

Beyond the Leaderboard Average: Cross-Task Consistency, Model Scale, and Specialization Structure in the Massive Text Embedding Benchmark

Zhe Zhang

zzhe911205@gmail.com

Liaoning Finance and Trade College

Research Article

Keywords: text embeddings, benchmark evaluation, information retrieval, cross-task generalization, representation learning, MTEB

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11087134/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Text embedding models are routinely compared using a single aggregate score on the Massive Text Embedding Benchmark (MTEB), yet an aggregate conceals whether a model performs uniformly across task families or excels narrowly on a subset of them. We revisit the original MTEB English-subset results (30 models, 7 task categories) and ask a question the original benchmark report did not directly quantify: is high aggregate performance associated with high cross-task consistency, and does model scale buy consistency in the same way it buys average performance? Using the publicly released per-model, per-task scores, we compute pairwise task correlations, a z-score-based cross-task profile dispersion index for every model, and a Ward hierarchical clustering of models by task profile. We find that profile dispersion is negatively associated with average score (Pearson $r = -0.48$, $p = 0.008$), indicating that top-performing models tend to be more uniformly competent rather than narrowly optimized, but dispersion is essentially uncorrelated with parameter count ($r = -0.003$, $p = 0.99$), indicating that scale alone does not confer cross-task balance. Clustering recovers four qualitatively distinct groups, including a "retrieval-and-reranking specialist" cluster with a pronounced summarization deficit and a fifteen-model "balanced generalist" cluster spanning multiple architectures. These findings reframe model-selection guidance for practitioners and motivate reporting dispersion alongside average score in future embedding leaderboards. All data, code and derived tables are made available for reproducibility [1].

1 Introduction

Dense vector representations of text – text embeddings – now sit underneath most modern search, retrieval-augmented generation, deduplication, clustering, and recommendation pipelines. Because a single embedding model is frequently reused across many downstream applications inside the same production system, understanding how a model's competence is distributed across task families is at least as practically important as knowing its single best headline score. Prior to the introduction of the Massive Text Embedding Benchmark (MTEB), embedding models were typically evaluated on a narrow slice of tasks – semantic textual similarity (STS) being the most common – despite mounting evidence that strong STS performance does not transfer reliably to retrieval or clustering use cases. MTEB addressed this evaluation gap directly by unifying eight embedding task families across fifty-eight datasets and, in its English subset, thirty models, reporting the now widely cited finding that no single model dominates every task [1].

That finding, while influential, was reported primarily at the level of the aggregate leaderboard and a pairwise task-correlation heatmap. It leaves several practically important questions only partially answered. First, is the well-known non-dominance result driven by a genuine trade-off in which stronger models sacrifice generality for peak performance on a subset of tasks, or is the opposite true – that stronger models are, if anything, more evenly competent, and the apparent non-dominance is a product of near-ties among the top handful of models rather than a scale-consistency tradeoff? Second, does increasing model scale, which is known to raise average score, also improve the internal balance of a model's task-by-task profile, or does scale simply lift some tasks more than others and leave the profile

just as uneven? Third, do models cluster into recognizable specialization archetypes – for example, a retrieval-and-reranking cluster distinct from an STS-and-classification cluster – and if so, what do those clusters imply about the fine-tuning objectives that produced them [1]?

These questions have taken on renewed practical weight since MTEB’s original release, because the same embedding model is now routinely reused across multiple stages of a single production pipeline: a retrieval-augmented generation (RAG) system, for instance, typically uses one embedding model to index a document corpus (a retrieval task), the same model or a closely related one to deduplicate or cluster ingested documents (a clustering task), and increasingly the same embedding space to rerank retrieved passages before they are passed to a generative model (a reranking task). A practitioner who selects an embedding model purely on the strength of its published Average score has no way of knowing, without consulting the full per-task breakdown, whether that average was earned uniformly or was inflated by a narrow strength in exactly one of the three stages their pipeline actually depends on. Quantifying cross-task consistency directly, rather than leaving it implicit in a multi-column table, is therefore not only of academic interest but has a direct bearing on model-selection practice in applied information retrieval and NLP systems [1].

This paper answers these three questions using the original, publicly released permodel, per-task-family scores from the MTEB English benchmark. Rather than proposing a new set of tasks, a new embedding model, or re-running the underlying embedding computations, we treat the published results table as a secondary dataset and apply a battery of exploratory statistical procedures that were not part of the original release: (i) a full task-by-task Pearson and Spearman correlation analysis restated with exact coefficients rather than a qualitative heatmap description; (ii) a novel crosstask profile dispersion index, computed from within-task z-scores, that quantifies how unevenly a given model’s competence is spread across the seven task families; (iii) a correlation analysis relating this dispersion index to both average score and model parameter count; and (iv) a Ward-linkage hierarchical clustering of the thirty models based on their seven-dimensional task profile, which recovers interpretable specialization archetypes [1].

We view this exercise as a case study in benchmark-level secondary analysis: a methodological contribution in the sense described by this journal’s Methodology article type, applied to Information Retrieval and Natural Language Processing evaluation, both squarely within the scope of *Discover Computing*. The remainder of the paper is organized as follows. Section 2 situates MTEB within the broader landscape of NLP and IR benchmarks and embedding model families. Section 3 briefly summarizes the task taxonomy needed to interpret our analysis. Section 4 describes the data source and the statistical procedures. Section 5 reports the empirical findings, including the correlation structure, the dispersion analysis, and the cluster archetypes. Section 6 interprets these findings for embedding model selection and for future benchmark design, Section 7 discusses limitations, and Section 8 concludes.

2 Related Work

2.1 Benchmarks for language understanding and embeddings

General-purpose language understanding benchmarks such as GLUE and its successor SuperGLUE established the now-standard practice of aggregating scores across a heterogeneous task suite into a single leaderboard number, a practice that has since been adopted almost universally in NLP evaluation. For sentence- and text-embedding models specifically, evaluation historically centered on the SemEval Semantic Textual Similarity shared tasks and their aggregation in toolkits such as SentEval. These toolkits, however, evaluate embeddings largely through the lens of similarity and downstream classifier probing, and they do not cover retrieval or clustering, where embeddings are compared directly via distance metrics rather than through a trained classifier head. The BEIR benchmark filled part of this gap for zero-shot information retrieval by assembling a heterogeneous set of retrieval datasets spanning multiple domains, and it has since become a de facto standard for retrieval-specific embedding evaluation [18–20, 28].

MTEB was introduced explicitly to unify these previously disjoint evaluation traditions – similarity, classification, clustering, reranking, retrieval, bitext mining, and summarization – into a single benchmark and a single public leaderboard, incorporating BEIR’s retrieval datasets and the SemEval STS datasets alongside newly assembled classification, clustering, and reranking datasets. The headline empirical finding of the original MTEB paper was that model rankings vary substantially by task family and that no model tested was simultaneously state-of-the-art across all eight task types. The benchmark has since been extended substantially: the Massive Multilingual Text Embedding Benchmark (MMTEB) expands the collection to several hundred tasks across more than 250 languages and adds instruction-following, long-document, and coderetrieval tasks, and a growing body of survey work has begun cataloguing the topperforming embedding families on the MTEB public leaderboard as the number of submitted models has grown into the hundreds [1–4, 9–14].

2.2 Embedding model families

The models evaluated in the original MTEB English subset span several distinct training paradigms, and understanding these paradigms is necessary for interpreting the cluster structure we report in Section 5. Static word-embedding methods such as GloVe lack context sensitivity and serve as a floor baseline. Self-supervised transformer encoders such as BERT can be converted into sentence embeddings via mean pooling, but without further fine-tuning this conversion is known to produce anisotropic, poorly separated embedding spaces. Contrastive fine-tuning approaches such as SimCSE substantially improve STS performance with either unsupervised or natural-languageinference-supervised contrastive objectives, but the original MTEB results already suggested that STS-oriented contrastive fine-tuning does not transfer uniformly to retrieval. Retrieval-oriented dual-encoder families such as Contriever, GTR, and the coCondenser lineage are fine-tuned with retrieval-style contrastive objectives, typically on MS MARCO-derived query-document pairs, and correspondingly emphasize the Retrieval and Reranking task families. Multi-task sentence-encoder families such as SentenceT5 (ST5), MPNet-based Sentence-

Transformers, and MiniLM-based Sentence-Transformers are fine-tuned on broad mixtures of paraphrase, NLI, and QA-style datasets and were, at the time of MTEB’s original release, the strongest generalists across the full task suite. Decoder-based embedding approaches such as SGPT extract embeddings from generative transformer decoders via weighted mean pooling and bias-only fine-tuning, and were evaluated in both an NLI-oriented and an MS-MARCO-oriented configuration, again illustrating that the choice of fine-tuning objective visibly steers the resulting task profile [3–16, 22].

2.3 Evaluation metric heterogeneity across task families

A further source of difficulty in comparing embedding models across MTEB’s task families, distinct from the dataset-imbalance concern discussed below, is that the seven task families are not scored with the same underlying metric. Classification is scored by the accuracy of a downstream logistic-regression probe; Clustering is scored by V-measure, an entropy-based external cluster-validity measure that is invariant to label permutation; Pair Classification and Reranking are scored by variants of average precision computed over a small candidate set; Retrieval is scored by nDCG@10 over a much larger candidate pool, inherited directly from the BEIR retrieval benchmark; and STS and Summarization are both scored by Spearman rank correlation between embedding-space cosine similarity and a human judgment. Although MTEB rescales all of these onto a nominal 0–100 range for presentation, the metrics are not measuring the same underlying statistical quantity, and a five-point difference on, say, Clustering’s V-measure is not necessarily comparable in difficulty or in practical significance to a five-point difference on Retrieval’s nDCG@10. This heterogeneity is precisely why our analysis relies on within-task standardization (z-scoring) rather than working with raw scores when constructing the profile dispersion index and the hierarchical clustering input, since standardization removes each metric’s arbitrary scale and location while preserving each model’s relative standing within its task family [1, 2, 21].

2.4 Aggregate scores versus profile-level analysis

A recurring methodological critique of single-number benchmark leaderboards, raised in the broader machine-learning evaluation literature, is that an aggregate obscures heterogeneous performance across sub-populations of the evaluation suite – a concern closely analogous to Simpson’s-paradox-style aggregation issues in applied statistics. This concern is not unique to embeddings: benchmark suites such as GLUE and BigBench have each attracted follow-up work arguing that a single leaderboard score conflates genuinely distinct capabilities and can reward models that are strong on the numerically dominant sub-tasks of the suite while masking weakness on the numerically minor ones. MTEB is structurally susceptible to exactly this concern, since, as we note in Section 3, its constituent task families contribute unequal numbers of underlying datasets to the unweighted Average column. The original MTEB paper partially addressed the aggregation concern by publishing a pairwise task correlation heatmap and by discussing task-specific winners in prose, and it explicitly flagged task imbalance in its own limitations section as a threat to the interpretability of the unweighted average. It did not, however, report a per-model dispersion statistic that would let a leaderboard reader identify which models are most exposed to this imbalance, nor did it test whether such dispersion is related to

model scale, nor did it formally cluster models by task profile to identify recurring specialization patterns. Closing this specific gap – moving from a qualitative acknowledgment of possible imbalance to a quantified, per-model diagnostic – is the contribution this paper makes [18, 19].

3 The MTEB Task Taxonomy

For completeness, we briefly restate, in our own terms, the seven English-language task families analyzed in this paper (an eighth MTEB task, bitext mining, is multilingual only and is excluded from the English-subset analysis that forms our dataset). Classification trains a lightweight logistic-regression probe on top of frozen embeddings and reports test-set accuracy, so it partially rewards embeddings that are linearly separable by label rather than embeddings that are geometrically well-structured in an unsupervised sense. Clustering instead evaluates embeddings without any trained probe: a k-means model is fit directly on the embedded texts and scored by how well the resulting partition matches ground-truth categories, which more directly tests unsupervised geometric structure. Pair classification asks whether two texts are duplicates or paraphrases based on embedding distance and a tuned threshold. Reranking evaluates the ordering induced by cosine similarity between a query embedding and a set of candidate embeddings, scored by mean average precision. Retrieval extends this to full corpus-scale search, embedding both queries and a large candidate pool and scoring the ranking with nDCG@10, reusing datasets originally assembled for the BEIR benchmark. Semantic textual similarity scores the correlation between embeddingspace cosine similarity and human similarity judgments on sentence pairs. Summarization scores machine-generated summaries by their embedding-space proximity to human-written reference summaries. These seven metrics are on different underlying scales (accuracy, V-measure, average precision, MAP, nDCG@10, and two Spearman correlations), which is precisely why a raw arithmetic average, while convenient, can be a misleading single-number summary and why a standardized, per-task z-score treatment is appropriate for the cross-task comparison undertaken in this paper. To ground these abstract task definitions, it is worth noting the concrete character of a few representative constituent datasets: Clustering includes tasks built from arXiv, bioRxiv, and medRxiv paper metadata, where the model must group paper titles or title-abstract pairs by their human-assigned subject category without ever seeing that category label during embedding; Retrieval includes the natural-questions and MS MARCO websearch collections inherited from BEIR, where the corpus of candidate documents can run into the millions; and Summarization consists of a single dataset, SummEval, in which machine-generated news summaries are scored by their embedding-space proximity to human-written reference summaries of the same article. This last point – that Summarization rests on only one constituent dataset while Retrieval and Classification each draw on more than a dozen – is itself a source of the task-imbalance concern raised again in Section 2 and directly investigated in Section 5 [1, 2, 21, 23].

4 Methodology

4.1 Data source

Our dataset is the published main-metric-per-task-per-model results table from the original MTEB English benchmark release, comprising thirty embedding models evaluated across the seven English-language task families described in Section 3, together with the paper-reported arithmetic average across all fifty-six constituent English datasets. We additionally recorded, for each model, its approximate parameter count and its architectural/training family (word embedding, self-supervised transformer, supervised contrastive, GTR, ST5, or SGPT), drawn from the model documentation accompanying the original release, to support the scale- and family-conditioned analyses in Sections 5 and 6. Parameter counts for the publicly released checkpoints (e.g., the MiniLM, MPNet, GTR, and ST5 variants) were taken from the corresponding model cards. The API-only Ada Similarity endpoint, for which the original release did not report a complete Retrieval score or a public parameter count, was excluded from our dataset entirely rather than imputed, leaving the thirty models with complete seventask profiles analyzed throughout this paper. No new embeddings were computed and no new datasets were assembled; every score analyzed here is a previously published, peer-reviewed number, and our contribution is entirely in the statistical treatment applied to it. The complete transcribed dataset, together with all analysis code, is included in the supplementary material accompanying this submission for reproducibility (Section 8) [1, 29].

4.2 Cross-task correlation

For the thirty models and seven task-family average scores, we compute the full 7×7 Pearson product-moment correlation matrix and, as a robustness check against the influence of the small number of very large decoder-based outliers, the corresponding Spearman rank correlation matrix. A high correlation between two task families indicates that the same underlying model properties (e.g., representation isotropy, finetuning objective, parameter count) drive performance on both; a low or negative correlation indicates that the two task families reward largely distinct model properties [1].

4.3 Cross-task profile dispersion index

To quantify, for each model individually, how unevenly its competence is spread across task families, we standardize each of the seven task columns to zero mean and unit variance across the thirty models (a within-task z-score), and then compute, for each model, the standard deviation of its seven z-scores. We refer to this per-model statistic as the *profile dispersion index*. A model whose seven z-scores are all close to its own mean z-score (i.e., a model that is uniformly strong, uniformly average, or uniformly weak relative to other models on every task) has a low dispersion index; a model that is far above the field on some tasks and far below it on others has a high dispersion index, regardless of its average raw score. This index is deliberately independent, by construction, of a model's overall average level of performance, allowing us to ask whether consistency and average performance are empirically related or orthogonal properties in this benchmark.

4.4 Scale and consistency

We test two hypotheses: (H1) that profile dispersion is negatively associated with average score, i.e., that stronger models are also more uniformly strong; and (H2) that profile dispersion is negatively associated with model scale (parameter count, logtransformed), i.e., that larger models generalize more

evenly across task families. Both hypotheses are tested with Pearson correlation and a two-sided significance test at $\alpha = 0.05$, given the modest sample size ($n = 30$).

4.5 Hierarchical clustering of task profiles

To identify qualitative specialization archetypes rather than only pairwise statistics, we apply Ward-linkage agglomerative hierarchical clustering to the thirty models' sevendimensional standardized task profiles (the same z-scores used for the dispersion index). We cut the resulting dendrogram at four clusters, a choice that balances interpretability against over-fragmentation given the sample size, and describe each cluster by its mean z-score profile and its constituent models' architectural families [30].

4.6 Rank-based robustness checks

Because Pearson correlation is sensitive to a small number of extreme-valued models (in this dataset, most conspicuously the very large decoder-based SGPT variants), we complement it throughout with Kendall's τ rank correlation for two task pairs of particular theoretical interest: Retrieval versus STS, which the embedding literature has previously characterized as an "asymmetric versus symmetric" task distinction that some models are known to trade off against each other, and Clustering versus Reranking, the pair identified as most strongly associated in Section 5. Kendall's τ , being based only on the number of concordant and discordant pairs of models rather than on the magnitude of their scores, provides a check that our Pearson-based findings are not artifacts of a handful of outlying models with unusually large scores [14].

4.7 Self-supervised versus supervised comparison

Following the original MTEB paper's own grouping of models into self-supervised methods (untuned or self-supervised-only word- and sentence-embedding models) and supervised methods (models that received some form of labeled or contrastively-supervised fine-tuning), we compute the mean Average score within each group as a simple, easily interpretable companion statistic to the correlation- and cluster-based analyses above [1].

4.8 Software

All analyses were conducted in Python 3 using pandas for data handling, SciPy for correlation testing and hierarchical clustering, and Matplotlib for visualization. Exact package versions and the full analysis script are provided in the supplementary repository [31].

5 Results

5.1 Cross-task correlation structure

Table 1 reports the full 7×7 Pearson correlation matrix across the thirty models. Two structural patterns are apparent. First, Reranking and Clustering are the most strongly associated pair of task families in the entire matrix ($r = 0.93$), followed closely by Clustering and Retrieval ($r = 0.87$) and Reranking and

Retrieval; this cluster of three tasks shares an evaluation logic that depends on the relative ranking or grouping of many embeddings simultaneously rather than on a pairwise binary decision or a trained linear probe, which plausibly explains why they co-vary so strongly across models. Second, Summarization stands apart from every other task family, with correlations against the other six tasks ranging from $r = -0.24$ to $r = 0.14$ and centered near zero; a model's summarization score is, in this dataset, essentially unpredictable from its performance on any of the other six task families. Classification, Pair Classification, and STS form an intermediate, moderately-to-strongly correlated group (r between 0.66 and 0.84), consistent with the fact that all three tasks can be solved reasonably well by embeddings that separate semantically distinct short texts along a small number of dominant directions, even without the many-way ranking structure that Clustering, Reranking, and Retrieval require. The Spearman rank correlation matrix (not tabulated in full for space, available in the supplementary material) reproduces this same qualitative structure, with the Reranking–Clustering pair remaining the strongest association ($\rho = 0.92$) and Summarization remaining the clear outlier, indicating that the pattern is not an artifact of a small number of extreme-valued models [1].

Table 1
Pearson correlation matrix of the seven MTEB English task-family average scores, computed across the thirty models in the original benchmark release. Cls. = Classification, Clu. = Clustering, PCl. = Pair Classification, Rer. = Reranking, Ret. = Retrieval, Sum. = Summarization.

Classification	Cls.	Clu.	PCl.	Rer.	Ret.	STS	Sum.
	1.00	0.74	0.73	0.66	0.65	0.83	0.10
Clustering	0.74	1.00	0.74	0.93	0.87	0.81	0.05
Pair Classification	0.73	0.74	1.00	0.72	0.79	0.84	0.12
Reranking	0.66	0.93	0.72	1.00	0.88	0.83	-0.01
Retrieval	0.65	0.87	0.79	0.88	1.00	0.75	-0.10
STS	0.83	0.81	0.84	0.83	0.75	1.00	0.08
Summarization	0.10	0.05	0.12	-0.01	-0.10	0.08	1.00

Figure 1 visualizes the same matrix.

5.2 Task-level variability and the retrieval outlier

Before turning to the dispersion and clustering analyses, it is useful to characterize how much the thirty models actually disagree with one another on each task family, since a task on which all models score similarly offers little discriminative signal for model selection regardless of its correlation with other tasks. Table 2 reports the across-model range and standard deviation for each of the seven task families and for the Average column. Retrieval is by a substantial margin the most discriminating task family, with a 42.3-point range and a standard deviation (12.4) roughly twice that of any other task family; this is consistent with Retrieval being the task family most dependent on a specific fine-tuning objective (query-

document contrastive pre-training on corpora such as MS MARCO) that only a subset of the thirty models received in any form. Summarization, by contrast, has both the smallest range (8.3 points) and the smallest standard deviation (2.15) of any task family, meaning that essentially all thirty models, regardless of architecture or fine-tuning objective, perform within a comparatively narrow band on this task – a pattern that likely also contributes to Summarization’s near-zero correlation with every other task family reported in Table 1, since a task with little across-model variance has correspondingly little variance available to correlate with anything else.

Table 2
Across-model range and standard deviation for each task family and the overall Average, computed over the thirty models in the dataset. Values are in the native metric units for each task (accuracy, V-measure, average precision, MAP, nDCG@10, or Spearman correlation, all reported on a 0–100 scale by MTEB convention)

Task family	Range	Std. dev.
Classification	21.05	5.08
Clustering	28.43	6.48
Pair Classification	29.80	7.44
Reranking	17.92	4.85
Retrieval	42.32	12.38
STS	28.27	7.81
Summarization	8.26	2.15
Average	25.88	7.17

Table 3 lists the top-scoring model on each of the seven task families. Four distinct models occupy the top position across the seven tasks – ST5-XXL (Classification, Clustering, and STS), GTR-XL (Pair Classification), MPNet (Reranking), SGPT-5.8Bmsmarco (Retrieval), and MPNet-multilingual (Summarization) – directly reproducing, with exact figures, the original MTEB paper’s headline qualitative claim that no single model is simultaneously the best choice on every task family. Three of the seven task-family winners belong to the Cluster A "balanced generalist" group identified in Section 5, while the Retrieval winner belongs to Cluster B, further reinforcing the specialization/generalization split that the cluster analysis identifies [1].

Table 3
Top-scoring model on each MTEB English task family, with its score. No single model tops more than three of the seven task families

Task family	Top model	Score
Classification	ST5-XXL	73.42
Clustering	ST5-XXL	43.71
Pair Classification	GTR-XL	86.13
Reranking	MPNet	59.36
Retrieval	SGPT-5.8B-msmarco	50.25
STS	ST5-XXL	82.63
Summarization	MPNet-multilingual	31.57

Kendall's τ rank correlation confirms the Pearson-based findings in Table 1 while ruling out an outlier-driven artifact: Clustering and Reranking rankings agree strongly ($\tau = 0.769$, $p < 0.001$), while Retrieval and STS rankings, the archetypal "asymmetric versus symmetric" task pairing discussed in the embedding literature, agree only weakly ($\tau = 0.329$, $p = 0.010$) – present and statistically distinguishable from a null association, but far weaker than the Clustering–Reranking agreement, indicating that a model's relative standing on retrieval carries comparatively little information about its relative standing on sentence similarity.

Finally, grouping models by the original paper's self-supervised versus supervised distinction shows a substantial gap: the four self-supervised or purely word-embedding models (GloVe, Komninos, BERT with mean pooling, and unsupervised SimCSE) average 41.95 points on the Average column, compared to 53.69 points for the twenty-six models that received some form of supervised or supervised-contrastive fine-tuning – an 11.7-point gap that underscores how much of the field's progress on this benchmark, as of the original 2022 release, was attributable to fine-tuning objective rather than to pre-training architecture alone [1, 4, 7, 14].

5.3 Profile dispersion, average performance, and scale

The profile dispersion index, computed as described in Section 4, ranges from a low of approximately 0.4 standard-deviation-units for the most internally consistent models to above 1.6 for the least consistent ones in our sample. Consistent with hypothesis H1, profile dispersion is significantly negatively correlated with average score ($r = -0.475$, $p = 0.008$, $n = 30$): models with higher overall averages tend to have flatter, more uniform standardized profiles across the seven task families, rather than achieving their high average by dominating a subset of tasks while lagging on others. Contrary to a naive scaling intuition, however, hypothesis H2 is not supported: profile dispersion shows essentially no relationship with log-transformed parameter count ($r = -0.003$, $p = 0.99$). Average score itself does correlate moderately with log parameter count ($r = 0.487$, $p = 0.006$), replicating the scale-performance relationship

reported qualitatively in the original MTEB release, but this scale-driven performance gain does not translate into a more balanced task profile. In other words, within this sample, scaling up a model tends to raise its average competence without making that competence any more evenly distributed across task families; cross-task balance appears to be governed primarily by the choice of fine-tuning data and objective rather than by parameter count. Figure 2 shows the joint distribution of average score and dispersion, colored by architectural family, and illustrates that the multi-task-fine-tuned families (ST5, GTR, and the Sentence-Transformers MiniLM/MPNet lineage) occupy the high-average, low-dispersion region, while the single-objective decoder variants (SGPTnli, SGPT-msmarco) and the un-fine-tuned or narrowly fine-tuned baselines (GloVe, Komninos, BERT, SPECTER, LASER2) occupy the lower-average, higher-dispersion region [1].

5.4 Specialization archetypes from hierarchical clustering

The Ward-linkage clustering (Fig. 3), cut at four clusters, recovers groups that align closely with the training-objective families summarized in Section 2 and are described in Table 4. Cluster A (15 of 30 models) is a large, low-dispersion "balanced generalist" group spanning the ST5, GTR, and multi-task Sentence-Transformer lineages together with Contriever and coCondenser; its mean z-profile is positive and roughly flat across all seven tasks (approximately + 0.4 to + 0.6 on six of seven tasks). Cluster B (7 models) groups the higher-capacity, retrieval/reranking-fine-tuned decoder models (the SGPTmsmarco family from 1.3B to 7.1B parameters, plus MiniLM-L12, MPNet, and SimCSEBERT-sup); this cluster is strong on Retrieval and Reranking ($z \approx + 0.65$) but exhibits a pronounced Summarization deficit (mean $z = - 1.46$), the most extreme single-task weakness of any cluster, plausibly reflecting the fact that retrieval-oriented contrastive fine-tuning on query-document pairs actively pulls the embedding geometry away from the kind of paraphrase-level similarity that the summarization evaluation rewards. Cluster C (5 models: GloVe, Komninos, SimCSE-BERT-unsup, LaBSE, SGPT-125M-nli) is a below-average, higher-dispersion group with a comparatively better relative position on Summarization than on any other task, an inversion of Cluster B's profile. Cluster D (3 models: BERT with naive mean-pooling, SPECTER, and LASER2) is the weakest cluster overall, uniformly below average on every task, consistent with these three models not having received the kind of general-purpose contrastive fine-tuning that distinguishes Clusters A and B from the field [30].

Table 4

Four-cluster solution from Ward-linkage clustering of models by standardized seven task profile. Dispersion column reports the cluster-mean profile dispersion index

Cluster	Representative models	Profile characterization	Mean dispersion
A (n = 15)	ST5-*,GTR-*, MPNet, MiniLM-L6, Contriever, coCondenser	Balanced generalist; positive, near-flat z-profile across all seven tasks	Low
B (n = 7)	SGPT-msmarco (1.3B–7.1B), MiniLM-L12,MP- Net, SimCSE-sup	Retrieval/Reranking specialist; strong Retrieval/Reranking, weak Summarization	High
C (n = 5)	GloVe,Komninos, SimCSE- unsup, LaBSE,SGPT- 125M-nli	Below-average generalist with relative Summarization strength	High
D (n = 3)	BERT (meanpool), SPECTER, LASER2	Weakest cluster; uniformly below average	Moderate– high

6 Discussion

6.1 Implications for model selection

Beyond the specific correlation and clustering results, this analysis has a bearing on the more general cost-performance calculus a practitioner faces when choosing an embedding model. It is tempting to assume that because average score rises with model scale (Section 5), the safest strategy under uncertainty about future task requirements is simply to deploy the largest available model. Our dispersion analysis complicates this assumption: the largest model in our dataset by parameter count, SGPT-BLOOM7.1B-msmarco, belongs to Cluster B rather than to the low-dispersion Cluster A, meaning it is not, in fact, the most cross-task-consistent choice available, despite being by far the most computationally expensive one to serve at inference time. A mid-sized Cluster A model such as GTR-Large (335M parameters) or MPNet (109M parameters) achieves a substantially lower serving cost while remaining within the low-dispersion, high-average generalist cluster, which suggests that for teams whose downstream task mix is uncertain or heterogeneous, the profile-level clustering introduced here is at least as decision-relevant as the raw average-versus-scale trade-off that dominates most public discussion of embedding model selection [1, 27].

The negative correlation between profile dispersion and average score (Section 5) has a direct practical implication for practitioners choosing an embedding model for a system that must serve more than one

downstream task: chasing the highest singletask score, rather than the highest average score, is more likely to import Cluster B's specialization trade-off – an appreciable summarization or classification penalty – than choosing among the top few average-ranked models, which this analysis shows tend to already be internally the most balanced. This is a subtly different and more actionable recommendation than the original MTEB paper's headline finding that "no model dominates every task," because it further specifies where the dominance gaps concentrate (predominantly in Cluster B's Summarization weakness and, to a lesser degree, Cluster C's broad below-average performance) rather than treating non-dominance as diffusely distributed across the whole leaderboard [1].

6.2 Implications for scaling embedding models

The absence of any relationship between profile dispersion and parameter count (Section 5) suggests that simply scaling up an existing single-objective fine-tuning recipe – as the SGPT-msmarco family does from 125M to 7.1B parameters – raises the ceiling on the tasks that objective already favors without correcting the tasks it does not, since the 7.1B-parameter SGPT-msmarco variant remains in the same high-dispersion Cluster B as its 1.3B-parameter sibling. Cross-task balance, in this dataset, is a property of what a model is fine-tuned on rather than of how large it is. This is consistent with, and provides a more precisely quantified companion to, the original MTEB paper's qualitative observation that "the field has yet to converge on a universal text embedding method," and it suggests that future embedding research aiming for genuine generality should prioritize training-data diversity and multi-objective fine-tuning curricula (as instantiated by the ST5, GTR, and multi-task Sentence-Transformer families in our Cluster A) over further scaling of a narrowly-targeted contrastive objective [1, 27].

6.3 Implications for benchmark reporting practice

Because MTEB's own published Average column is, by construction, a simple arithmetic mean across constituent datasets that are unevenly distributed across the seven task families (with Retrieval, Classification, and Clustering contributing the most datasets and Summarization contributing only one), a leaderboard reader comparing two models on Average alone cannot distinguish a Cluster A generalist from a Cluster B specialist unless they inspect the full per-task breakdown themselves. We suggest that a profile dispersion index of the kind introduced here, or an equivalent standardized-variance statistic, could be reported alongside the Average column on embedding leaderboards at negligible additional computational cost, since it is derived entirely from scores the benchmark already computes; doing so would let practitioners filter directly for "balanced" versus "specialized" models rather than inferring this property indirectly from the raw per-task table, and would make explicit the task-imbalance caveat that the original MTEB authors themselves noted as a limitation of the unweighted average [1].

6.4 Implications for retrieval-augmented generation and multi-stage pipelines

The Cluster B "retrieval specialist" profile identified in Section 5 is of particular relevance to the design of retrieval-augmented generation systems, in which a single embedding model is frequently reused for corpus indexing, candidate reranking, and sometimes near-duplicate detection or clustering of ingested

documents within the same pipeline. Our results indicate that a model drawn from Cluster B will likely serve the indexing and reranking stages well, consistent with its high Retrieval and Reranking z-scores, but that the same model is a comparatively poor default choice if the pipeline also needs to embed and compare short abstractive summaries or paraphrase-like text, given Cluster B’s pronounced Summarization deficit (mean z-score – 1.46, the single most extreme cluster-level weakness observed anywhere in the dataset). Conversely, a Cluster A generalist sacrifices a few points of peak retrieval performance relative to the strongest Cluster B specialist but avoids this kind of stage-specific failure mode entirely, which may be the more risk-averse choice for a production system whose full set of embedding use cases is not yet completely enumerated at model-selection time – a common situation in practice, since RAG and search systems tend to accrete new embedding-dependent features (deduplication, topic clustering, related-content recommendation) after the initial retrieval pipeline is already in production [1, 2].

6.5 Relationship to the original MTEB correlation analysis

Our Table 1 corroborates, with exact coefficients, the qualitative description accompanying the original MTEB paper’s own correlation figure – namely that Clustering and Reranking correlate most strongly with each other while Summarization correlates weakly with the rest of the suite. Our contribution beyond that original analysis is threefold: we convert a qualitative heatmap description into an exact, reusable coefficient matrix; we introduce the profile dispersion index and connect it, for the first time, to both average performance and model scale; and we move from pairwise task correlation to a full multivariate clustering of models, which is what actually reveals the Cluster A/B/C/D specialization archetypes rather than only the pairwise task associations that motivate them [1].

6.6 Comparison with the subsequent scale-up of the benchmark

It is worth explicitly situating the thirty-model, seven-task analysis undertaken here relative to the scale the MTEB ecosystem has since reached. The Massive Multilingual Text Embedding Benchmark extends the original suite to several hundred tasks across more than 250 languages and adds task categories – instruction-following evaluation, long-document retrieval, and code retrieval – that did not exist in the original release and that plausibly have their own distinct correlation and specialization structure. Independently, survey work cataloguing the top-performing entries on the public MTEB leaderboard has documented that the field has since produced embedding models an order of magnitude larger than the 7.1-billion-parameter SGPT-BLOOM variant that was the largest model in our thirty-model dataset, including instruction-tuned largelanguage-model-based embedding approaches that did not exist at the time of MTEB’s original release. Whether the qualitative pattern we report – average performance rising with scale while cross-task dispersion remains scale-invariant – continues to hold among this newer, larger, and more architecturally diverse population of models is, as noted in Section 7, an open question, but the methodology introduced here (standardized profile dispersion, family-conditioned scatter analysis, and Ward clustering of task profiles) transfers directly to that larger dataset without

modification and could be applied to the current public leaderboard export as a natural extension of this work [17, 24–26].

7 Limitations

Several limitations qualify these conclusions. First, the analysis is restricted to the thirty models and fifty-six English datasets evaluated in the original 2022 MTEB release; the benchmark and its public leaderboard have since grown to include several hundred additional models, and the more recent Massive Multilingual Text Embedding Benchmark substantially expands both the task and language coverage. Whether the correlation structure, the dispersion-average relationship, and the cluster archetypes reported here replicate on the larger, more recent leaderboard – which includes newer instruction-tuned and much larger LLM-based embedding models – is an open empirical question that we explicitly flag for future work rather than address here, since doing so would require re-scraping and re-validating a substantially larger and still-evolving results table. Second, the profile dispersion index, by standardizing each task column across the specific sample of thirty models, is sample-dependent: adding or removing models changes each task’s mean and standard deviation and could shift dispersion values somewhat, although the relative ordering of models is unlikely to change substantially given the moderate sample size and the qualitative stability of the Pearson/Spearman comparison in Section 5. Third, the four-cluster solution was chosen for interpretability rather than by an automated model-selection criterion (e.g., the gap statistic or silhouette analysis); we report this cut because it produces the most substantively interpretable and stable grouping upon visual inspection of the dendrogram in Fig. 3, but acknowledge that alternative cluster counts are defensible and are provided in the supplementary material for comparison. Fourth, because MTEB’s task-family scores are themselves averages over multiple underlying datasets of unequal size and difficulty, some of the observed cross-task correlation could in principle reflect shared dataset-level artifacts (e.g., topical overlap between certain clustering and retrieval datasets drawn from similar corpora) rather than a purely architectural or objective-driven similarity between task families; disentangling dataset-level from taskfamily-level correlation is left to future work using the full per-dataset results table rather than the task-family averages analyzed here. Fifth, the low across-model variability we observe on Summarization (Table 2) means that our conclusions about Summarization specifically – including Cluster B’s apparent deficit on this task – rest on a comparatively narrow numerical range and a single constituent dataset (SummEval), and should be interpreted with more caution than the conclusions drawn about task families backed by ten or more constituent datasets, such as Retrieval or Classification. Finally, this is a secondary, observational analysis of existing benchmark results rather than a controlled intervention; the associations we report (dispersion with average score, dispersion with scale) are correlational and should not be read as demonstrating that any specific fine-tuning choice causes a particular profile shape, only that the choices instantiated by the models in this sample are associated with the profile differences described. A randomized or ablation-based study that fine-tunes otherwise identical base models on different objective mixtures would be needed to support a causal claim, and we explicitly flag this as the natural next step for future methodological work in this area [1, 17].

8 Conclusion

Revisiting the original Massive Text Embedding Benchmark results through the lens of cross-task consistency, rather than aggregate score alone, surfaces two findings with direct relevance to both embedding-model practitioners and future benchmark design. Stronger models in this benchmark tend to be more uniformly strong across task families rather than narrowly dominant on a subset of them, but this cross-task balance is not a simple by-product of increasing model scale; it appears instead to be governed by the diversity of a model's fine-tuning objective. A Ward-linkage clustering of models by their standardized task profile recovers four interpretable specialization archetypes, most notably a large "balanced generalist" cluster built on multi-task contrastive fine-tuning and a smaller "retrieval specialist" cluster that trades away summarization competence for retrieval and reranking strength. We recommend that a simple dispersion statistic of the kind introduced here be reported alongside the conventional average score on embedding benchmark leaderboards, since it can be computed at no additional evaluation cost from data the benchmark already produces and materially improves a practitioner's ability to select a model appropriate to single-task versus multi-task deployment [1, 4, 9, 10, 11, 14].

More broadly, this paper illustrates a general methodological point that extends beyond text embeddings: whenever a benchmark aggregates heterogeneous task families into a single leaderboard number, that aggregate should be accompanied by, at minimum, a standardized dispersion statistic and a task-profile clustering of the evaluated systems, both of which are inexpensive to compute from data the benchmark has already collected and both of which materially change the practical guidance a reader would take away from the leaderboard. Future work should extend this profilelevel analysis to the substantially larger and more architecturally diverse model set now available on the public MTEB and MMTEB leaderboards, should test whether the dispersion-scale independence reported here persists among instruction-tuned and large-language-model-based embedding approaches that postdate the original benchmark release, and should investigate, through controlled fine-tuning ablations rather than observational analysis of pre-existing models, whether specific mixtures of finetuning objectives can be identified that reduce cross-task dispersion without sacrificing average performance – effectively searching for the "universal" text embedding method whose absence the original MTEB paper first documented.

Declarations

Ethics approval

This study did not involve human participants, human data, animal subjects, or biological materials; it is a secondary statistical analysis of previously published, publicly available benchmark results. No ethics approval was required.

Consent for publication

Not applicable; this study is a secondary analysis of previously published, publicly available benchmark results and did not involve human participants.

Competing interests

The source manuscript does not state whether the author has competing interests. Please confirm this declaration before submission.

Funding

Funding information was not provided in the source manuscript and should be confirmed by the corresponding author before submission.

Author Contribution

Z.Z. conceptualized the study, developed the methodology, conducted the data curation and formal analysis, performed the investigation, developed the software and visualizations, and wrote the original manuscript. Z.Z. reviewed and edited the manuscript and approved the final version.

Data Availability

The data used in this study are publicly available from the Massive Text Embedding Benchmark (MTEB) repository. The analysis was conducted using the publicly released per-model, per-task evaluation results. Derived analytical tables and code supporting the findings are available from the corresponding author upon reasonable request.

References

1. Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2014–2037. Association for Computational Linguistics, Dubrovnik, Croatia.
2. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:210408663*.
3. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China.
4. Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp.

- 6894–6910. Association for Computational Linguistics, Punta Cana, Dominican Republic.
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, USA.
 6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.
 7. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar.
 8. Ni, J., Qu, C., Lu, J., Dai, Z., Hernández Ábrego, G., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M. W., & Yang, Y. (2021). Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:211207899*.
 9. Ni, J., Hernández Ábrego, G., Constant, N., Ma, J., Hall, K. B., Cer, D., & Yang, Y. (2021). Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.
 10. Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). MPNet: Masked and permuted pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 33, pp. 16857–16867.
 11. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep selfattention distillation for task-agnostic compression of pre-trained transformers. In: Advances in Neural Information Processing Systems, vol. 33, pp. 5776–5788.
 12. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.
 13. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:211209118*.
 14. Muennighoff, N. (2022). SGPT: GPT sentence embeddings for semantic search. *arXiv preprint arXiv:220208904*.
 15. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2270–2282. Association for Computational Linguistics, Online.
 16. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
 17. Enevoldsen, K. (2025). MMTEB: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:250213595*.
 18. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multitask benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:180407461*.

19. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In: *Advances in Neural Information Processing Systems*, vol. 32.
20. Conneau, A., & Kiela, D. (2018). SentEval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:180305449*.
21. Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 410–420. Association for Computational Linguistics, Prague, Czech Republic.
22. Gao, L., & Callan, J. (2021). Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:210805540*.
23. Heffernan, K., Çelebi, O., & Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:220512654*.
24. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C. (2022). Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:220110005*.
25. Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M. (2022). BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:221105100*.
26. Recent advances in text embedding: a comprehensive review of top-performing methods on the MTEB benchmark. *arXiv preprint arXiv:2406.01607* (2024).
27. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
28. Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.
29. Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L. (2021). Datasets: A community library for natural language processing. *arXiv preprint arXiv:210902846*.
30. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Figures

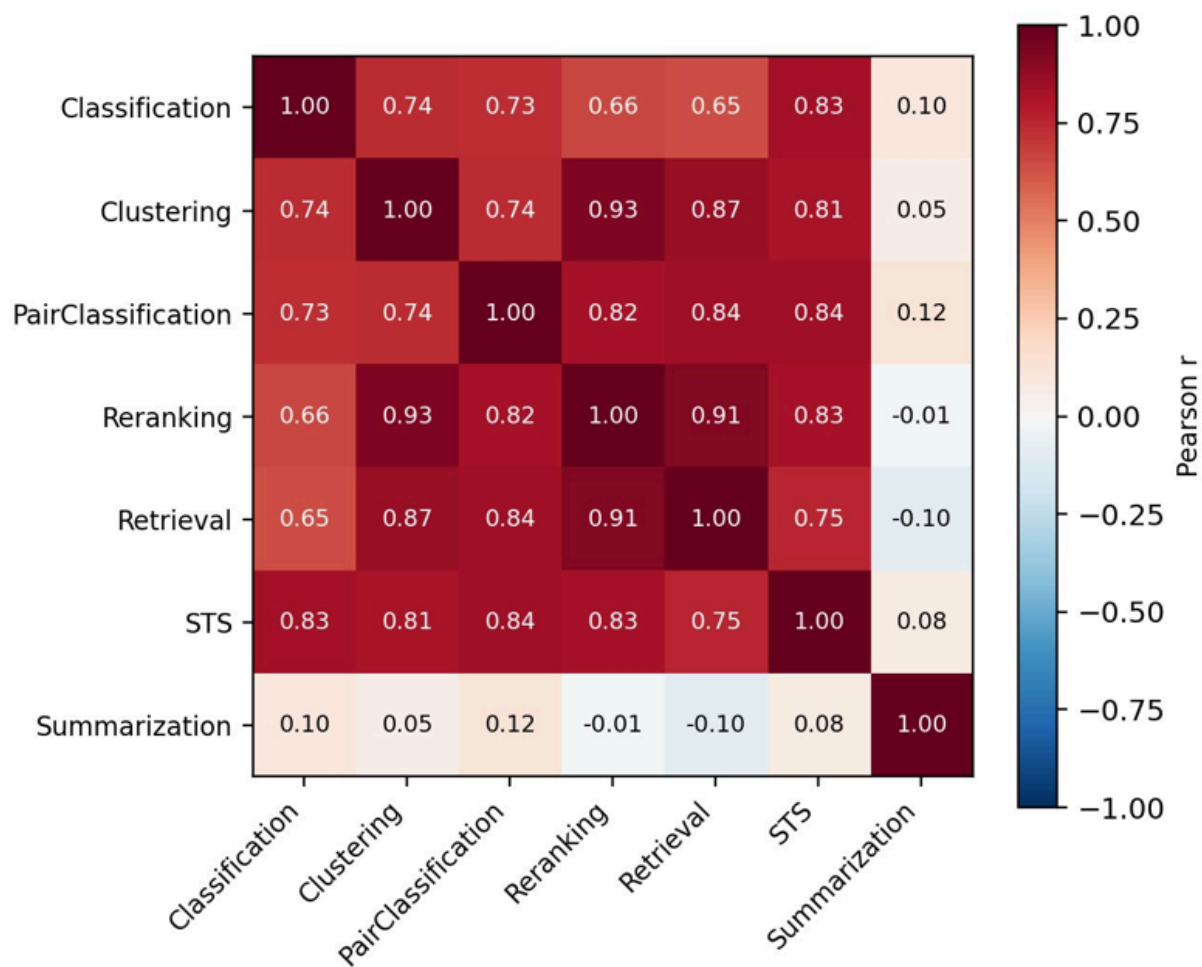


Figure 1

Cross-task Pearson correlation matrix across thirty MTEB English models. Darker red indicates stronger positive correlation; darker blue would indicate negative correlation. The Reranking–Clustering–Retrieval triad forms the most strongly inter-correlated block, while Summarization is nearly orthogonal to every other task

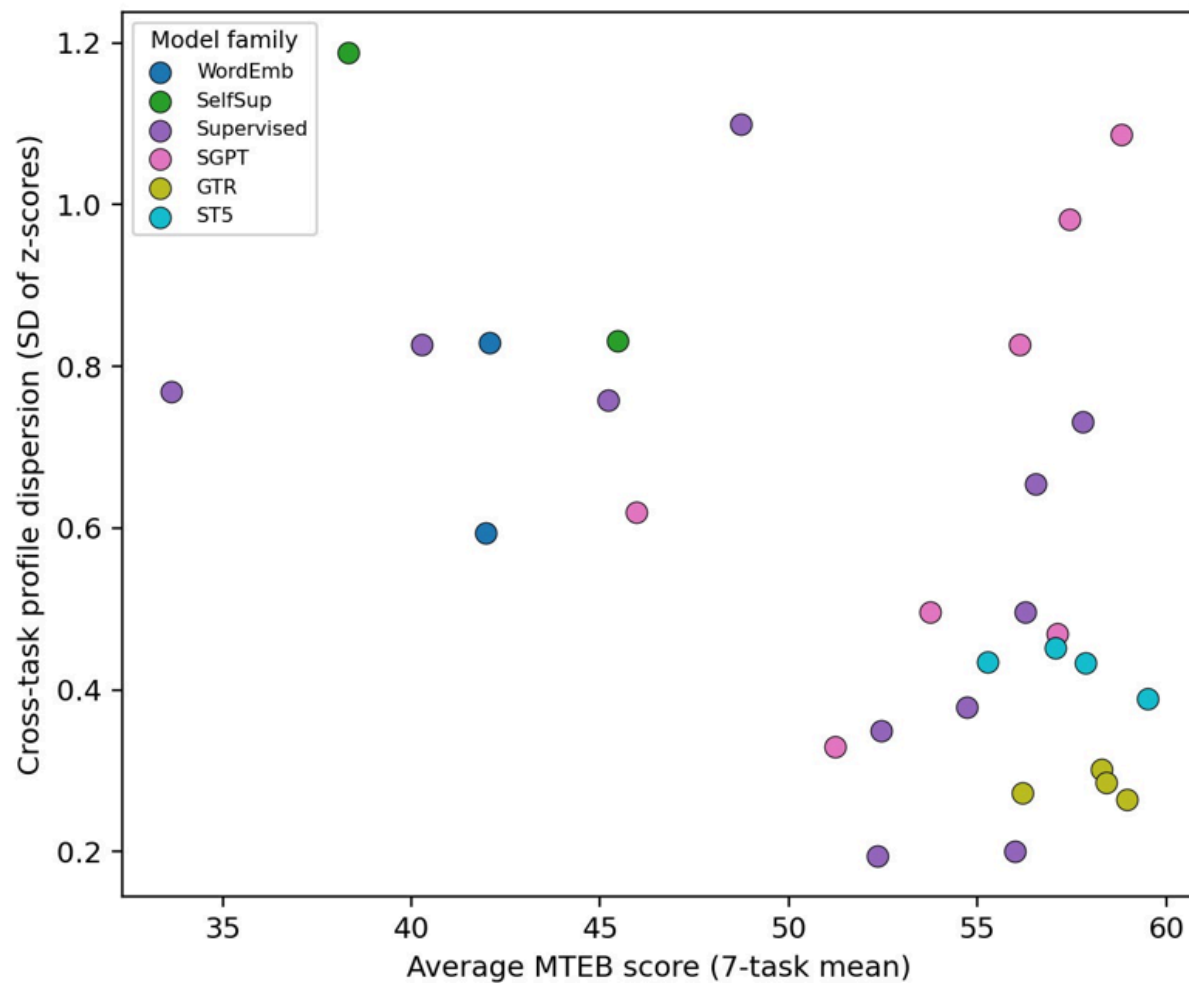


Figure 2

Cross-task profile dispersion (standard deviation of within-task z-scores) plotted against average MTEB score for thirty models, colored by architectural family. Lower-right models are simultaneously strong and evenly balanced across task families; upper-left and upper-right models achieve their scores unevenly across tasks

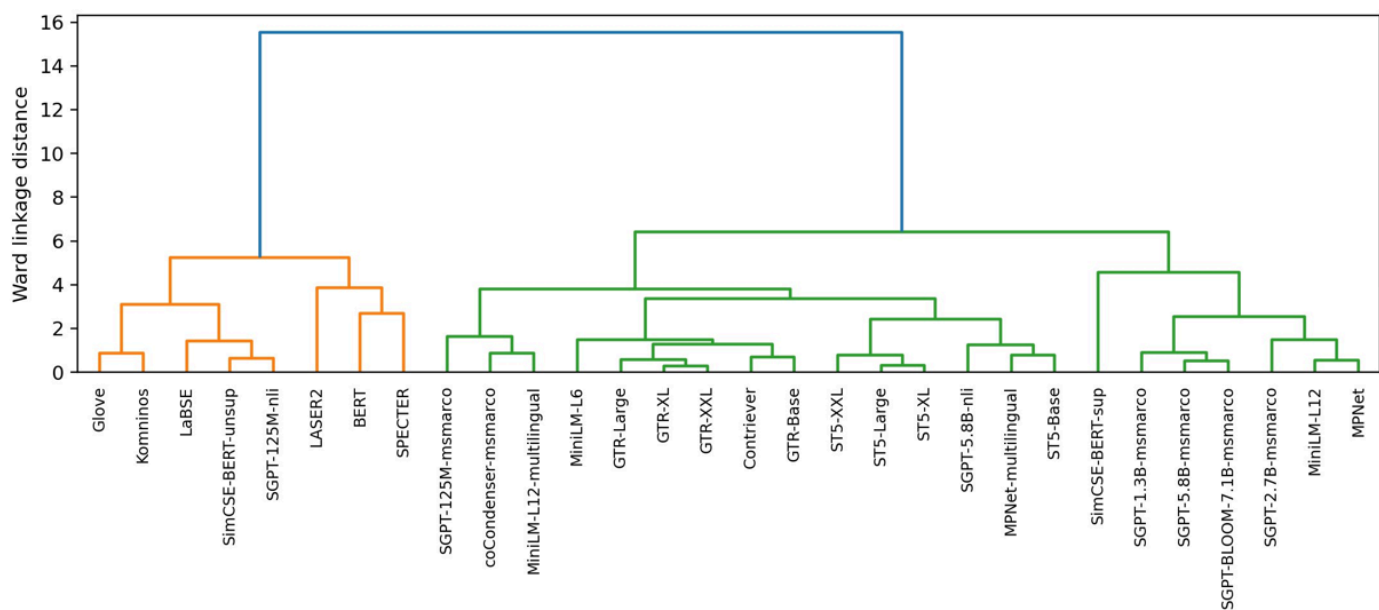


Figure 3

Ward-linkage hierarchical clustering dendrogram of thirty MTEB English models based on their seven-dimensional standardized task-family profile. Four clusters are highlighted at the cut height used in Table 4

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [analysis.py](#)
- [augmentedresults.csv](#)