qwen3-32B	0.52	0.022
Qwen3-Coder	0.57	0.011
Average (all LLMs)	0.52	0.023

Table 13: Spearman rank correlation between human and LLM performance across 19 task-language pairs (clustering excluded). A positive ρ indicates that tasks where humans score high also tend to be tasks where LLMs score high. Values with $p < 0.05$ indicate statistically significant correlations. The moderate positive correlations suggest partially shared task difficulty patterns between humans and LLMs.

J LLM USAGE STATEMENT

Large language models were used to assist with formatting, citation integration, and writing polish during the preparation of this manuscript. Specifically, we used LLMs for:

- Formatting assistance for LaTeX tables and mathematical notation
- Integration and standardization of citation formats
- Minor writing improvements for clarity and flow
- Code documentation and data processing script organization

All substantive content, including research design, data analysis, interpretation of results, and scientific conclusions, was developed entirely by the authors. The core contributions, methodology, and findings presented in this work are the original intellectual contribution of the research team. LLM assistance was limited to technical formatting and presentation improvements that did not influence the scientific content or conclusions of the study.