Beyond Redundancy: Diagnosing Information Distribution in Text Embeddings via Task-Aware Dimension Selection

Anonymous ACL submission

Abstract

Text embedding models often remain robust after substantial random dimension removal, suggesting considerable redundancy in high-dimensional representations. However, such robustness does not reveal how task-relevant information is distributed across dimensions: whether it is concentrated in a few critical dimensions or spread broadly across the embedding space. We diagnose this information distribution across 11 embedding models and 34 MTEB tasks by comparing random, task-aware optimized, and anti-optimized dimension selection, and by transferring dimension rankings across tasks. At 256 retained dimensions, optimized selection outperforms random by only +2.4–5.3% for Retrieval-optimized Embedders and most Instruction-conditioned Embedders, but by +8.2–11.8% for RoBERTa-InBedder and General-purpose Language Model Backbones. Under stronger compression, the optimized–random gap increases, indicating that broad redundancy coexists with exploitable structure across embedding dimensions. Cross-task transfer shows that dimension rankings are largely uncorrelated across tasks, yet transferred selections remain slightly above random across retained dimensionalities, revealing task-specific rankings together with shared redundancy. Entropy analysis shows near-uniform yet model-dependent importance distributions, explaining why many dimension subsets preserve performance while some models exhibit stronger concentration. Overall, task-aware dimension selection reveals a training-paradigm-dependent information distribution.

1 Introduction

Dense text embeddings are widely used in semantic search, clustering, classification, and reranking (Reimers and Gurevych, 2019; Karpukhin et al., 2020; Khattab and Zaharia, 2020), but their 768–4096-dimensional vectors impose substantial storage and computation costs at scale (Johnson et al., 2021). Recent work has shown that randomly removing half of the dimensions from modern text embeddings causes only modest degradation on retrieval and classification tasks (Takeshita et al., 2025), with similar robustness reported under more aggressive compression settings (Tsukagoshi and Sasano, 2025). These findings suggest considerable redundancy in text embedding representations.

However, redundancy alone does not reveal whether task-relevant information is concentrated in a small set of critical dimensions or broadly distributed across the embedding space. This distinction has a direct empirical implication: if information is broadly distributed, random dimension selection should perform close to task-aware optimized selection; if information is more concentrated, optimized selection should yield a clear advantage. We therefore use the optimized–random gap as a diagnostic of information distribution, and measure this gap across 11 embedding models and 34 MTEB tasks (Muennighoff et al., 2023) at different compression levels.

Beyond asking how concentrated task-relevant information is within individual tasks, we also ask whether dimension importance is task-specific. If each task relies on its own set of important dimensions, then rankings derived from one task should transfer poorly to another. If tasks share redundant structure, transferred selections may preserve performance even when rankings disagree. We test this through cross-task transfer, ranking dimensions on a source task and evaluating the retained dimensions on a target task.

We observe that random selection already approaches task-aware selection under moderate compression for Retrieval-optimized Embedders and most Instruction-conditioned Embedders. For RoBERTa-InBedder and General-purpose Language Model Backbones, task-aware selection yields larger gains, suggesting stronger information concentration. Cross-task transfer further shows

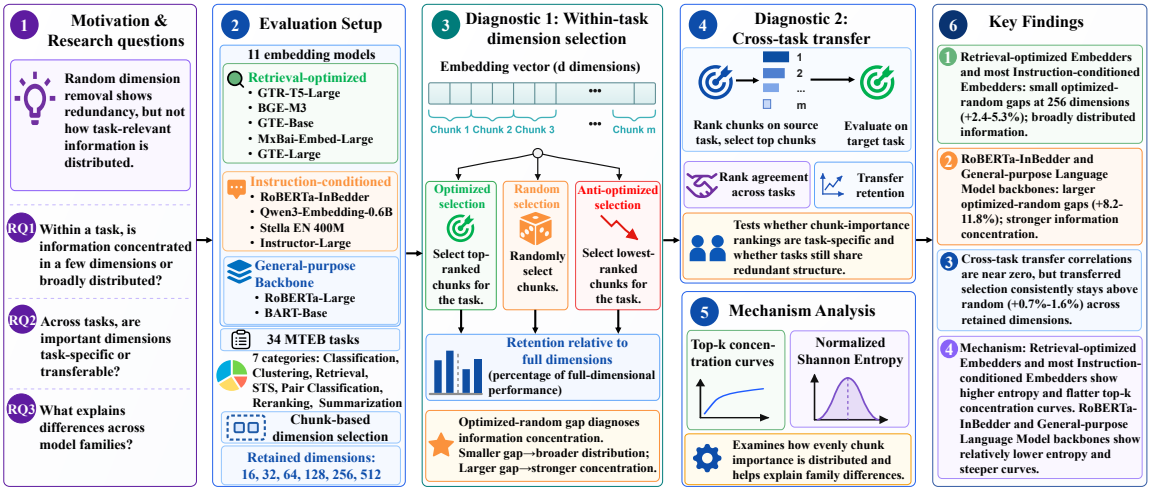

Figure 1: Research framework and main findings. We diagnose how task-relevant information is distributed across embedding dimensions by comparing optimized, random, and anti-optimized chunk selection within tasks, transferring chunk rankings across tasks, and analyzing chunk-importance concentration and entropy. Results show that task-relevant information is broadly distributed but not structureless: concentration varies by training paradigm, and task-specific rankings coexist with shared redundancy.

that dimension rankings vary substantially across tasks without causing proportional performance loss, indicating that shared redundancy buffers the impact of task-specific rankings. Figure 1 summarizes our research framework and key findings.

Our contributions are:

1. We propose dimension selection as a diagnostic for studying how task-relevant information is distributed across embedding dimensions.
2. We show that task-relevant information distribution differs by training paradigm and is clearer under stronger compression.
3. We demonstrate that task-specific dimension rankings coexist with shared redundancy.

2 Related Work

2.1 Embedding Redundancy and Compressibility

Evidence of redundancy. Recent work shows that text embeddings contain substantial dimensional redundancy. Takeshita et al. (2025) find that randomly removing 50% of dimensions causes only modest degradation on retrieval and classification tasks. Tsukagoshi and Sasano (2025) show that prompt-based LLM embeddings remain robust under aggressive dimensionality reduction. Zhang et al. (2024b) compare unsupervised dimensionality reduction methods, showing PCA can halve embedding size with limited loss.

Compression-oriented methods. Several methods take advantage of the redundancy of embedding vectors. Kusupati et al. (2022) train embeddings where the first k dimensions are optimized to be a good representation at any truncation level, with follow-up works adapting this idea to LLM embedding compression (Yoon et al., 2024; Zhang et al., 2025a). Sparse-coding approaches further relax the contiguous-prefix constraint (Wen et al., 2025).

Our distinction. These works show that embeddings are highly redundant and compressible, but treat compression as an objective. We instead use dimension selection as a diagnostic for studying whether task-relevant information is concentrated in a few critical dimensions or broadly distributed.

2.2 Task-specific Compression and Representation Structure

Task-specific compression. Several methods try to exploit task-specific dimension importance. Faggioli et al. (2024) use retrieval feedback to estimate dimension utility, and Wu et al. (2026) predict dimension importance from query embeddings, while Tang and Yang (2026) use gradient alignment for domain-specific pruning.

Representation structure. The broader literature examines geometric and effective-dimensional properties of neural representations (Ethayarajh, 2019; Li et al., 2018; Aghajanyan et al., 2021), showing that high-dimensional representations of-

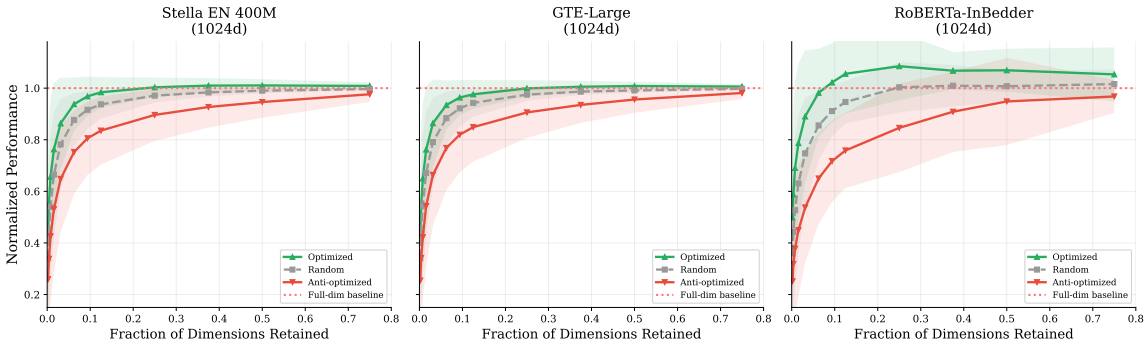

Figure 2: Retention across retained dimensionalities for three representative models. Optimized (green) retains the highest-ranked chunks, Anti-optimized (red) retains the lowest-ranked chunks, and Random (gray) provides the random baseline. The optimized–random gap remains modest for GTE-Large and Stella EN 400M under moderate compression, but is larger for RoBERTa-InBedder and widens under stronger compression. Shaded bands show 95% bootstrap confidence intervals over tasks.

ten exhibit lower-dimensional or anisotropic structure across models and tasks.

Our distinction. Rather than only estimating which dimensions are important for one task, we ask whether dimension importance is task-specific, and whether different tasks still share redundant structure. We use cross-task transfer as a diagnostic, ranking dimensions on a source task and evaluating them on target tasks.

3 Methodology

3.1 Models and Tasks

Models. We evaluate 11 text embedding models with diverse architectures and training paradigms, which we group into 3 families. *General-purpose Language Model Backbones* include RoBERTa-Large (Liu et al., 2019) and BART-Base (Lewis et al., 2020), which are not originally designed as embedding models. *Instruction-conditioned Embedders* include RoBERTa-InBedder (Peng et al., 2024), Qwen3-Embedding-0.6B (Zhang et al., 2025b), Stella EN 400M (Zhang et al., 2024a), and Instructor-Large (Su et al., 2023), whose embedding ability is added or strengthened through post-training adaptation. *Retrieval-optimized Embedders* include GTR-T5-Large (Ni et al., 2022), BGE-M3 (Chen et al., 2024), GTE-Base, MxBai-Embed-Large, and GTE-Large (Li et al., 2023), which are trained directly for retrieval-style representation learning. Appendix A summarizes which evaluations are available for each model.

Tasks. We use 34 benchmark tasks from MTEB (Muennighoff et al., 2023) covering 7 categories: 11 classification tasks (e.g., AmazonReviewsClassification, Banking77Classification), 3 clustering tasks (BiorxivClusteringS2S, MedrxivClusteringS2S, TwentyNewsgroupsClustering), 5 retrieval tasks (ArguAna, CQADupstackEnglishRetrieval, NFCorpus, SCIDOCS, SciFact), 3 reranking tasks (AskUbuntuDupQuestions, SciDocsRR, StackOverflowDupQuestions), 8 STS tasks (BIOSSES, SICK-R, STS12–STS16, STSBenchmark), 3 pair classification tasks (SprintDuplicateQuestions, TwitterSemEval2015, TwitterURLCorpus), and 1 summarization task (SummEval).

3.2 Dimension Selection as a Diagnostic

For a D-dimensional embedding, a reduced representation is defined by a retained dimension set $A \subseteq \{1, \ldots, D\}$ with $|A| = k$. For model M, task T, and retained dimensions A, let $S^{M,T}(A)$ denote the MTEB main metric (e.g., accuracy, nDCG@10) using only dimensions in A. We report retention relative to the full-dimensional representation:

$$R^{M,T}(A) = \frac{S^{M,T}(A)}{S^{M,T}(\text{Full})} \times 100\% \qquad (1)$$

$R^{M,T}(A) > 100\%$ indicates that dimension selection may improve task performance relative to the full-dimensional baseline. Task averages are reported with 95% bootstrap confidence intervals, and paired comparisons are conducted using paired t-tests and Wilcoxon signed-rank tests, with Holm–Bonferroni correction applied.

Information concentration. To test whether task-relevant information is concentrated in a few critical dimensions or broadly distributed, we compare random selection with task-aware optimized selection. Random selection samples k dimensions

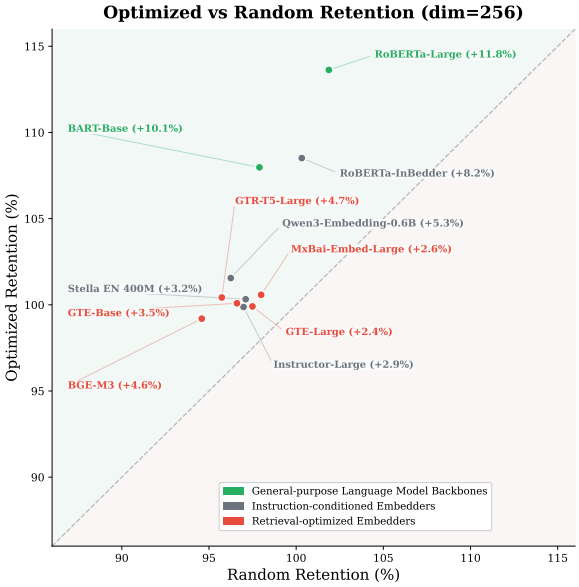

Figure 3: Mean random retention (x-axis) versus mean optimized retention (y-axis) at 256 retained dimensions. The dashed identity line marks parity, and vertical distance above the line corresponds to optimized–random gap. Green marks General-purpose Language Model Backbones, gray marks Instruction-conditioned Embedders, and red marks Retrieval-optimized Embedders.

uniformly without replacement. We repeat this $B = 10$ times with different random seeds and report the mean retention:

$$\bar{R}^{M,T}_{\text{rand}}(k) = \frac{1}{B}\sum_{b=1}^{B} R^{M,T}(A^{(b)}_{\text{rand}}(k)) \quad (2)$$

where $A^{(b)}_{\text{rand}}(k)$ is the retained dimension set in the b-th trial. For optimized selection, we divide the D dimensions into $N = D/w$ contiguous chunks and retain $m = k/w$ chunks. We primarily use $w = 2$ to balance detail and cost, and further evaluate sensitivity to chunk size as an auxiliary validation. Each chunk is scored by keeping only that chunk:

$$q^{M,T}_i = S^{M,T}(C_i) \quad (3)$$

where C_i denotes the i-th chunk. We rank chunks by $q^{M,T}_i$ and retain the m highest-scoring chunks:

$$A^{M,T}_{\text{opt}}(k) = \text{Top}_m\left(\{q^{M,T}_i\}_{i=1}^{N}\right) \quad (4)$$

We define the optimized–random gap as:

$$\Delta^{M,T}_{\text{gap}}(k) = R^{M,T}(A^{M,T}_{\text{opt}}(k)) - \bar{R}^{M,T}_{\text{rand}}(k) \quad (5)$$

where a smaller gap indicates that random subsets are nearly as informative as task-aware ones, indicating broadly distributed task-relevant information. A larger gap instead suggests stronger concentration in top-ranked chunks.

Lower bound. We also evaluate anti-optimized selection, retaining m lowest-scoring chunks:

$$A^{M,T}_{\text{anti}}(k) = \text{Bottom}_m\left(\{q^{M,T}_i\}_{i=1}^{N}\right) \quad (6)$$

If anti-optimized selection substantially underperforms random selection, then dimensions do carry task-relevant information, although that information is broadly distributed enough for random selection to remain strong.

3.3 Cross-task Transfer

To test whether dimension importance is task-specific and whether different tasks nevertheless share redundant structure, we transfer dimension rankings across tasks. For source task T_s and target task T_t, we rank chunks by source-task scores and evaluate the selected chunks on the target task:

$$R^{M}_{s\to t}(k) = R^{M,T_t}\left(\text{Top}_m\left(\{q^{M,T_s}_i\}_{i=1}^{N}\right)\right) \quad (7)$$

We compare each transfer result with random selection on the same target task:

$$\Delta^{M}_{s\to t}(k) = R^{M}_{s\to t}(k) - \bar{R}^{M,T_t}_{\text{rand}}(k) \quad (8)$$

We also compute source–target rank agreement:

$$\rho^{M}_{s,t} = \text{Spearman}\left(\{q^{M,T_s}_i\}_{i=1}^{N}, \{q^{M,T_t}_i\}_{i=1}^{N}\right) \quad (9)$$

Together, $\Delta^{M}_{s\to t}(k)$ and $\rho^{M}_{s,t}$ disentangle functional transfer from ranking agreement. The former measures how well a source-derived ranking transfers to the target task relative to random, while the latter captures the rank correlation between source and target importance scores.

3.4 Auxiliary Experiments

We conduct auxiliary experiments to test whether the main diagnostic conclusions are robust under alternative compression settings, data splits, and importance definitions. The pruning-ratio sweep examines how the optimized–random gap and cross-task transfer behavior change as fewer dimensions are retained. Train/test split validation tests whether optimized rankings generalize beyond the data used to construct them. Mechanism analysis further uses entropy and cumulative importance to characterize whether dimension importance is near-uniform or concentrated.

We also test robustness to chunk-scoring granularity, random subset variability, practical selection baselines, and alternative ways of estimating dimension importance.

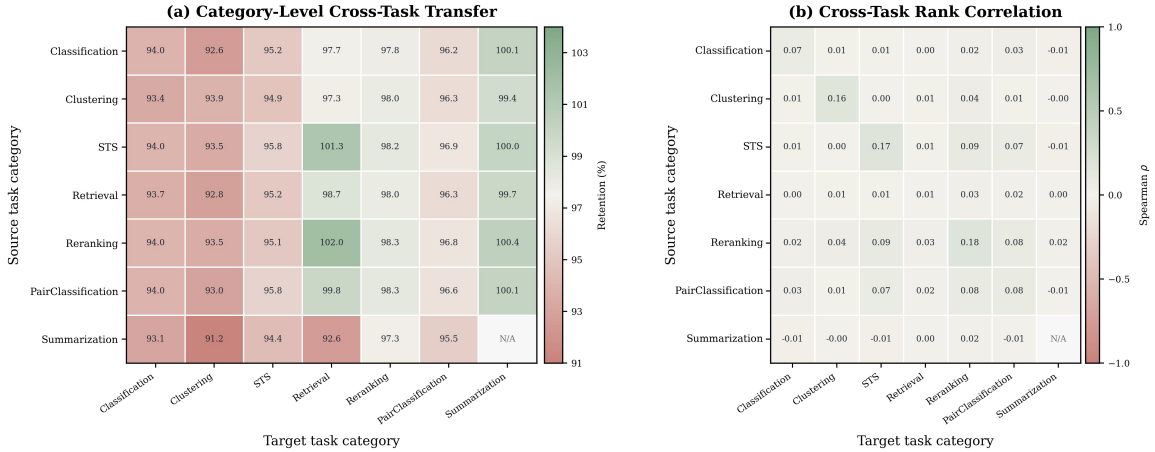

Figure 4: Category-level cross-task transfer and rank agreement at 256 retained dimensions, excluding same-task source–target pairs. Rows denote source task categories and columns denote target task categories. N/A appears because the Summarization category contains only one task. (a) Transfer retention averaged over source–target task categories. Retention remains stable across category pairs. (b) Spearman rank correlations between source-task and target-task chunk-importance rankings, averaged over category pairs. Correlations are close to zero across most pairs. Together, these results show that task-specific rankings coexist with shared redundancy across tasks.

Table 1: Optimized–random retention gap at 256 retained dimensions. Δ reports the mean retention difference between optimized and random selection for each model with 95% bootstrap CI. d is the Cohen's d for paired per-task differences. p_{holm} reports Holm-Bonferroni corrected p values across 11 tests.

Model	Δ	95% CI	d	p_{holm}
Retrieval-optimized Embedders				
GTR-T5-Large	+4.69%	[+2.91%, +6.61%]	0.84	<0.001
BGE-M3	+4.61%	[+3.17%, +6.07%]	1.04	<0.001
GTE-Base	+3.49%	[+2.15%, +4.97%]	0.82	<0.001
MxBai-Embed-Large	+2.59%	[+1.51%, +3.83%]	0.83	<0.001
GTE-Large	+2.41%	[+1.48%, +3.40%]	0.83	<0.001
Instruction-conditioned Embedders				
RoBERTa-InBedder	+8.19%	[+5.96%, +10.62%]	1.17	<0.001
Qwen3-Embedding-0.6B	+5.30%	[+3.54%, +7.28%]	1.00	<0.001
Stella EN 400M	+3.22%	[+2.16%, +4.44%]	0.93	<0.001
Instructor-Large	+2.89%	[+1.80%, +4.06%]	0.83	<0.001
General-purpose Language Model Backbones				
RoBERTa-Large	+11.75%	[+8.61%, +15.53%]	1.16	<0.001
BART-Base	+10.08%	[+6.30%, +13.72%]	0.90	<0.001

4 Experimental Results and Analysis

4.1 Training Paradigms Exhibit Different Concentration Profiles

We first compare optimized, random and anti-optimized selection across compression levels to diagnose how concentrated task-relevant information is across embedding dimensions. Figure 2 shows retention from a single retained dimension up to 75% retention (768 of 1024 dimensions) for three representative models, covering moderate to aggressive compression regimes. For GTE-Large at 256 retained dimensions, Table 2 reports category-level retention, while Appendix B provides the full task-level results across all 34 tasks.

Information concentration differs by training paradigm. Optimized selection consistently outperforms random selection across all 11 models, but the optimized–random gap differs sharply by model family at 256 retained dimensions, as shown in Figure 3 and Table 1. The gaps are smallest among Retrieval-optimized Embedders (+2.4%–4.8%), with random selection retaining 95–98% of full-dimensional performance. Instruction-conditioned Embedders show a similar pattern except for RoBERTa-InBedder: their gaps remain modest at +2.9%–5.3%, with random retention of 95–97%. These small gaps therefore provide diagnostic evidence that task-relevant information is broadly distributed across dimensions for these models. In contrast, RoBERTa-InBedder and General-purpose Language Model Backbones show larger optimized–random gaps that would materially affect downstream performance (+8.2%–11.8%). For these models, task-relevant information is more concentrated in top-ranked chunks, suggesting that dimension concentration is closely associated with training paradigm.

Smaller retained subsets increase sensitivity to chunk choice. The optimized–random gap widens as fewer dimensions are retained. At 256 retained dimensions, the gap is only +2.4–5.3% for Retrieval-optimized Embedders and Instruction-conditioned Embedders other than RoBERTa-InBedder. At 64 retained dimensions, the gap widens to +5.1–9.4% for the same model group.

Table 2: Mean retention (%) by task category for GTE-Large at 256 retained dimensions. Random, Optimized, and Anti-optimized columns report retention ratios for the three strategies relative to the full-dimensional baseline, averaged over tasks within each category. Gap is Optimized–Random gap.

Task Category	Random	Optimized	Gap	Anti-optimized
Classification	97.3	101.4	+4.1	84.2
Clustering	95.1	99.4	+4.3	81.5
Retrieval	94.3	94.3	+0.0	93.3
Reranking	98.8	100.4	+1.6	96.0
STS	99.1	101.1	+2.0	95.5
Pair Classification	99.4	100.9	+1.5	97.2
Summarization	99.8	98.7	-1.1	98.8

Table 3: Pooled cross-task transfer vs. random selection across 11 models. Transfer consistently exceeds random selection across retained dimensionalities. All gap includes diagonal self-transfer, while off-diagonal gap excludes same-task source–target pairs. p_{holm}: Holm-corrected p-values for off-diagonal gaps. Per-model and category breakdowns are shown in Appendix D.

Retained Dims	Transfer	Random	All gap	Off-diagonal gap	p_{holm}
16	64.8%	63.8%	+1.0%	+0.7%	<0.001
32	76.3%	75.3%	+1.0%	+0.7%	<0.001
64	86.7%	85.6%	+1.0%	+0.8%	<0.001
128	94.3%	93.4%	+0.9%	+0.7%	<0.001
256	99.1%	98.3%	+0.8%	+0.7%	<0.001
512	101.5%	99.8%	+1.7%	+1.6%	<0.001

This pattern suggests that performance becomes more sensitive to chunk choice as the retained subset shrinks, making non-uniform dimension importance more visible under stronger compression. The transition point differs across model families, as detailed in Appendix C.

Anti-optimized selection exposes weakly informative dimensions. Anti-optimized selection provides a complementary view of the same information-distribution question by probing the least informative dimensions. At 256 retained dimensions, it degrades retention to 84–91% across the tested models, well below random selection. This shows that dimensions do carry task-relevant information. For Retrieval-optimized Embedders and most Instruction-conditioned Embedders, the small optimized–random gap indicates broadly distributed task-relevant information, but the much lower anti-optimized retention shows that bottom-ranked chunks contribute less to task-relevant information. For RoBERTa-InBedder and General-purpose Language Model Backbones, the larger optimized–random gap indicates stronger concentration in high-ranked chunks, making anti-optimized selection more damaging.

4.2 Task-specific Rankings Coexist with Shared Redundancy

The optimized–random gap diagnoses how task-relevant information is distributed within each task. Cross-task transfer asks a complementary question: whether dimension-importance rankings are task-specific, and whether selections derived from one task still preserve information for another. We therefore rank chunks on a source task, retain the top-ranked chunks, and evaluate the resulting subset on a target task.

Dimension rankings are task-specific. To quantify ranking agreement, we compute Spearman correlations between source-task and target-task standalone chunk-score vectors over the same N chunks. Figure 4(b) reports the results, averaged within each source–target task-category pair. The correlations are close to zero across most category pairs, with low ranking agreement even for within-category transfer, indicating that source-task rankings provide little information about the relative ordering of chunks for target tasks. Dimension importance is therefore strongly task-specific.

Transferred performance reveals shared redundancy. Cross-task transfer consistently exceeds random selection across retained dimensionalities, as shown in Table 3. The transfer gap is positive at every retained dimensionality, ranging from $+0.7$–1.6% after excluding same-task source–target pairs, with Holm-corrected $p < 0.001$ throughout. Appendix D reports per-model breakdowns. Given the near-zero rank agreement above, tasks largely disagree on how dimensions should be ranked. Nevertheless, the consistently above-random transfer shows that source-derived selections can still preserve information useful for other tasks. This indicates that shared redundancy buffers the impact of task-specific rankings.

The shared-redundancy pattern is not category-specific. Figure 4(a) examines whether transfer retention is concentrated in specific source–target category pairs. The heatmap shows no clear diagonal advantage. Instead, retention varies more by target category than by source category, suggesting that category-level differences mainly reflect target-side difficulty rather than special source–target compatibility. Thus, shared redundancy persists across task categories.

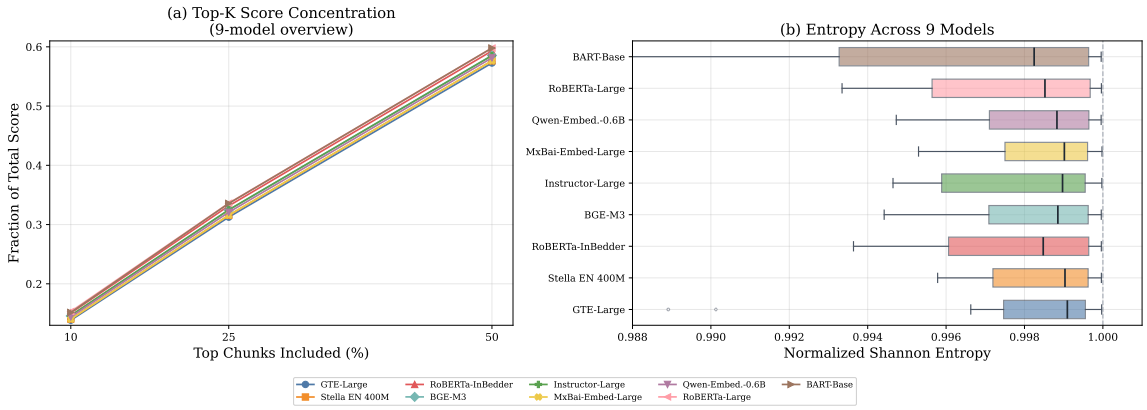

Figure 5: (a) Top-k concentration curves for 9 models. (b) Boxplots of normalized Shannon entropy for the same 9 models. Although all models have high entropy, RoBERTa-InBedder and General-purpose backbones show relatively higher top-k concentration and lower entropy, indicating stronger concentration.

Table 4: Held-out validation of optimized selection. Chunks are ranked on 70% of each task and evaluated on the remaining 30%. **Holdout Gap** is optimized–random gap on held-out data; **Same-Data Gap** is the drop from same-data to held-out evaluation.

Model	Holdout Gap	Win%	Same-Data Gap
Retrieval-optimized Embedders			
GTE-Large	1.74%	74.5%	0.67%
BGE-M3	2.70%	79.9%	1.91%
MxBai-Embed-Large	1.69%	75.0%	0.90%
Instruction-conditioned Embedders			
RoBERTa-InBedder	4.02%	81.3%	4.17%
Qwen3-Embedding-0.6B	2.42%	77.3%	2.88%
Stella EN 400M	2.21%	80.5%	1.01%
Instructor-Large	1.52%	71.4%	1.37%
General-purpose Language Model Backbones			
RoBERTa-Large	5.15%	87.2%	6.60%
BART-Base	4.26%	81.8%	5.82%

4.3 Auxiliary Robustness Checks

Held-out validation controls for same-data ranking bias. A potential concern is that optimized selection may overstate dimension concentration because chunks are ranked and evaluated on the same data. We therefore conduct a train/test split validation, ranking chunks on a 70% split and evaluating them on the held-out 30%. As shown in Table 4, the same training-paradigm pattern persists: Retrieval-optimized Embedders and most Instruction-conditioned Embedders continue to show smaller holdout gaps (+1.52%–2.70%), whereas RoBERTa-InBedder and General-purpose Language Model Backbones show larger ones (+4.02%–5.15%). The held-out advantage is also consistent across tasks: optimized selection beats random in 71.4–87.2% of held-out comparisons. These results support our diagnostic interpretation: the paradigm-dependent concentration pattern persists beyond the data used for ranking.

Additional robustness checks. We include several additional checks in the appendices. Appendix E varies the chunk size used for scoring dimensions, testing whether the diagnosed concentration pattern is sensitive to chunk granularity. Appendices F and G quantify random-seed variability and task-level lower-tail outcomes, showing that moderate-compression conclusions are not driven by unstable seeds or hidden task-level failures. Appendix I compares practical selection methods such as sequential, magnitude, gradient saliency, activation variance, learned masks, as well as projection-based compression baselines including PCA and random projection. These experiments verify the robustness of our main conclusions.

5 Mechanism

To further understand why the optimized–random gap differs across model families, we analyze the distribution of standalone chunk importance through normalized Shannon entropy. For each model-task pair, the standalone importance of chunk i is its task-specific evaluation score $q_i^{M,T}$, and $p_i^{M,T} = q_i^{M,T} / \sum_j q_j^{M,T}$ is the normalized importance. We then compute:

$$H_{\text{norm}} = -\frac{1}{\log_2 N} \sum_{i=1}^{N} p_i^{M,T} \log_2 p_i^{M,T} \qquad (10)$$

where N is the total number of chunks. Higher entropy indicates that importance is spread more evenly across chunks. Appendix H repeats the analysis with alternative importance definitions beyond standalone chunk scoring. Appendix L provides a

basis-dependence analysis, showing the same pattern appears under random orthogonal rotations, PCA, and whitening.

Top-k concentration and entropy track model-family gaps. Retrieval-optimized embedders and most instruction-conditioned embedders exhibit flatter top-k concentration curves and higher entropy, indicating that task-relevant information is distributed broadly. This pattern is consistent with their smaller optimized–random gaps, since random subsets are likely to retain enough useful signal when importance is near-uniform. By contrast, RoBERTa-InBedder and General-purpose Language Model Backbones consistently show relatively higher top-k concentration curves and lower entropy, suggesting a more concentrated importance distribution and explaining why optimized selection provides larger gains for these models.

High-entropy rankings support transferability. The same mechanism also helps explain why cross-task transfer remains consistently above random despite low rank agreement. When many chunks have similar standalone importance, task-specific differences can substantially reorder the chunk rankings, producing low Spearman correlations across tasks. However, source-task rankings can still give higher priority to chunks that carry information useful across tasks. This produces a small but stable advantage over random selection.

6 Discussion

Dimension selection reveals structured information distribution. Dimension selection provides a unified diagnostic for how task-relevant information is distributed within tasks and shared across tasks. Within a task, the optimized–random gap measures the degree of information concentration: smaller gaps indicate broadly distributed information, whereas larger gaps indicate stronger reliance on top-ranked chunks. Across tasks, weak rank agreement indicates that exact chunk-importance rankings are task-specific, while consistently above-random transfer gaps show shared redundancy in the embedding space. Rather than sharing a universal chunk ordering, tasks appear to share a coarser redundant structure: many chunks carry transferable task-relevant signal, while some chunks are repeatedly assigned low standalone importance. Together, these diagnostics give redundancy a more precise interpretation.

Training paradigms as a source of information distribution differences. Prior work shows that pretrained contextual representations and BERT sentence embeddings often exhibit anisotropic, non-uniform geometry, and that fine-tuning can further reshape embedding-space structure (Ethayarajh, 2019; Li et al., 2020; Rajaee and Pilehvar, 2021). This helps explain why General-purpose Language Model Backbones, not trained specifically for similarity or retrieval, show larger optimized–random gaps: task-relevant information may follow a more uneven geometry, making top-ranked dimensions more informative. By contrast, Retrieval-optimized Embedders are trained directly for similarity-based representation learning, often with contrastive learning that affects alignment and uniformity (Wang and Isola, 2020; Gao et al., 2021). This is consistent with their smaller optimized–random gaps, suggesting broader distribution of task-relevant information. RoBERTa-InBedder is a useful boundary case, with its stronger dimension structure potentially reflecting its short-answer-conditioned fine-tuning (Appendix K).

7 Conclusion

Prior work has shown that text embeddings are highly compressible, but compressibility alone does not reveal the representational structure behind this robustness. We use task-aware dimension selection to diagnose whether task-relevant information is broadly distributed or concentrated, how this pattern varies across training paradigms, and whether cross-task transfer reveals task-specific dimension importance or shared redundant structure.

Across 11 embedding models and 34 MTEB tasks, we find that Retrieval-optimized Embedders and most Instruction-conditioned Embedders show small optimized–random gaps, higher entropy, and flatter top-k concentration curves, indicating broadly distributed task-relevant information. RoBERTa-InBedder and General-purpose Language Model Backbones show stronger concentration, suggesting that information distribution is closely associated with training paradigm. Cross-task transfer further shows that chunk-importance rankings are task-specific, yet transferred selections remain above random, indicating shared redundancy beyond exact rank agreement. Overall, dimension selection reveals a training-paradigm-dependent information distribution. Code and results will be released upon publication.

Limitations

Our study provides a broad diagnostic analysis of task-relevant information distribution across 11 embedding models and 34 MTEB tasks. This coverage enables us to identify consistent patterns across model families, task categories, and compression levels, but it does not cover every possible deployment setting.

We focus on frozen, off-the-shelf text embedding models because this setting isolates how information is already organized in existing representations. Future work could extend the same diagnostic framework to models trained explicitly for adaptive dimensionality.

Our experiments study text embedding models. The same question may also be relevant for multimodal embeddings, where representation geometry may differ.

The observed association with training paradigm is empirical rather than theoretical. We do not provide a formal proof of this.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. 2024. Dimension importance estimation for dense information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. 2024. Answer is all you need: Instruction-following text embedding via answering the question. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. 2025. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yixuan Tang and Yi Yang. 2026. Gaprune: Gradient-alignment pruning for domain-aware embeddings. In *International Conference on Learning Representations (ICLR)*.

Hayato Tsukagoshi and Ryohei Sasano. 2025. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. 2025. Beyond matryoshka: Revisiting sparse coding for adaptive representation. In *International Conference on Machine Learning (ICML)*.

Zhanyu Wu, Richong Zhang, and Zhijie Nie. 2026. Learning to select: Query-aware adaptive dimension selection for dense retrieval. *arXiv preprint arXiv:2602.03306*.

Jinsung Yoon, Rajarishi Sinha, Sercan O Arik, and Tomas Pfister. 2024. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. 2025a. Smec: Rethinking matryoshka representation learning for retrieval embedding compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024a. Jasper and stella: Distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Gaifan Zhang, Yi Zhou, and Danushka Bollegala. 2024b. Evaluating unsupervised dimensionality reduction methods for pretrained sentence embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Evaluation Coverage by Model

Table 5 reports the output dimensionality of each model and summarizes which analyses are available for each of the 11 models in the study. For Stella EN 400M, which provides linear heads for multiple output sizes (512–8192), we use the default 1024-dimensional head before post-hoc dimension selection.

B Per-Task Results for GTE-Large

Table 6 reports per-task retention at 256 dimensions (corresponding to a 75% dimensionality reduction) for all 34 tasks in our benchmark on GTE-Large. The source average asks whether a task provides a generally useful dimension ranking for other tasks, while the target average asks how well a task can be served by rankings derived from other tasks. Figure 6 further shows the corresponding cross-task transfer matrix. For every source-target task pair, using the source task's importance ranking retains 93–100% of the target task's performance. These detailed results complement the main-text averages by showing how retention varies across individual tasks for each pruning method.

C Optimized–Random Gap Across the Number of Retained Dimensions

Table 7 reports the optimized–random gap across retained dimensionalities. The gap is positive throughout the sweep, confirming that task-aware rankings identify more informative chunks than random selection. The gap widens as fewer dimensions are retained, indicating that performance becomes more sensitive to chunk choice when the selected subset is smaller.

This effect differs by model family. Retrieval-optimized Embedders and most Instruction-conditioned Embedders show modest gaps at 256 retained dimensions, but the gaps become larger at 64, 32, and 16 dimensions. RoBERTa-InBedder and General-purpose Language Model Backbones show larger gaps throughout the sweep, indicating that their task-relevant information is less evenly distributed across dimensions. Figure 10(b) further shows that the normalized gap between optimized and worst-chunk selections shrinks quickly as more chunks are retained, meaning that once the retained subset is large enough, the exact chunk choice matters much less.

D Full Cross-Task Transfer Results Across Retained Dimensionalities

Table 8 extends the main body results (Table 3) to individual models. All values are mean retention across all source-target task pairs, including diagonal self-transfer.

E Chunk Size Sensitivity

We use chunk size $w = 2$ throughout the main experiments as a practical compromise between resolution and evaluation cost. For a typical 1024-dimensional model, evaluating single dimensions ($w = 1$) would require 1024 full task evaluations, while $w = 2$ cuts that to 512. Smaller chunks provide finer standalone scores but increase cost; larger chunks reduce cost but may hide within-chunk information. To verify that our conclusions are not sensitive to this choice, we sweep chunk sizes $w \in \{1, 2, 4, 8, 16, 32, 64\}$ on 5 models across 34 MTEB tasks.

Table 9 shows the full results. Three findings emerge: (1) Optimized retention is nearly identical at $w = 1$ and $w = 2$ for all 5 models, indicating that the main optimized–random gaps are not an artifact of using paired dimensions. (2) The optimized–random gap degrades gracefully as chunk size increases, and even $w = 64$ retains a positive gap for all models. (3) The association between the optimized–random gap and training paradigm persists across chunk sizes. The gaps remain small for Retrieval-optimized Embedders and Stella EN 400M (+1.4–4.8%), but are larger for RoBERTa-InBedder (+6.7–8.4%). Thus, varying the chunk size does not change the main empirical conclusion: the value of task-aware pruning remains associated with training paradigm. These results support our choice of $w = 2$ for the main experiments.

F Random Variance and Tail Risk

We quantify the full distribution of random selection performance using 100 seeds per combination across 9 models and 5 retained dimensionalities ($d \in \{16, 32, 64, 128, 256\}$).

Table 10 summarizes the representative results. At 256 retained dimensions, GTE-Large and Stella EN 400M are stable, with CV around 1.2–1.3% and P5 about 2 percentage points below the mean. Seed sensitivity increases at dim=16, but the lower tail remains moderate for these models. In contrast, General-purpose Language Model Backbones

Table 5: Model coverage by experiment. “Gap” = optimized–random gap comparison and anti-optimized experiment. “Transfer” = cross-task transfer. “Sweep” = pruning-ratio sweep. “Mechanism” = mechanism analysis on entropy. “Chunk” = chunk size sweep analysis. “Split” = train/test split experiment. “Tail” = random-variance and tail-risk analysis. “Magnitude” = magnitude pruning evaluation. “More Baselines” = Practical baselines including Gradient Saliency, Activation Variance, PCA, and Learned Mask. “More Definitions” = alternative importance definitions analysis. “Cost” = retrieval cost analysis.

Model	Dims	Gap	Transfer	Sweep	Mechanism	Chunk	Split	Tail	Magnitude	More Baselines	More Definitions	Cost
RoBERTa-Large	1024	✓	✓	✓	✓		✓	✓		✓	✓	✓
BART-Base	768	✓	✓	✓	✓		✓	✓	✓			
RoBERTa-InBedder	1024	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Qwen3-Embedding-0.6B	1024	✓	✓	✓	✓		✓	✓	✓			
Stella EN 400M	1024	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Instructor-Large	768	✓	✓	✓	✓		✓	✓				
GTR-T5-Large	768	✓	✓	✓		✓						
BGE-M3	1024	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
GTE-Base	768	✓	✓	✓		✓			✓			
MxBai-Embed-Large	1024	✓	✓	✓	✓		✓	✓	✓			
GTE-Large	1024	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

show much larger variance: at 256 retained dimensions, CV reaches 5.7–6.9%, and at dim=16 it rises to 15.1–18.5%. Thus, random selection has low tail risk for embedding-specific models at moderate compression, while tail risk becomes more severe for General-purpose Language Model Backbones and aggressive compression.

G Task-Level Low-Retention Outcomes

Retention ratio is useful for comparing tasks with different raw metrics, but ratio averaging may hide failures. Therefore, Table 11 reports low-retention summaries at 256 retained dimensions. P25, P10, and P5 describe progressively lower parts of the task distribution, while the worst-task column identifies where each model is most sensitive.

At 256 retained dimensions, 10 of 11 models have P10 retention above 90%, and 7 of 11 models have at most one task below 90%. This confirms that the mean retention is not driven only by a few extremely easy or difficult tasks: most of the task distribution remains close to full-dimensional performance.

The lower-retention cases are also structured rather than arbitrary. Across all 11 models, the task-model pairs below 90% come mostly from retrieval tasks (9 pairs) and classification tasks (7 pairs), with only two clustering pairs and none from other task categories. Worst-task entries show the same pattern: NFCorpus (retrieval) is the lowest-retention task for four models, SCIDOCS (retrieval) for two models, and TwentyNewsgroupsClustering (clustering) for two models. Thus, the low end of the distribution points to a small set of sensitive benchmark tasks rather than a collapse of selection strategy.

H Analysis of Alternative Importance Definitions

We compare the standalone chunk score used in the main analysis with three alternative importance definitions. As in the main text, $S^{M,T}(A)$ denotes the task score of model M on task T when retaining dimension set A, and C_i denotes the i-th chunk, Full denotes the full set of embedding dimensions. We define $\mathcal{A}_{-i}$ as the collection of retained dimension sets that are unions of chunks and do not contain C_i.

Leave-One-Out importance. Leave-One-Out (LOO) importance measures the performance drop caused by removing chunk C_i from the full representation:

$$\ell_i^{M,T} = S^{M,T}(\text{Full}) - S^{M,T}(\text{Full} \setminus C_i) \quad (11)$$

Thus, LOO evaluates whether C_i is necessary in the presence of all other chunks.

Marginal importance. Marginal importance measures the expected gain from adding chunk C_i to a retained dimension set that does not already contain it:

$$m_i^{M,T} = \mathbb{E}_{A \sim \mathcal{A}_{-i}} \left[S^{M,T}(A \cup C_i) - S^{M,T}(A) \right] \quad (12)$$

This captures whether C_i provides additional value when combined with other retained dimensions.

Shapley importance. Shapley importance averages the marginal contribution of C_i over all possible retained dimension sets that exclude it:

$$\phi_i^{M,T} = \sum_{A \in \mathcal{A}_{-i}} w(A) \left[S^{M,T}(A \cup C_i) - S^{M,T}(A) \right] \quad (13)$$

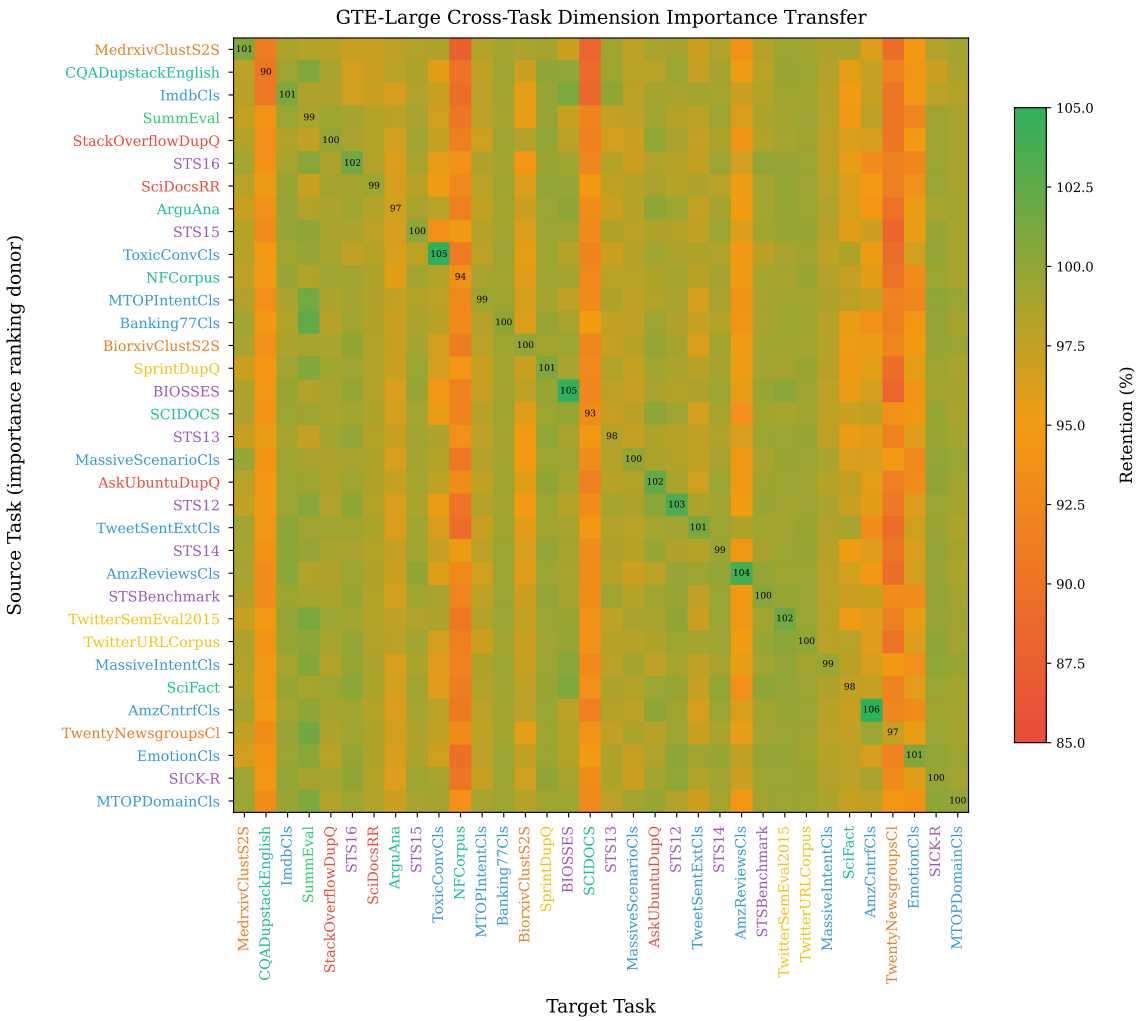

Figure 6: Cross-task dimension importance transfer matrix for GTE-Large at 256 retained dimensions. Each cell shows retention when using task i's ranking to prune for task j.

where $w(A) = |A|_c!(N-|A|_c-1)!/N!$ is the standard Shapley weight and $|A|_c$ denotes the number of chunks contained in A. This gives a context-averaged attribution of chunk importance and accounts for interactions among chunks. In practice, marginal and Shapley importance are estimated by sampling retained dimension sets from $\mathcal{A}_{-i}$.

Table 12 reports model-level Spearman correlation between standalone and alternative importance definitions and mean entropy per definition. Table 13 further shows per-task results on RoBERTa-Large. The weak correlation between standalone and Leave-One-Out/marginal/Shapley importance suggests that chunk importance is context-dependent. Standalone entropy is near-maximum (0.985–1.000), while Shapley entropy is substantially lower (0.655–0.933), revealing that interaction-aware importance detects much more structure.

Why standalone scoring? We use standalone chunk scores in the main text because our goal is to diagnose whether individual chunks can carry task-relevant information. In highly redundant embeddings, evaluating the effect of adding or removing one chunk from a large retained set often produces only minor changes in MTEB metrics, since many other chunks can compensate for the removed information. Such low-contrast marginal scores can make chunk rankings unstable. Standalone evaluation removes this compensation effect by asking how much task-relevant signal each chunk can support on its own, yielding a more discriminative diagnostic of isolated chunk informativeness.

I Practical Selection Baselines

The main experiments use optimized selection as a diagnostic, but it requires evaluating every chunk on each target task. We therefore ask whether

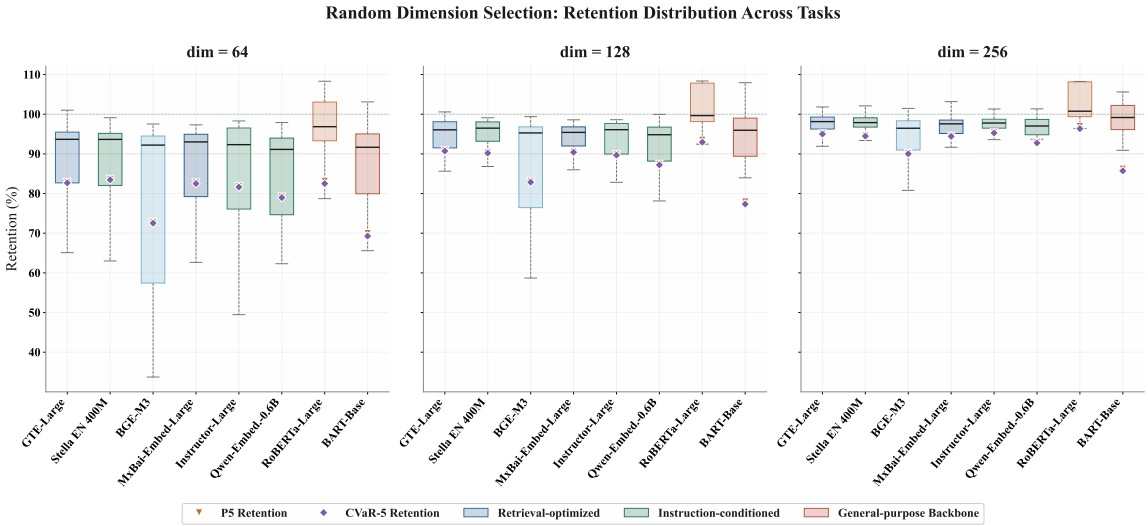

Figure 7: Retention distributions under random dimension selection across tasks. Each box plot shows the task-level spread of mean retention for retained dimensions $\in \{64, 128, 256\}$. Triangle markers ($\triangle$) indicate P5 retention; diamond markers ($\diamond$) indicate CVaR-5 retention. Tail risk increases sharply as the retained dimension decreases.

lower-cost selection rules can approximate the same task-relevant dimension structure. As in the main text, C_i denotes the i-th contiguous chunk of size w, and let $m = k/w$ be the number of retained chunks. For a baseline that assigns each chunk a score u_i, we retain

$$A_{\text{base}}(k) = \text{Top}_m \left(\{u_i\}_{i=1}^{N} \right) \tag{14}$$

where the selected chunks correspond to the retained dimensions.

Sequential selection. Sequential selection retains the first m chunks in the model's original dimension order:

$$A_{\text{seq}}(k) = \{C_1, \ldots, C_m\} \tag{15}$$

This tests whether the native dimension order already encodes a useful importance ranking, as in Matryoshka-style representations.

Magnitude-based selection. Magnitude-based selection scores each dimension by the ℓ_2 norm of its corresponding column in the embedding weight matrix. Let $\mathbf{W} \in \mathbb{R}^{V \times D}$ be the embedding matrix. The dimension score is:

$$u_d^{\text{mag}} = \|\mathbf{W}_{:,d}\|_2 \tag{16}$$

This baseline tests whether dimensions with larger embedding-weight magnitude are also more important for downstream task performance.

Gradient saliency. Gradient saliency uses local task-loss sensitivity to score dimensions. For task T, embedding $\mathbf{e}(x)$, and task loss $\mathcal{L}_T$, we define

$$u_d^{\text{grad}} = \mathbb{E}_x \left[\left| e_d(x) \frac{\partial \mathcal{L}_T}{\partial e_d(x)} \right| \right] \tag{17}$$

The corresponding chunk score is

$$u_i^{\text{grad}} = \sum_{d \in C_i} u_d^{\text{grad}} \tag{18}$$

This method tests whether dimensions with larger local influence on the task loss also preserve more task-relevant information after selection.

Activation variance. Activation variance scores dimensions by how much their embedding values vary across input examples:

$$u_d^{\text{var}} = \text{Var}_x \left[e_d(x) \right] \tag{19}$$

The corresponding chunk score is

$$u_i^{\text{var}} = \sum_{d \in C_i} u_d^{\text{var}} \tag{20}$$

The intuition is that dimensions with larger variation across examples may carry more discriminative information.

Learned mask selection. The learned-mask baseline directly optimizes a continuous mask over embedding dimensions. Let $\mathbf{m} \in [0, 1]^D$ be applied to the embedding:

$$\tilde{\mathbf{e}}(x) = \mathbf{m} \odot \mathbf{e}(x) \tag{21}$$

We optimize the mask with a task loss and an ℓ_1 sparsity penalty:

$$\min_{\mathbf{m}\in[0,1]^D} \mathbb{E}_{(x,y)} \left[\mathcal{L}_T(\mathbf{m} \odot \mathbf{e}(x), y)\right] + \lambda\|\mathbf{m}\|_1 \tag{22}$$

After optimization, dimensions are ranked by their mask values.

Projection-based compression. PCA and random projection are included as projection-based compression baselines, although they do not retain a subset of original dimensions. For PCA, given centered embeddings $\mathbf{e} - \boldsymbol{\mu}$ and the top k principal directions $\mathbf{U}_k \in \mathbb{R}^{D\times k}$, we use

$$\mathbf{z}^{\mathrm{PCA}} = \mathbf{U}_k^\top(\mathbf{e} - \boldsymbol{\mu}) \tag{23}$$

For random projection, with projection matrix $\mathbf{G} \in \mathbb{R}^{D\times k}$, we use

$$\mathbf{z}^{\mathrm{RP}} = \mathbf{G}^\top \mathbf{e} \tag{24}$$

We evaluate these compressed representations using the same retention ratio as in the main text, replacing the retained-dimension representation with the projected representation.

Results. Table 15 reports the representative-task results for gradient saliency, activation variance, and learned-mask selection, while Table 16 compares PCA, random projection, and random selection. Sequential and magnitude-based results are reported in Figure 8 and 9.

Across these baselines, simple task-agnostic signals (prefix order, weight magnitude, activation variance, and random projections) and lightweight task-aware signals (gradient saliency and learned masks) do not consistently recover a substantially better retained subset than random selection at moderate compression. This supports the main diagnostic conclusion: for many embedding-specific models, task-relevant information is distributed broadly enough that many retained subsets perform similarly.

J Retrieval System Cost

Using FAISS FlatIP, Random-256 retains 87.5–98.4% of full-dim nDCG while reducing raw vector storage by 4× (Table 17). Random projection is included as a projection-based comparator, but it requires applying a projection matrix. For GTE-Large, Random-256 reduces P50 latency from 0.16–0.23ms to 0.04–0.08ms, a 2.9–4.3× speedup.

K Short-Answer-Conditioned Supervision Is Associated with Greater Dimension Structure

Characteristics of the exception. To understand why RoBERTa-InBedder differs, we compare its dimension importance distribution with one Retrieval-Optimized Embedder (GTE-Large) and one Instruction-conditioned Embedder (Stella EN 400M) (Table 14, Figure 10). Its Gini coefficient is 0.124 vs. 0.101–0.107 for GTE-Large and Stella EN 400M, its chunk-importance coefficient of variation is 0.242 vs. 0.186–0.195. Its normalized entropy is 0.990 vs. 0.993–0.994. While these absolute differences are small, the relative differences are consistent: RoBERTa-InBedder's importance distribution is measurably less uniform. Table 14 also demonstrates that General-purpose Language Model Backbones show even lower entropy and higher Gini, confirming that training paradigm influences dimension structure.

Mechanism: short-answer-conditioned supervision reduces uniformity. RoBERTa-InBedder shows consistently higher concentration of importance across all metrics than Retrieval-optimized Embedders and other Instruction-conditioned Embedders. A plausible interpretation is that its supervision differs from standard embedding training in both objective and signal breadth. InBedder instantiates instruction-following embedding by treating the instruction as a question and learning representations through expected answers; in practice, it is fine-tuned on abstractive QA data, with answers simplified to an average length of only 2.89 tokens (Peng et al., 2024). Such short-answer-conditioned supervision may provide a narrower training signal than contrastive or multi-task embedding objectives, causing the model to align a smaller subset of dimensions with latent answer features while leaving other dimensions less consistently useful. This interpretation is consistent with prior evidence that fine-tuning can substantially reshape embedding geometry rather than simply improving isotropy (Rajaee and Pilehvar, 2021).

L Theoretical perspective: near-uniform importance as a training-induced pattern

Under a contrastive loss $\mathcal{L} = -\log[\exp(\mathbf{z}_i \cdot \mathbf{z}_j)/\sum_k \exp(\mathbf{z}_i \cdot \mathbf{z}_k)]$ where $\mathbf{z}_i$ is the representation of anchor i, $\mathbf{z}_j$ is the positive, and $\mathbf{z}_k$ ranges

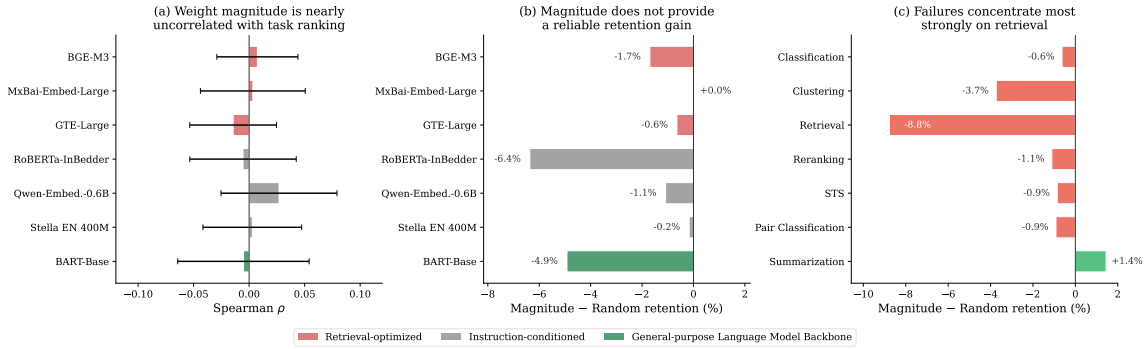

Figure 8: (a) Spearman rank correlation between magnitude-based dimension ranking and task-specific importance ranking across five models. All mean correlations remain near zero. (b–c) Per-task magnitude vs. random retention. Points below the diagonal indicate tasks where random outperforms magnitude.

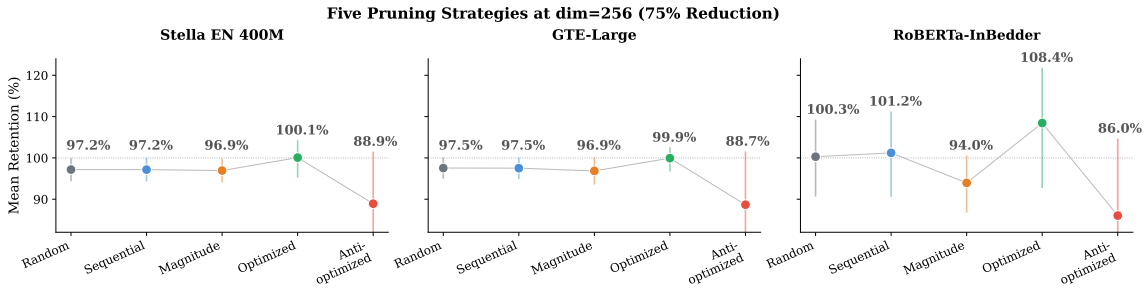

Figure 9: Comparison of five pruning methods at 256 retained dimensions for three representative models. Error bars show standard deviations across tasks. Random, Sequential, and Magnitude cluster together, while Anti-optimized drops to 86–89%. The optimized selection strategy lies only slightly above the cluster for GTE-Large and Stella EN 400M, but noticeably farther above it for RoBERTa-InBedder. Magnitude slightly underperforms Random on GTE-Large and Stella EN 400M, but shows a much larger drop on RoBERTa-InBedder.

over all positives and negatives in the batch, each dimension d contributes additively to the dot product. When the training data is broad and diverse, no single dimension can specialize too strongly without hurting performance elsewhere, so the contributions tend to balance out. The RoBERTa-InBedder exception fits this picture in that its non-contrastive, answer-conditioned fine-tuning on very short QA answers is associated with a less even importance distribution and a larger task-aware advantage. Taken together, this reasoning suggests a clear prediction for future work: models trained with less diverse supervision should display less uniform importance distributions and, as a result, larger optimized–random gaps. A rigorous test of this prediction would require controlling for confounding factors by holding architecture and data fixed while varying the training objective, or by comparing matched model pairs that share the same embedding head and normalization. We leave this controlled investigation to future work.

We also repeat the pruning analysis after several basis transformations, including random orthogonal rotations, PCA, and whitening. The sequential–random gap stays below 1% in every basis (Figure 11). This suggests that the observed robustness of random selection is not merely a quirk of the native coordinate basis.

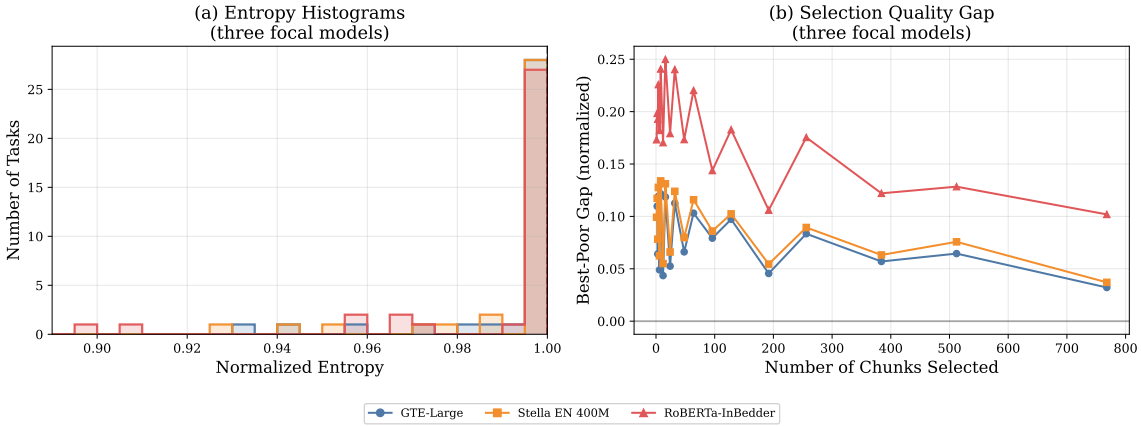

Figure 10: Representative views of the redundancy mechanism for the three representative models. (a) Histograms of per-task normalized entropy. (b) Normalized optimized–worst selection gap as the number of selected chunks increases.

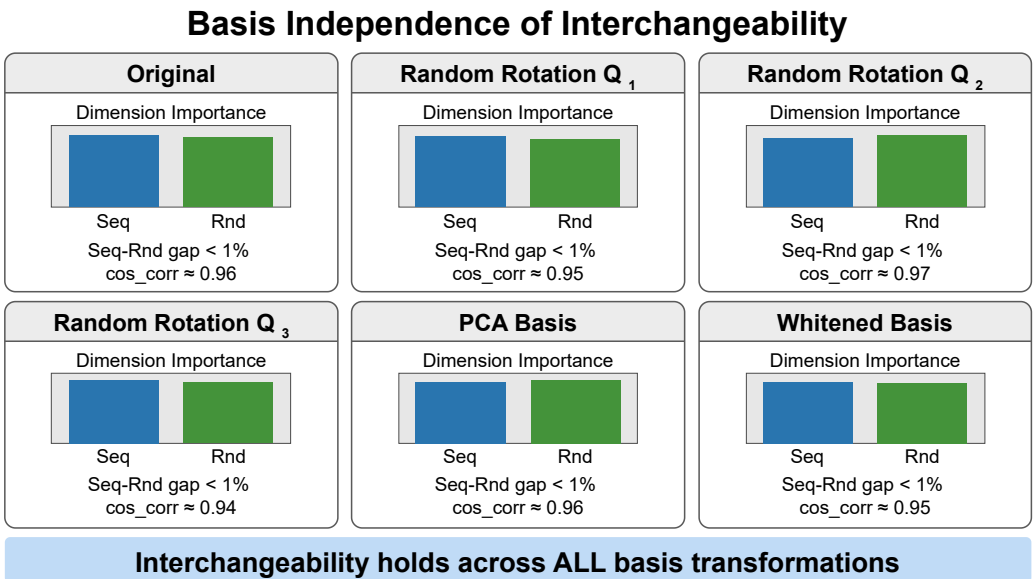

Figure 11: Basis-dependence analysis of random selection. We apply six basis transformations to the embedding space (original, three random orthogonal rotations, PCA, and whitening) and measure the sequential–random gap and cosine–correlation preservation after pruning. In every basis, the gap stays below 1% and cosine correlation exceeds 0.94, suggesting that the broad distribution of task-relevant information is not merely an artifact of one coordinate system.

Table 6: Per-task retention (%) at 256 retained dimensions for GTE-Large. Optimized and Anti-optimized use chunk size $w = 2$. Source average is the mean retention when the row task supplies the dimension ranking and all 34 tasks are targets. Target average is the mean retention when all 34 tasks supply rankings and the row task is the target.

Task	Random	Optimized	Anti-optimized	Source average	Target average
Classification					
AmazonCounterfactualClassification	95.8	105.6	70.0	97.9	96.5
AmazonReviewsClassification	95.0	103.9	69.4	97.8	95.5
Banking77Classification	99.3	100.0	98.1	97.5	99.3
EmotionClassification	94.4	100.9	82.0	97.9	94.8
ImdbClassification	99.3	100.9	77.5	97.1	99.4
MTOPDomainClassification	99.0	99.9	95.5	98.0	99.0
MTOPIntentClassification	97.8	99.2	94.7	97.5	97.9
MassiveIntentClassification	98.0	99.5	96.7	97.8	98.1
MassiveScenarioClassification	97.4	99.8	91.2	97.6	98.0
ToxicConversationsClassification	96.9	105.0	66.6	97.5	97.3
TweetSentimentExtractionClassification	97.9	101.2	84.8	97.7	97.8
Clustering					
BiorxivClusteringS2S	96.1	99.9	85.0	97.5	96.6
MedrxivClusteringS2S	98.5	101.0	95.6	97.0	98.3
TwentyNewsgroupsClustering	90.8	97.3	63.9	97.9	91.4
Retrieval					
ArguAna	96.7	96.7	96.7	97.4	96.8
CQADupstackEnglishRetrieval	93.8	90.5	94.4	97.1	93.8
NFCorpus	91.4	93.9	89.3	97.5	91.7
SCIDOCS	92.9	93.0	89.7	97.6	92.6
SciFact	96.9	97.6	96.3	97.9	96.7
Reranking					
AskUbuntuDupQuestions	99.2	102.2	96.6	97.6	99.1
SciDocsRR	98.3	99.4	94.7	97.4	98.3
StackOverflowDupQuestions	98.8	99.6	96.7	97.2	98.8
STS					
BIOSSES	98.8	105.1	94.0	97.6	99.2
SICK-R	99.5	99.8	97.1	97.9	99.5
STS12	99.6	103.4	96.9	97.7	99.1
STS13	98.3	98.3	94.4	97.6	98.1
STS14	99.1	99.5	97.6	97.7	99.1
STS15	99.1	100.2	92.5	97.5	98.9
STS16	99.1	101.7	94.0	97.4	99.2
STSBenchmark	99.2	100.4	97.6	97.8	99.1
Pair Classification					
SprintDuplicateQuestions	99.5	100.9	97.0	97.5	99.6
TwitterSemEval2015	99.3	101.7	96.4	97.8	99.5
TwitterURLCorpus	99.4	100.0	98.3	97.8	99.4
Summarization					
SummEval	99.8	98.7	98.8	97.2	99.5
Mean	97.5	99.9	90.6	97.6	97.6

Table 7: Optimized–random gap across retained dimensions with chunk size $w = 2$.

Model	Retained dimensions					
	16	32	64	128	256	512
Retrieval-optimized Embedders						
GTR-T5-Large	+10.09%	+9.69%	+8.45%	+6.88%	+4.69%	+2.35%
BGE-M3	+11.06%	+9.68%	+8.10%	+6.11%	+4.61%	+3.22%
GTE-Base	+9.79%	+8.44%	+6.46%	+4.73%	+3.49%	+1.82%
MxBai-Embed-Large	+9.66%	+7.12%	+5.11%	+3.59%	+2.59%	+1.57%
GTE-Large	+9.12%	+7.31%	+5.09%	+3.31%	+2.41%	+1.72%
Instruction-conditioned Embedders						
RoBERTa-InBedder	+15.63%	+14.33%	+12.65%	+10.92%	+8.19%	+6.18%
Qwen3-Embedding-0.6B	+13.90%	+11.99%	+9.44%	+6.99%	+5.30%	+3.10%
Stella EN 400M	+9.82%	+8.15%	+6.09%	+4.68%	+3.22%	+2.03%
Instructor-Large	+9.20%	+7.97%	+6.05%	+4.30%	+2.89%	+1.44%
General-purpose Language Model Backbones						
RoBERTa-Large	+20.33%	+19.09%	+18.45%	+15.01%	+11.75%	+11.73%
BART-Base	+19.22%	+16.81%	+14.90%	+13.87%	+10.08%	+3.73%

Table 8: Cross-task transfer across six different retained dimensions for all 11 models with chunk size $w = 2$. Each cell is mean retention over all source-target pairs, including diagonal self-transfer.

Model	Transfer@16	Transfer@32	Transfer@64	Transfer@128	Transfer@256	Transfer@512
RoBERTa-Large	67.0%	79.5%	93.5%	105.5%	114.0%	117.8%
RoBERTa-InBedder	64.0%	75.9%	87.2%	96.2%	102.0%	104.8%
BART-Base	64.4%	75.5%	85.5%	94.2%	100.1%	102.1%
GTE-Large	67.8%	79.6%	88.8%	94.5%	97.6%	99.2%
Stella EN 400M	67.5%	79.2%	88.5%	94.3%	97.5%	99.2%
MxBai-Embed-Large	67.2%	78.8%	88.0%	93.9%	97.3%	99.1%
Instructor-Large	64.4%	76.0%	86.1%	93.0%	97.0%	99.2%
Qwen3-Embedding-0.6B	64.9%	76.6%	86.2%	92.9%	96.8%	99.1%
GTE-Base	63.1%	75.0%	85.5%	92.7%	96.8%	99.1%
GTR-T5-Large	62.2%	72.9%	83.2%	91.0%	96.1%	98.9%
BGE-M3	60.3%	70.7%	81.0%	89.2%	94.9%	98.3%

Table 9: Optimized retention (%) at dim=256 across chunk sizes averaged over 34 MTEB tasks. Random and Sequential baselines shown for reference. Bold marks the best chunk size per model.

Model	Optimized retention at different chunk sizes w							Random	Seq
	w=1	w=2	w=4	w=8	w=16	w=32	w=64		
Retrieval-optimized Embedders									
GTE-Large	99.9	**99.9**	99.6	99.4	99.4	99.2	98.9	97.5	97.5
GTE-Base	99.9	**100.1**	99.8	99.1	98.8	98.3	98.2	96.6	96.6
GTR-T5-Large	**100.5**	100.4	100.3	99.6	99.0	98.3	98.1	95.7	95.4
Instruction-conditioned Embedders									
RoBERTa-InBedder	**108.7**	108.5	107.4	107.0	108.1	108.0	107.4	100.3	101.3
Stella EN 400M	100.1	**100.3**	100.2	100.2	99.8	99.2	99.0	97.1	97.1

Table 10: Random selection distribution statistics across 100 seeds (mean over 20 tasks per model). CV = coefficient of variation. P5 = 5th percentile. CVaR-5 = conditional value at risk at 5%.

Model	Dim	Mean $\pm$ Std	CV	P5	CVaR-5
Retrieval-optimized Embedders					
GTE-Large	256	$(97.2 \pm 1.3)\%$	1.3%	95.1%	95.0%
GTE-Large	64	$(86.9 \pm 2.1)\%$	2.6%	82.9%	82.7%
GTE-Large	16	$(63.9 \pm 3.1)\%$	6.8%	58.7%	58.4%
Instruction-conditioned Embedders					
Stella EN 400M	256	$(96.6 \pm 1.2)\%$	1.2%	94.6%	94.5%
Stella EN 400M	64	$(87.3 \pm 2.3)\%$	2.8%	83.7%	83.5%
Stella EN 400M	16	$(63.7 \pm 3.4)\%$	7.9%	58.2%	58.0%
General-purpose Language Model Backbones					
RoBERTa-Large	256	$(114.3 \pm 8.0)\%$	5.7%	97.0%	96.4%
RoBERTa-Large	64	$(99.4 \pm 8.7)\%$	8.4%	83.2%	82.5%
RoBERTa-Large	16	$(76.2 \pm 7.7)\%$	15.1%	62.4%	61.7%
BART-Base	256	$(97.8 \pm 6.6)\%$	6.9%	86.2%	85.7%
BART-Base	64	$(82.9 \pm 8.0)\%$	11.5%	70.0%	69.3%
BART-Base	16	$(62.1 \pm 7.0)\%$	18.5%	50.0%	49.4%

Table 11: Random low-retention summaries at 256 retained dimensions. P25, P10, and P5 are task-level percentiles. Worst task reports the lowest-retention task and its retention.

Model	Mean	P25	P10	P5	Worst task	< 90	> 100
Retrieval-optimized Embedders							
GTR-T5-Large	95.7%	92.6%	90.3%	89.3%	SCIDOCS (88.4%)	3	1
BGE-M3	94.6%	91.9%	86.5%	84.0%	SCIDOCS (80.2%)	7	0
GTE-Base	96.6%	94.4%	93.1%	92.5%	MTOPIntentClassification (91.2%)	0	0
MxBai-Embed-Large	97.2%	96.1%	93.4%	92.7%	NFCorpus (91.1%)	0	0
GTE-Large	97.5%	96.8%	94.0%	92.4%	TwentyNewsgroupsClustering (90.8%)	0	0
Instruction-conditioned Embedders							
RoBERTa-InBedder	100.3%	96.3%	94.2%	93.3%	TwentyNewsgroupsClustering (91.4%)	0	11
Qwen3-Embedding-0.6B	95.5%	92.6%	90.6%	88.8%	NFCorpus (83.6%)	3	0
Stella EN 400M	97.1%	95.7%	93.0%	92.2%	NFCorpus (89.8%)	1	1
Instructor-Large	97.0%	95.8%	93.6%	91.9%	NFCorpus (89.0%)	1	1
General-purpose Language Model Backbones							
RoBERTa-Large	111.5%	96.1%	90.8%	89.9%	Banking77Classification (87.2%)	2	18
BART-Base	97.9%	95.1%	93.0%	91.7%	CQADupstackEnglishRetrieval (86.8%)	1	6

Table 12: Spearman correlation between standalone and alternative importance definitions at 256 retained dimensions (chunk size $w = 4$), and mean entropy per definition.

	Correlation with Standalone			Entropy			
Model	LOO	Marginal	Shapley	Standalone	LOO	Marginal	Shapley
GTE-Large	0.094	0.155	0.132	0.999	0.936	0.939	0.625
BGE-M3	0.175	0.251	0.181	0.998	0.932	0.939	0.597
Stella EN 400M	0.124	0.209	0.157	0.999	0.918	0.944	0.610
RoBERTa-Large	0.091	0.296	0.235	0.997	0.838	0.891	0.660
RoBERTa-InBedder	0.096	0.248	0.179	0.997	0.888	0.927	0.654

Table 13: Per-task importance definition comparison for RoBERTa-Large at 256 retained dimensions (chunk size $w = 4$). Spearman correlations and normalized entropy for standalone, Leave-one-out, marginal, and Shapley importance. Low standalone-to-alternative correlations confirm importance is context-dependent.

	Spearman ρ with Standalone				Normalized Entropy			
Task	LOO	Marginal	Shapley	p (best)	Standalone	LOO	Marginal	Shapley
ImdbClassification	-0.080	0.252	0.176	<0.001	1.000	0.943	0.942	0.707
Banking77Classification	-0.001	0.078	0.156	0.012	0.996	0.949	0.914	0.860
TwentyNewsgroupsClustering	-0.038	0.137	0.132	0.029	0.998	0.922	0.950	0.932
MedrxivClusteringS2S	0.018	0.039	0.073	0.247	0.999	0.952	0.945	0.842
STSBenchmark	0.387	0.704	0.412	<0.001	0.998	0.343	0.677	0.655
BIOSSES	0.260	0.568	0.459	<0.001	0.985	0.921	0.920	0.783

Table 14: Chunk importance distribution across five representative models. Optimized–random gap is measured at 256 retained dimensions, with chunk size $w = 2$.

Metric	GTE-Large	Stella EN 400M	RoBERTa-InBedder	BART-Base	RoBERTa-Large
Normalized entropy	0.994	0.993	0.990	0.989	0.989
Gini coefficient	0.101	0.107	0.124	0.130	0.131
Chunk importance CV	0.186	0.195	0.242	0.252	0.255
Top-50% concentration	0.573	0.576	0.593	0.598	0.598
Optimized–random gap	+2.41%	+3.22%	+8.19%	+10.08%	+11.75%

Table 15: Mean retention of gradient saliency, activation variance, and learned-mask selection on 8 representative tasks at 256 dimensions. Gradient and activation variance can be actively harmful for RoBERTa-InBedder and RoBERTa-Large.

Model	Random	Gradient	Activation Variance	Mask
Retrieval-optimized Embedders				
GTE-Large	99.5%	99.5%	99.7%	98.9%
BGE-M3	96.8%	97.1%	97.6%	96.4%
Instruction-conditioned Embedders				
Stella EN 400M	98.8%	99.3%	98.4%	98.3%
RoBERTa-InBedder	100.1%	94.8%	93.3%	98.5%
General-purpose Language Model Backbones				
RoBERTa-Large	102.9%	93.4%	92.4%	102.0%

Table 16: Random selection vs. projection-based compression, comparing random selection (RS), PCA, and random projection (RP) at 64 and 256 dimensions across representative models.

Model	Dim	PCA	RP	RS
GTE-Large	64	90.2%	92.9%	93.1%
	256	98.0%	98.7%	99.0%
BGE-M3	64	85.9%	87.2%	87.8%
	256	95.8%	96.4%	96.1%
Stella EN 400M	64	92.0%	93.9%	94.4%
	256	98.5%	98.6%	98.9%
RoBERTa-InBedder	64	80.8%	90.4%	96.7%
	256	95.6%	98.4%	98.0%
RoBERTa-Large	64	80.6%	91.2%	98.4%
	256	99.9%	99.6%	100.2%

Table 17: Retrieval cost–quality tradeoff at dim=256 for GTE-Large using FAISS FlatIP. nDCG@10 and Recall report retention relative to the full-dimensional baseline for each task.

Task	Method	nDCG@10	Mem (MB)	Latency (ms)	Recall
NFCorpus	Full-1024	100.0%	14.2	0.16	100.0%
	Random-256	98.4%	3.5	0.04	96.7%
	RP-256	85.3%	3.5	0.38	82.9%
ArguAna	Full-1024	100.0%	19.5	0.22	100.0%
	Random-256	93.7%	4.9	0.05	95.9%
	RP-256	94.0%	4.9	0.40	93.8%
SciFact	Full-1024	100.0%	19.5	0.23	100.0%
	Random-256	87.5%	4.9	0.08	93.8%
	RP-256	83.4%	4.9	0.80	88.1%