Do We Really Need All Those Dimensions? An Intrinsic Evaluation Framework for Compressed Embeddings

Nathan Inkiriwang[1,2], **Necva Bölücü**[1], **Garth Tarr**[2], **Maciej Rybinski**[3]
[1]CSIRO Data61, Sydney, Australia
[2]University of Sydney, Sydney, Australia
[3]ITIS, University of Málaga, Málaga, Spain

Abstract

High-dimensional text embeddings are foundational to modern NLP but costly to store and use. While embedding compression addresses these challenges, selecting the best compression method remains difficult. Existing evaluation methods for compressed embeddings are either expensive or too simplistic. We introduce a comprehensive intrinsic evaluation framework featuring a suite of task-agnostic metrics that together provide a reliable proxy for downstream performance. A key contribution is EOS_k, a novel spectral fidelity measure specifically designed to be robust to embedding anisotropy. Through extensive experiments on diverse embeddings across four downstream tasks, we demonstrate that our intrinsic metrics reliably predict extrinsic performance and reveal how different embedding architectures depend on distinct geometric properties. Our framework provides a practical, efficient, and interpretable alternative to standard evaluations for compressed embeddings[1].

1 Introduction

Word, sentence, and, more generally, text embeddings[2] have become central to Natural Language Processing (NLP), enabling a range of tasks from semantic search to classification and clustering (Muennighoff et al., 2023; Wang et al., 2024a; Chen et al., 2024; Wang et al., 2024b). As embedding models have evolved from static embeddings (e.g., GloVe (Pennington et al., 2014)) to contextualised ones (e.g., BERT (Devlin et al., 2019)) and more recently, large language model (LLM)-based (e.g., E5 (Wang et al., 2022)), the dimensionality and complexity of these embeddings have increased significantly. Although higher-dimensional embeddings often capture richer linguistic information, they incur substantial computational costs in terms of memory consumption, inference time, energy usage and carbon emissions (Strubell et al., 2019; Schwartz et al., 2020; Liu and Yin, 2024). Such high dimensionality also poses practical challenges, particularly in low-resource settings or efficiency-critical environments, where memory, computational cost, and latency are major constraints (Sanh et al., 2019; Turc et al., 2019).

To address these challenges, dimensionality reduction (DR) and quantisation have been increasingly adopted to compress embeddings (Raunak et al., 2019; Sherki et al., 2021; Liu et al., 2022; Rosa et al., 2022; Yamagiwa et al., 2023; Hwang et al., 2023; Xue et al., 2024; Bibi et al., 2024; Lang et al., 2024; Hina et al., 2024). This trend is motivated, in part, by the finding that many embedding models possess an inherently low intrinsic dimensionality (Kataiwa et al., 2025). This property indicates significant redundancy, which compression[3] methods can exploit to substantially reduce the computational burden while preserving, or even improving, downstream performance (Raunak et al., 2019; Zhang et al., 2024). However, despite growing adoption for compressing embeddings, significant gaps remain in both the theoretical understanding and systematic empirical evaluation of these methods.

The evaluation of embedding compression has largely relied on two *limited* practices: (i) using extrinsic downstream performance metrics (e.g., accuracy or retrieval scores) (Yamagiwa et al., 2023; Hwang et al., 2023; Xue et al., 2024; Bibi et al., 2024; Lang et al., 2024); and (ii) relying on a single intrinsic metric (May et al., 2019). Neither offers a complete or reliable picture of embedding quality.

Extrinsic evaluations are computationally de-

[1]The framework and EOS_k implementation are available at `https://github.com/nathaninkiriwang/TextEmbedCompress`.

[2]"Embedding" and "representation" are used interchangeably in the literature.

[3]Throughout, 'compression' covers both dimensionality reduction and quantisation.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13305–13323
November 4-9, 2025 ©2025 Association for Computational Linguistics

manding, given the large combinatorial space of models, tasks, and compression techniques, and are highly sensitive to dataset and configuration choices (e.g., classifier design, retrieval settings). More importantly, they provide limited insight. Performance scores do not show which structural properties are preserved or lost. This results in a fragmented and opaque understanding of compression, especially across diverse embedding types (Yamagiwa et al., 2023; Hwang et al., 2023; Xue et al., 2024; Bibi et al., 2024).

Intrinsic evaluations based on a single metric are similarly limited (May et al., 2019). Such approaches fail to generalise across embedding architectures and thus offer a limited view that restricts practical applicability, especially when compression methods behave inconsistently across tasks.

To address these limitations, we propose a comprehensive and scalable evaluation framework for compressed embeddings[4]. Our framework includes a set of theoretically grounded **intrinsic metrics** that are task-agnostic, and, crucially, provide a consistently *robust* proxy for overall downstream utility. These metrics are motivated by key goals in embedding design, preserving local neighbourhood structure (May et al., 2019; Wang and Isola, 2020), retaining global topology (Ethayarajh, 2019) and maintaining information fidelity (Abdi and Williams, 2010; Mu and Viswanath, 2018), and aim to capture distinct geometric and statistical properties that affect downstream performance. We also introduce EOS_k, a novel spectral fidelity metric designed to better measure semantic preservation. Unlike traditional metrics that focus on the entire eigenspectrum, EOS_k specifically analyses the residual eigenspace after removing the top$-k$ principal components. These top components often capture broad, anisotropic variance that can overshadow more subtle, task-relevant information.

We apply our framework to three widely used open-source embeddings, GloVe (Pennington et al., 2014) (static), BERT (Devlin et al., 2019) (contextual), and E5 (Wang et al., 2022) (contrastive), which vary in architectures, training objectives, and anisotropy levels. Through extensive correlation analysis on four downstream tasks across 21 datasets from the MTEB benchmark (Muennighoff et al., 2023), we find consistent patterns linking intrinsic properties with downstream performance: contextual embeddings benefit most from local structures, while static and contrastive embeddings align better with global and spectral fidelity.

Our framework provides a practical guide for selecting compression methods based on embedding type and downstream tasks. By measuring key intrinsic metrics: local, global, and spectral structure, practitioners can determine which properties are most important for their task. This enables more informed decisions when selecting compression methods, allowing them to balance compression ratios with the preservation of structurally important features for their applications.

Using our evaluation framework, we identify Random Projection and `int8` quantisation as consistently effective compression strategies. This benchmarked approach will allow users to compress embeddings effectively while maintaining task-relevant performance and avoiding exhaustive benchmarking. In addition, our novel EOS_k metric outperforms standard spectral metrics in scenarios with anisotropic embeddings, enabling more reliable intrinsic evaluation of structure-preserving quality under varying model architectures.

2 Compression Methods

Preliminaries and Notation Let $\mathbf{X} \in \mathbb{R}^{n \times D}$ denote the original embedding matrix, where n is the number of samples (e.g., words, sentences or documents) and D is the original embedding dimension. Each row $\mathbf{x}_i \in \mathbb{R}^D$ is an individual embedding vector. The objective of embedding compression is to transform $\mathbf{X}$ into a representation that requires less storage and/or computational resources, while preserving its utility. We focus on two classes of operations: Dimensionality Reduction (DR) and Quantisation (Q).

Dimensionality Reduction (DR): A function f_{DR} maps $\mathbf{X}$ to a lower-dimensional space $\mathbb{R}^{n \times d}$, where $d < D$:

$$\mathbf{X}_{DR} = f_{DR}(\mathbf{X})$$

Quantisation (Q): Given a real-valued matrix $\mathbf{M} \in \mathbb{R}^{n \times k}$ (e.g., $\mathbf{X}$ or $\mathbf{X}_{DR}$), quantisation maps it to B-bit integers using scale S and zero-point ZP:

$$f_Q : \mathbb{R}^{n \times k} \to (\mathbb{I}_B^{n \times k}, \mathcal{P}_Q)$$

where k is the dimension of the input matrix (either D or d), and $\mathcal{P}_Q = \{S, ZP\}$ represents the set of quantisation parameters. The quantised matrix is:

$$(\mathbf{M}_Q, \mathcal{P}_Q) = f_Q(\mathbf{M}).$$

[4] Related work is given in Appendix A.

Here, $\mathbf{M}_Q \in \mathbb{I}_B^{n \times k}$. For this study, we focus on $B = 8$ (i.e., 'int8' quantisation).

DR followed by Quantisation (DR+Q): First, DR is applied to $\mathbf{X}$ to obtain $\mathbf{X}_{DR}$. Then, $\mathbf{X}_{DR}$ is quantised:

$$((\mathbf{X}_{DR})_Q, \mathcal{P}_Q) = f_Q(\mathbf{X}_{DR}) = f_Q(f_{DR}(\mathbf{X})).$$

Appendix B provides detailed explanations of the compression methods.

3 Evaluation Framework

To comprehensively evaluate the effectiveness of compression methods, we propose a unified evaluation framework that captures both the **structural and spectral fidelity** of compressed representations. Given original embeddings $\mathbf{X} \in \mathbb{R}^{n \times D}$ and their compressed representations $\mathbf{Z} \in \mathbb{R}^{n \times d}$ with $d \ll D$, we evaluate how well the low-dimensional space preserves the geometric and informational properties of original embedding space.

Notation for Evaluation Metrics. Let n be the number of samples. For each sample i, $\mathbf{x}_i \in \mathbb{R}^D$ is its original D-dimensional embedding, and $\mathbf{z}_i \in \mathbb{R}^d$ is its compressed d-dimensional representation. Pairwise Euclidean distances in the original and compressed spaces are $\delta_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ and $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$, respectively. The set of k nearest neighbours of sample i in the high-dimensional space is $\mathcal{N}_k(i)$, and in the low-dimensional space is $\mathcal{N}'_k(i)$. The rank of sample j in the neighbourhood of i is $r(i,j)$ in the original space and $r'(i,j)$ in the compressed space.

We categorise our intrinsic metrics along three orthogonal axes: **local neighbourhood fidelity**, **global geometric structure**, and **spectral and information-theoretic content**. This multidimensional perspective ensures a comprehensive characterisation of compression effects, from fine-grained local relationships to broader manifold structures and core informational content.

3.1 Local neighbourhood Fidelity

The preservation of local neighbourhood structures is critical, as these structures often encode subtle semantic similarities vital for many tasks.

Trustworthiness (T_k) and Continuity (C_k) (Venna and Kaski, 2001) These two metrics evaluate the reliability of local neighbourhoods. Trustworthiness (T_k) measures how many false neighbours are introduced by the DR process. Specifically, it measures the extent to which points that appear close in the compressed space ($\mathbf{Z}$) were not actually close in the original space ($\mathbf{X}$). A high T_k value indicates that the DR method does not create spurious local relationships.

$$T_k = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in U_k(i)} (r(i,j) - k)$$

where $U_k(i) = \mathcal{N}'_k(i) \setminus \mathcal{N}_k(i)$. Continuity ($C_k$), in contrast, measures how many true neighbours from the original space $\mathbf{X}$ are lost in the compressed space $\mathbf{Z}$. A high C_k indicates that the DR method successfully preserves original local relationships.

$$C_k = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in V_k(i)} (r'(i,j) - k)$$

where $V_k(i) = \mathcal{N}_k(i) \setminus \mathcal{N}'_k(i)$. Together, T_k and C_k provide a robust measure of how faithfully local manifold structures are maintained.

Mean Relative Rank Error (MRRE$_k$) (Lee and Verleysen, 2007) Beyond simple neighbourhood overlap, MRRE$_k$ measures the average proportional change in the ranks of those neighbours that are *preserved* within the top-k set after compression. A lower MRRE$_k$ value indicates that the relative ordering of neighbours is mostly unchanged. This suggests that the compression preserves fine-grained local distances. It also means that the local metric structure experiences minimal distortion.

$$\mathrm{MRRE}_k = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_k(i)} \frac{|r(i,j) - r'(i,j)|}{r(i,j)}.$$

Neighbourhood Precision at k (NP$_k$) This metric measures the overlap between the top-k neighbours in the original and compressed spaces. It quantifies how many true neighbours are retained after compression, offering a direct and intuitive measure of local structure preservation.

$$\mathrm{NP}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{|\mathcal{N}_k(i) \cap \mathcal{N}'_k(i)|}{k}.$$

Local Average Procrustes Measure (LPro) (Schönemann, 1966) This metric measures the preservation of local neighbourhood geometry by averaging Procrustes disparities across all points. For each point i, its k-nearest neighbours in $\mathbf{X}$ and $\mathbf{Z}$ forming sets $\mathcal{N}_X(i)$ with embeddings and $\mathcal{N}_Z(i)$ with embeddings $\mathbf{Z}_{\mathcal{N}_Z(i)}$.

A local Procrustes alignment is performed between these two neighbourhoods. Each set is centred, and an optimal local rotation $\mathbf{R}_i$, and local scaling factor ρ_i are computed to best align the centred neighbourhood $\mathbf{Z}_{\mathcal{N}_Z(i),c}$ to $\mathbf{X}_{\mathcal{N}_X(i),c}$, minimizing the Frobenius norm of their differences. The normalised disparity for point i is then calculated as:

$$\text{Disparity}_i = \frac{\left\|\mathbf{X}_{\mathcal{N}_X(i),c} - \rho_i \mathbf{Z}_{\mathcal{N}_Z(i),c}\mathbf{R}_i\right\|_F^2}{\left\|\mathbf{X}_{\mathcal{N}_X(i),c}\right\|_F^2}.$$

A low LPro indicates that the geometric structure of the local neighbours around each point is well-preserved. This indicates robustness against local distortions such as shearing or anisotropic scaling, thereby maintaining the relative distances, angles, and overall configuration within the neighbourhood. Such preservation of fine-grained local structure is often critical for tasks that rely on nuanced semantic similarity and precise neighbour identification.

3.2 Global Geometry Fidelity

Preserving the global geometry of the embedding space is crucial for tasks that rely on broader semantic relationships, such as clustering or topic modelling. This involves maintaining the overall shape of the data manifold and the relative positions of distant points or clusters.

Kruskal's Stress (KS) (Kruskal, 1964) Kruskal's Stress (KS) measures the overall distortion of pairwise distances among all samples. It calculates the normalised sum of squared differences between distances in $\mathbf{X}$ (δ_{ij}) and $\mathbf{Z}$ (d_{ij}). A lower KS indicates better preservation of the global metric structure, meaning that the large-scale arrangement of embeddings and inter-cluster separations are well maintained.

$$\text{KS} = \sqrt{\frac{\sum_{i<j}(\delta_{ij} - d_{ij})^2}{\sum_{i<j}\delta_{ij}^2}}.$$

Distance Correlation (Spearman's ρ and Pearson's r) We compute Spearman's rank correlation and Pearson's linear correlation between all pairwise distances $\{\delta_{ij}\}$ and $\{d_{ij}\}$. High positive correlations indicate that the relative ordering (Spearman) and linear relationship (Pearson) of inter-sample distances are preserved, maintaining the global similarity structure after compression.

Global Procrustes Measure (GPro) (Schönemann, 1966) This metric measures the overall structural difference between $\mathbf{X}$ and $\mathbf{Z}$. It finds an optimal rigid transformation (including orthogonal rotation $\mathbf{R}$, uniform scaling ρ, and translation, though translation is handled by centring the data) that minimises the sum of squared differences between the transformed $\mathbf{Z}$ and $\mathbf{X}$. A low GPro indicates that the overall shape and orientation of the point cloud are well-preserved after this optimal alignment, showing robustness to global distortions. The error is calculated as the sum of squared Frobenius norms of the differences, normalised by the sum of squared Frobenius norm of the centred original embeddings:

$$\text{GPro} = \frac{\|\mathbf{X}_c - \rho\mathbf{Z}_c\mathbf{R}\|_F^2}{\|\mathbf{X}_c\|_F^2}.$$

3.3 Spectral Retention

This dimension evaluates how well statistical information and dominant data directions are preserved, which often correspond to key semantic axes within the embedding space.

Explained Variance Ratio (EVR) When the compression method allows (e.g., PCA, or by comparing $\mathbf{Z}$ to a PCA of $\mathbf{X}$), EVR measures the proportion of total variance in $\mathbf{X}$ that is captured by $\mathbf{Z}$. A high EVR indicates that the principal components of semantic variation are preserved, minimising significant information loss. This metric is most directly interpretable for linear DR methods. For non-linear methods, EVR is computed based on the variance of $\mathbf{Z}$ and $\mathbf{X}$.

$$\text{EVR} = \frac{\text{tr}(\text{Cov}(\mathbf{Z}))}{\text{tr}(\text{Cov}(\mathbf{X}))}.$$

Pairwise Inner-Product (PIP) Loss (Yin and Shen, 2018) Inner products are fundamental to many similarity measures (e.g., cosine similarity). The PIP loss measures the squared Frobenius norm of the difference between the Gram matrices ($\mathbf{X}\mathbf{X}^\top$ and $\mathbf{Z}\mathbf{Z}^\top$). A low PIP loss indicates that key angular relationships and dot product magnitudes are well-preserved across the dataset.

$$\text{PIP} = \|\mathbf{X}\mathbf{X}^\top - \mathbf{Z}\mathbf{Z}^\top\|_F^2.$$

Eigenspace Overlap (EOS) (May et al., 2019) The comparison of linear subspaces, typically defined by the principal eigenvectors or singular vec-

Algorithm 1 Residual Eigenspace Overlap Score (EOS_k)

Require: $X \in \mathbb{R}^{n \times D}$, $Z \in \mathbb{R}^{n \times d}$, k, N_{sub}
Ensure: EOS_k
for $B \in \{X, Z\}$ **do**
  $(\_, \_, V_B^\top) \leftarrow \text{SVD}(B)$
  $V_B^{(k)} \leftarrow V_{B,1:k}$
  $B' \leftarrow B - B\,V_B^{(k)}(V_B^{(k)})^\top$
  $(U_{B'}, \_, \_) \leftarrow \text{SVD}(B')$
  $r_B \leftarrow \text{rank}(B')$
  $m_B \leftarrow \min(N_{\text{sub}}, r_B)$
end for
$N \leftarrow \min(m_X, m_Z)$
if $N = 0$ **then**
  if $r_X = 0$ and $r_Z = 0$ **then**
    return 1.0
  else
    return 0.0
  end if
end if
$U_X^* \leftarrow U_{X',1:N}, \quad U_Z^* \leftarrow U_{Z',1:N}$
$M \leftarrow (U_X^*)^\top U_Z^*$
$(\sigma_1, \ldots, \sigma_N) \leftarrow \text{SingularValues}(M)$
return $\frac{1}{N} \sum_{i=1}^{N} \sigma_i^2$

tors of data matrices, is a common method to understand structural similarities. EOS measures the degree of alignment or shared variance between these subspaces. It indicates whether different datasets or data representations of the same data emphasise similar underlying factors or directions of maximum variance. A high EOS shows that the main geometric or statistical features captured by one space are also prominent in the other. This alignment is quantified by first identifying the primary directional axes (eigenvectors) for each data representation. Then, one set of these axes is projected onto the other, and the sum of the squared strengths of these projections indicates the total overlap between the two subspaces.

4 EOS_k

While the spectral metrics (Section 3.3) offer valuable insights, their effectiveness can be compromised in the context of modern, anisotropic, text embeddings. In such cases, the metric may overstate the quality of preservation by capturing alignment in high-variance directions that lack meaningful semantic content, thereby obscuring the degradation of more subtle, task-relevant structures.

4.1 Anisotropy and Rogue Dimensions

A core assumption of many spectral metrics—that preserving high-variance directions ensures the retention of salient information—is often misleading for modern embeddings (e.g., BERT, E5). This is due to *anisotropy*, a property where variance is concentrated in a few dominant "rogue dimensions." These dimensions disproportionately inflate similarity scores while contributing little to downstream tasks.

This discrepancy is illustrated in Figure 1. The **top panel** shows each dimension's contribution to cosine similarity, revealing how a handful of rogue dimensions dominate the score while most contribute almost nothing. Formally, the contribution of dimension i to the cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$ is $CC_i = \frac{u_i v_i}{\|\mathbf{u}\|\|\mathbf{v}\|}$ (Timkey and van Schijndel, 2021). Rogue dimensions consistently have large-magnitude values, thus dominating this sum. The **bottom panel**, in contrast, shows logistic regression weights ($\mathbf{w}$) for a downstream classification task. Here, importance is spread more evenly across dimensions, and the rogue dimensions are appropriately down-weighted, as their large, task-agnostic variance provides little predictive power.

This fundamental misalignment between what is structurally dominant and what is semantically useful necessitates a more robust evaluation approach. To address this, we propose the **Residual Eigenspace Overlap Score** (EOS_k), a novel metric designed to look beyond these confounding high-variance components. EOS_k concentrates on the semantic content embedded within the *residual* eigenspace—the subspace remaining after the top-k dominant principal components are removed from both the original and compressed embeddings. This approach is motivated by prior work (Raunak et al., 2019; Timkey and van Schijndel, 2021) showing that leading components often capture task-agnostic noise. By intentionally excluding them, EOS_k offers a more faithful measure of meaningful semantic preservation, as detailed in Algorithm 1.

4.2 Determining the k Parameter for EOS_k

A critical aspect of EOS_k is the choice of k, the number of top principal components to remove. This choice is not arbitrary; it is grounded in a data-driven analysis of the embedding space's anisotropy. Specifically, k is determined by analysing the geometry of the embedding space

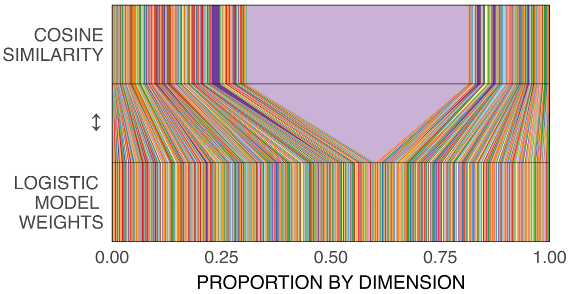


Figure 1: Comparison of each dimension's contribution to cosine similarity (top) and its logistic-regression weight (bottom) for E5 (1024d) on `ToxicConversationsClassification` dataset.

by observing how each dimension contributes to a standard similarity measure like cosine similarity.

In Figure 1, the **top panel** reveals a clear anisotropic pattern: only 1–3 dimensions dominate similarity, while the majority contribute negligibly. The **bottom panel**, however, shows that these dominant components are down-weighted by the classifier, whereas lower-variance dimensions play a meaningful role. Our experiments consistently revealed a single overwhelmingly dominant component in both BERT and E5 embeddings. Accordingly, we set $k = 1$ in all experiments reported in this paper.

5 Experimental Setup

We evaluate the proposed framework by performing a correlation analysis using Spearman's rank correlation coefficient to examine how well intrinsic evaluation metrics align with performance on extrinsic tasks, consistent with prior intrinsic evaluation studies (May et al., 2019). High correlation indicates that intrinsic metrics align well with downstream task performance and can therefore serve as reliable and cost-efficient proxies for evaluating embedding compression quality.

For each dataset and embedding family, the experimental pipeline proceeds as follows: (i) generate original sentence embeddings; (ii) apply a range of compression techniques; (iii) compute the proposed intrinsic scores; and (iv) evaluate the compressed embeddings on standard downstream tasks. The intrinsic and extrinsic outcomes are subsequently correlated to assess alignment. Figure 2 provides a schematic overview of this pipeline, from embedding generation through compression, evaluation, and correlation analysis.

Beyond correlation, the framework analyses how compression perturbs embedding classes, identifying which structural properties—local, global, or spectral—are most critical to preserve.

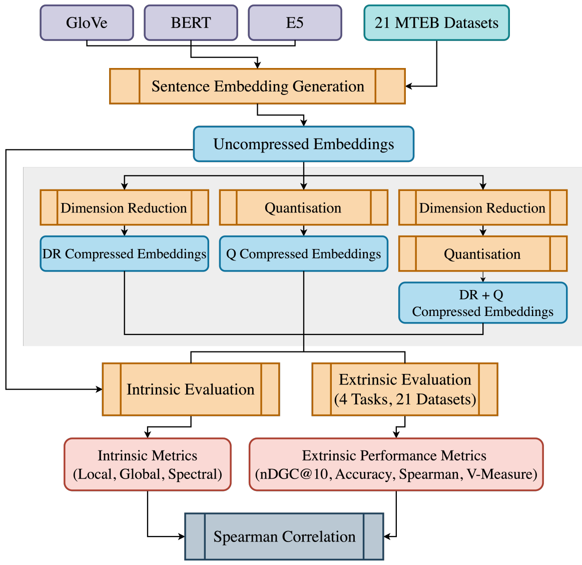


Figure 2: Overview of the experimental workflow designed to validate the proposed intrinsic evaluation framework.

5.1 Downstream Tasks

To validate our framework, we follow the MTEB benchmark (Muennighoff et al., 2023) and use four tasks: Retrieval, Semantic Textual Similarity (STS), Clustering, and Classification.

For STS and Retrieval, we compute cosine similarity between sentence embeddings. Clustering is performed using mini-batch k-means (batch size 32, number of clusters = number of gold labels), and Classification uses logistic regression with a maximum of 100 iterations.

For each task, we select representative datasets to ensure broad domain coverage while maintaining computational efficiency. Dataset details are provided in Appendix D; full statistical details can be found at (Muennighoff et al., 2023). We follow MTEB's standard evaluation protocols and primary metrics (Retrieval: *nDCG@10*, STS: *Spearman correlation*, Clustering: *V-measure*, Classification: *Accuracy*). The summary of evaluation metrics is provided in Appendix C.

5.2 Embedding Models and Sentence Representation

We select three representative embedding models covering static, contextual, and LLM-based types:

1. **GloVe (Static)** (Pennington et al., 2014): We

Group	Metric	GloVe				BERT				E5			
		CLF	CLU	IR	STS	CLF	CLU	IR	STS	CLF	CLU	IR	STS
Local	T_k	**0.67**†	0.24	0.55†	0.62†	**0.84**†	0.33	0.27	0.34	0.75†	0.28	**0.76**†	0.83†
	C_k	0.65†	0.49	**0.69**†	0.64†	**0.81**†	0.22	0.15	0.34	0.64†	0.35	**0.80**†	0.87†
	MRRE_k	0.65†	0.52†	**0.71**†	0.65†	**0.80**†	0.19	0.16	0.42	0.64†	0.35	**0.80**†	0.88†
	NP_k	0.67†	0.59†	**0.70**†	0.65†	**0.80**†	0.28	0.20	0.45	0.65†	0.35	0.80†	**0.87**†
	LPro	0.65†	**0.76**†	**0.76**†	**0.76**†	0.51†	**0.90**†	0.86†	0.81†	0.36	**0.86**†	0.66†	0.50
Global	KS	0.18	0.30	0.27	**0.58**†	0.20	0.08	**0.30**	0.18	0.17	0.19	0.24	**0.32**
	SDC	**0.34**	0.11	**0.34**	0.27	**0.55**†	0.44	0.54†	0.46	**0.72**†	0.37	0.57†	0.48
	PDC	0.35	0.20	**0.40**	0.34	**0.64**†	0.35	0.40	0.36	**0.75**†	0.30	0.63†	0.56†
	GPro	**0.70**†	0.59†	**0.70**†	0.69†	**0.74**†	0.36	0.44	0.48	0.48	0.46	0.81†	**0.84**†
Info. Ret.	EVR	**0.29**	0.12	0.24	0.13	**0.30**	0.16	0.18	0.16	0.41	0.21	0.44	**0.53**†
	PIP Loss	0.16	0.30	0.28	**0.60**†	0.13	0.13	**0.26**	0.24	0.12	**0.27**	0.26	0.26
	EOS	0.60†	0.71†	**0.72**†	0.63†	0.18	0.32	**0.38**	0.24	0.15	0.09	0.12	0.30
	EOS_k (Ours)	0.52†	0.42	**0.66**†	0.47	0.45	0.44	**0.57**†	0.60†	0.35	0.50	**0.54**†	0.49

Table 1: Spearman Correlation between intrinsic evaluation metrics and downstream task performance. Each correlation value is computed across all datasets within each task category (classification, clustering, IR, and STS), covering all dimensionality reduction, quantisation, preprocessing methods, and embedding dimensions as described in Section 5. Results for EOS_k are reported with $k = 1$. Task abbreviations: CLF = Classification, CLU = Clustering, IR = Information Retrieval, STS = Semantic Textual Similarity. $\dagger$ denotes statistical significance ($p < 0.01$). **Boldfaced** values indicate the highest correlation for each metric-task pair.

use `glove.840B.300d`[5] embeddings. Sentence embeddings are computed by averaging word vectors of lowercased tokens obtained via simple whitespace tokenisation.

2. **BERT (Contextual)** (Devlin et al., 2019): We use `bert-base-uncased`[6] embeddings. Sentence embeddings are obtained using the 'sentence-transformers' library[7], via mean-pooling of the last hidden state's token embeddings or the [CLS] token representation, depending on the specific model's configuration.

3. **E5 (LLM-based)** (Wang et al., 2022): We use `E5-large-v2`[8] embeddings. We obtain sentence embeddings using the default MTEB framework mechanisms when loading the embedding via the 'sentence-transformers' library.

6 Results

Table 1 presents correlation analysis results across diverse embeddings and tasks, demonstrating that several intrinsic metrics align closely with downstream behaviour.

6.1 Local Structure Preservation

Metrics evaluating local structure preservation prove to be highly reliable indicators of downstream task performance, particularly for classification. Traditional neighbourhood-based metrics such as **Trustworthiness** (T_k), **Continuity** (C_k), and **Neighbourhood Precision** (NP_k) consistently yield strong correlations with classification performance across all models, especially for BERT ($\rho \approx 0.80 - 0.84$). For E5 embeddings, these metrics are also exceptionally predictive of Information Retrieval (IR) and Semantic Textual Similarity (STS) performance, with correlations reaching as high as $\rho = 0.88$. However, the standout metric in this category is the **Local Procrustes** (LPro) measure. It achieves remarkable correlations with clustering performance for BERT ($\rho = 0.90$) and E5 ($\rho = 0.86$), tasks where all other local metrics showed limited predictive power.

The strong performance of metrics like T_k and C_k suggests that for discriminative tasks like classification, maintaining the identity of immediate neighbours is paramount. The introduction of false neighbours or the loss of true ones directly degrades performance. Conversely, LPro's unique success in predicting clustering performance for BERT and E5 indicates that for these complex, anisotropic embeddings, preserving the local *geometric configuration* (the "shape" of the neighbourhood) is more critical than preserving the exact set

[5] https://nlp.stanford.edu/projects/glove/
[6] https://huggingface.co/google-bert/bert-base-uncased
[7] https://huggingface.co/sentence-transformers
[8] https://huggingface.co/intfloat/e5-large-v2

of neighbours. This insight is crucial, as it suggests that effective compression for clustering tasks must prioritise the retention of local manifold structures over simple neighbourhood overlap.

6.2 Global Geometric Fidelity

The utility of global structure metrics is more varied, with measures focused on geometric shape significantly outperforming those based on pairwise distances. Classical metrics like **Kruskal's Stress** (KS), which penalise all pairwise distance errors, are poor predictors for modern embeddings, showing negligible correlation with performance for BERT and E5. In contrast, the **Global Procrustes** (GPro) measure, which assesses the preservation of the entire point cloud's shape under optimal rigid alignment, emerges as a robust indicator. GPro shows consistently high correlations for GloVe across all tasks ($\rho \approx 0.70$), for BERT classification ($\rho = 0.74$), and is exceptionally predictive for E5 on IR ($\rho = 0.81$) and STS ($\rho = 0.84$).

The failure of KS suggests that a strict, global preservation of all pairwise distances is not aligned with how modern embeddings encode semantic information, likely due to their anisotropic nature. GPro's success, however, demonstrates that maintaining the overall geometric configuration of the embedding space is a far more meaningful objective. For models like E5, the high GPro correlations on IR and STS tasks imply that the large-scale thematic organisation of concepts is vital for performance. This highlights the importance of using metrics that are sensitive to the global "silhouette" of the data rather than to absolute distance fidelity.

6.3 Spectral Information Retention

Traditional spectral metrics that focus on dominant variance components are generally poor predictors of performance for modern transformer-based embeddings. Metrics such as **Explained Variance Ratio** (EVR) and **Pairwise Inner-Product (PIP) Loss** show weak correlations for BERT and E5. The most striking finding is the failure of the standard **Eigenspace Overlap Score** (EOS). While EOS is a strong, consistent predictor for GloVe embeddings across all tasks (ρ values between 0.60 and 0.72), its predictive power plummets for BERT and E5, with correlations often falling below $\rho = 0.30$.

This dramatic performance drop for modern embeddings provides strong evidence of the confounding effects of **anisotropy**. For models like BERT and E5, the top principal components, which carry the most variance, do not necessarily align with the most semantically informative directions. Consequently, metrics like EOS, which exclusively evaluate the alignment of these dominant (but potentially task-irrelevant) subspaces, are fundamentally limited. This finding underscores the inadequacy of standard spectral methods for evaluating contemporary embeddings and directly motivates the need for metrics that can analyse structure beyond these misleading dominant components.

6.4 Residual Eigenspace Overlap (EOS_k)

Our proposed **Residual Eigenspace Overlap Score** (EOS_k) was designed specifically to address the limitations of standard EOS for anisotropic embeddings. By first removing the top-k dominant principal components and then measuring the alignment of the remaining, or *residual*, eigenspaces, EOS_k focuses on less dominant but more semantically rich structural information. The results compellingly validate this approach. For BERT, EOS_k delivers a marked improvement over standard EOS, yielding substantial correlations for IR ($\rho = 0.57$) and STS ($\rho = 0.60$). A similar, significant improvement is observed for E5, with EOS_k showing notable correlations for Clustering ($\rho = 0.50$) and IR ($\rho = 0.54$).

The superior performance of EOS_k confirms that for models like BERT and E5, a significant amount of task-relevant information resides in the **residual eigenspace**, not the dominant one. By successfully capturing the preservation of this "deeper" structure, EOS_k provides a much more reliable signal of downstream performance for modern embeddings. This validates EOS_k as a critical tool for evaluating compression techniques, demonstrating that to accurately predict performance on anisotropic embeddings, it is essential to look beyond the statistically dominant, and often semantically noisy, directions of variance.

7 Framework in Practice: Less is More

Our multi-metric evaluation framework not only aligns well with extrinsic (downstream) performance but also offers practical guidance for embedding compression and dimensionality reduction.

Random Projections: Simple, Fast, Effective Random Projections (RP) perform consistently well across all embeddings (GloVe, BERT, E5), reduction ratios (25%, 50%, 75%), and even under `int8` quantisation (Figure 3). Despite its simplicity,

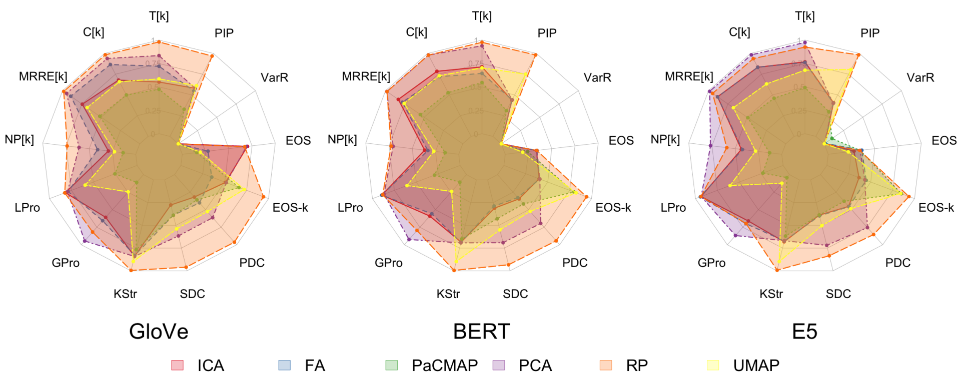


Figure 3: Comparative performance of dimensionality reduction techniques across intrinsic metrics for GloVe, BERT, and E5 sentence embeddings. Results shown are averaged over multiple reduction ratios (25%, 50%, 75%) and utilise int8 quantisation.

RP preserves both local (LPro, TW, CONT) and global (GPro) structures effectively, supported by the Johnson–Lindenstrauss lemma, which guarantees that RP can preserve pairwise distances well in high-dimensional spaces. Its inherent randomness and lack of complex learned transformations may also prevent over-fitting to specific geometric idiosyncrasies of the original embedding space, leading to a robust preservation of diverse structural qualities. This suggests that RP, often considered a baseline, can serve as a robust and efficient alternative to more complex DR methods, especially in scenarios where computational overhead or implementation simplicity are key constraints.

Quantisation: Compress Without Compromise We observed negligible differences between `int8` quantised and full-precision embeddings across all metrics and methods (see Section E). This suggests quantisation does not significantly distort embedding quality. This indicates that quantisation can be adopted as a lightweight yet effective compression step post-reduction, enabling substantial memory and storage savings without sacrificing the core representational qualities of the embeddings.

8 Conclusion

We introduced a scalable and interpretable intrinsic evaluation framework for compressed text embeddings over DR methods, combining local, global, and spectral fidelity metrics. We introduce a novel metric, EOS_k, that captures meaningful information beyond dominant principal components. We validate our framework using three embeddings across four tasks and 21 datasets. Experiments revealed that our framework robustly predicts downstream task performance across diverse downstream tasks.

Key findings highlight that optimal compression strategies are model-dependent: contextual embeddings benefit most from preserving local neighbourhood structures, while static and contrastive embeddings show stronger alignment with global and spectral fidelity. Notably, EOS_k revealed the importance of retaining information beyond dominant principal components, showing significant correlations for BERT and E5, particularly in tasks like STS. Our analysis also identified Random Projections as a highly efficient and effective dimensionality reduction technique, and we recommend the routine application of `int8` quantisation for further compression with minimal performance loss. Ultimately, this work provides a principled and interpretable framework, empowering more efficient and informed development of compressed embedding solutions.

Limitations

While this study introduces a robust framework for evaluating compressed embeddings, its scope has several limitations that provide clear avenues for future research. Our findings are based on a set of only three representative text embeddings, and our evaluation is **monolingually focused** on English language datasets. Consequently, the conclusions may not fully generalise to the entire landscape of available embeddings, especially those with different architectures, or to other languages with different morphological structures. Furthermore, our investigation of compression methods was not exhaustive. We focused on a select group of dimensionality reduction techniques and `int8` quantisation, leaving other promising techniques such as **pruning**, **knowledge distillation**, and more aggressive, **lower-bit quantisation** schemes unexplored. Finally, our reliance on **Spearman correlation** as the sole metric provides a valuable macro-level view of trends but may obscure practical utility, where a developer typically needs to select the single *best* compression method for a task. Future work should therefore incorporate a **top-1 accuracy** metric—how often the intrinsic framework's top-scored method aligns with the actual best-performing method—and compare it against a strong baseline to better quantify the practical, decision-making value of the framework.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Dimitris Achlioptas. 2003. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687.

Aishwarya Bhandare, Vamsi Sripathi, Deepthi Karkada, Vivek Menon, Sun Choi, Kushal Datta, and Vikram Saletore. 2019. Efficient 8-bit quantization of transformer neural machine language translation model. *arXiv preprint arXiv:1906.00532*.

Ummara Bibi, Mahrukh Mazhar, Dilshad Sabir, Muhammad Fasih Uddin Butt, Ali Hassan, Mustansar Ali Ghazanfar, Arshad Ali Khan, and Wadood Abdul. 2024. Advances in pruning and quantization for natural language processing. *IEEE Access*, 12:139113–139128.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Pierre Comon. 1994. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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: Human Language Technologies, volume 1 (long and short papers)*, pages 4171–4186.

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)*, pages 55–65.

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*, pages 6894–6910.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. 2022. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42, Berlin, Germany. ACL.

Nazish Hina, Mohand Boughanem, and Taoufiq Dkaki. 2024. Token pruning by dimensionality reduction methods on TCT-ColBERT for reranking. In *Foundations of Intelligent Systems*, pages 65–74, Cham. Springer Nature Switzerland.

GE Hinton and RR Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Naamán Huerga-Pérez, Rubén Álvarez, Rubén Ferrero-Guillén, Alberto Martínez-Gutiérrez, and Javier Díez-González. 2025. Optimization of embeddings storage for rag systems using quantization and dimensionality reduction techniques. *arXiv preprint arXiv:2505.00105*.

Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2023. Exploring dimensionality reduction techniques in multilingual transformers. *Cognitive Computation*, 15(2):590–612.

Dae Yon Hwang, Bilal Taha, and Yaroslav Nechaev. 2023. EmbedTextNet: Dimension reduction with weighted reconstruction and correlation losses for efficient text embedding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9863–9879.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713.

William B Johnson, Joram Lindenstrauss, and 1 others. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):1.

Takuya Kataiwa, Cho Hakaze, and Tetsushi Ohki. 2025. Measuring intrinsic dimension of token embeddings. *arXiv preprint arXiv:2503.02142*.

Daniyal Kazempour, Claudius Zelenka, Atakan Kara, Andreas Lohrer, and Peer Kröger. 2024. Do good scores imply good embeddings? On the necessity of inspecting manifold learning evaluation results by multiple criteria. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 317–324. IEEE.

Joseph B Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Jiedong Lang, Zhehao Guo, and Shuyu Huang. 2024. A comprehensive study on quantization techniques for large language models. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 224–231.

John A Lee and Michel Verleysen. 2007. *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Vivian Liu and Yiqiao Yin. 2024. Green AI: exploring carbon footprints, mitigation strategies, and trade offs in large language model training. *Discover Artificial Intelligence*, 4(1):49.

Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. Dimension reduction for efficient dense retrieval via conditional autoencoder. *arXiv preprint arXiv:2205.03284*.

Avner May, Jian Zhang, Tri Dao, and Christopher Ré. 2019. On the downstream performance of compressed word embeddings. *Advances in Neural Information Processing Systems*, 32.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loic 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*, pages 2014–2037.

Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2024. When text embedding meets large language model: A comprehensive survey. *arXiv preprint arXiv:2412.09165*.

Seyed Parsa Neshaei, Yasaman Boreshban, Gholamreza Ghassem-Sani, and Seyed Abolghasem Mirroshandel. 2024. The impact of quantization on the robustness of transformer-based text classifiers. *arXiv preprint arXiv:2403.16016*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. On effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243.

Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. No parameter left behind: How distillation and model size affect zero-shot retrieval. *arXiv preprint arXiv:2206.02873*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. In *Communications of the ACM*, volume 63, pages 54–63.

Praneet Sherki, Samarth Navali, Ramesh Inturi, and Vanraj Vala. 2021. Retaining semantic data in binarized word embedding. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 130–133.

Charles Spearman. 1904. "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–293.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4821–4836. ACL.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546. ACL.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Laurens van der Maaten. 2009. Learning a parametric embedding by preserving local structure. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 384–391. PMLR.

Jarkko Venna and Samuel Kaski. 2001. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks — ICANN 2001*, pages 485–491, Berlin, Heidelberg. Springer Berlin Heidelberg.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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*, volume 119, pages 9929–9939.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

Jintang Xue, Yun-Cheng Wang, Chengwei Wei, C-C Jay Kuo, and 1 others. 2024. Word embedding dimension reduction via weakly-supervised feature selection. *APSIPA Transactions on Signal and Information Processing*, 13(1).

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. 2023. Discovering universal geometry in embeddings with ICA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4647–4675.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183.

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. *Advances in Neural Information Processing Systems*, 31.

Gaifan Zhang, Yi Zhou, and Danushka Bollegala. 2024. 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 2024)*, pages 6530–6543.

A Related Work

Embeddings The evolution of embeddings has progressed from static models (e.g., Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017)) to contextual embeddings (e.g., BERT (Devlin et al., 2019), SimCSE (Gao et al., 2021), and E5 (Wang et al., 2024a)). Static embeddings capture global co-occurrence statistics, and contextual models provide dynamic representations sensitive to neighbour context. Nie et al. (2024); Zhang et al. (2024) explore LLMs, including decoder-based architectures' capability to generate embeddings and indicate that LLMs serve as competitive embedding generators, outperforming traditional models on downstream retrieval and classification tasks. These embeddings come with higher dimensionality and computational requirements, bringing the necessity of DR for real-world applications.

Dimensionality Reduction of Embeddings DR methods such as PCA (Hotelling, 1933), t-SNE (van der Maaten, 2009), UMAP (McInnes et al., 2018) and autoencoders (Hinton and Salakhutdinov, 2006) are used to compress vector spaces while aiming to preserve important features like semantic similarity and cluster structure. DR has been applied primarily to static embeddings such as Word2Vec and GloVe, showing that unsupervised methods (i.e., PCA) can substantially reduce the dimension of embeddings without performance degradation (Raunak et al., 2019; Mu and Viswanath, 2018). Zhang et al. (2024) explore the effectiveness of DR methods on sentence embeddings, showing that aggressive compression (e.g., to half the original size) can be achieved with minor downstream performance loss. Huertas-García et al. (2023) extend exploration of DR methods on embeddings into multilingual settings, indicating language-specific variance in DR behaviour. Huerga-Pérez et al. (2025) explore DR methods to RAG embeddings, showing that PCA (standard and Kernel) demonstrates the best performance in preserving retrieval quality evaluated on MTEB benchmark for the IR task.

Evaluation of Dimension Reduced Embeddings A key limitation in DR research is its heavy dependence on downstream task performance as the main evaluation criterion. Task-specific performance is informative, but it can not fully isolate the contribution of the reduced embeddings from classifier-specific effects (Zhang et al., 2024). This makes it hard to clearly evaluate how much of the performance is due to the DR method itself. On the other hand, intrinsic evaluation metrics such as T_k, continuity, and neighbourhood preservation, provide a more principled view of how well reduced embeddings maintain the geometric and structural properties of the original space. Gladkova and Drozd (2016) highlight the importance of integrating these metrics into evaluation to gain a deeper understanding of embedding quality. Finally, Kazempour et al. (2024) explores the quality not only with downstream tasks but also different evaluation criteria for Computer vision (CV).

Our study aims to address the following gaps in the literature: (i) compression embeddings are often evaluated in isolation, focusing narrowly on a single embedding type; (ii) evaluations frequently consider only a limited range of compression methods; and (iii) there is an over-reliance on downstream accuracy as the primary evaluation metric. To address these gaps, we propose a unified evaluation framework that systematically benchmarks multiple compression methods.

Method	Linearity	Local/Global	Time Complexity
PCA	Linear	Global	$\mathcal{O}(nd^2)$
ICA	Linear	Global	$\mathcal{O}(nd^2)$
RP	Linear	Global	$\mathcal{O}(ndk)$
FA	Linear	Global	$\mathcal{O}(nd^2)$
UMAP	Nonlinear	Local	$\mathcal{O}(n \log n)$
PaCMAP	Nonlinear	Local	$\mathcal{O}(n \log n)$

Table 2: Comparison of DR methods across key characteristics. n represents the number of samples, d represents the original input dimension and k the reduced output dimension.

B Compression Methods

B.1 Dimensionality Reduction

We explore a diverse set of DR methods, spanning a broad spectrum of algorithmic philosophies, encompassing classical linear projections, statistical decomposition methods, and contemporary non-linear manifold learning approaches. This diversity is critical for understanding the landscape of DR performance on text embeddings, which are known to possess complex, often non-linear, intrinsic structures (Kataiwa et al., 2025). A summary of DR methods, highlighting their key characteristics such as linearity, local/global preservation, and time complexity, is presented in Table 2.

Principal Component Analysis (PCA) (Hotelling, 1933): A cornerstone of linear DR, PCA identifies orthogonal directions (principal components) that capture the maximum variance in the data. The transformation is defined by projecting the data onto the subspace spanned by the top d principal components: $f_{\text{PCA}}(\mathbf{X}) = \mathbf{X}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{D\times d}$ is the matrix whose columns are the leading eigenvectors of the covariance matrix of $\mathbf{X}$, and $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. Its inclusion is motivated by its ubiquity, computational efficiency, and its utility as a baseline for variance-preserving linear transformations.

Independent Component Analysis (ICA) (Comon, 1994): Unlike PCA, ICA aims to decompose a multivariate signal into a set of statistically independent, non-Gaussian components. It seeks a linear transformation $f_{\text{ICA}}(\mathbf{X}) = \mathbf{X}\mathbf{W}$ such that the columns of the resulting $\mathbf{Z}$ are as statistically independent as possible, typically by maximising a measure of non-Gaussianity. ICA is selected for its potential to uncover underlying latent factors that may be more semantically meaningful than principal components, especially when the sources are not orthogonal.

Random Projection (RP) (Achlioptas, 2003): RP offers a computationally efficient, data-oblivious DR method grounded in the Johnson-Lindenstrauss lemma (Johnson et al., 1984). It projects data onto a lower-dimensional space using a random matrix $\mathbf{R} \in \mathbb{R}^{D \times d}$, where entries are typically drawn from a Gaussian or sparse Rademacher distribution: $f_{\text{RP}}(\mathbf{X}) = \frac{1}{\sqrt{d}}\mathbf{XR}$. RP is chosen for its scalability, theoretical guarantees on preserving pairwise distances (in expectation), and its utility in scenarios where constructing a data-dependent projection is computationally prohibitive.

Factor Analysis (FA) (Spearman, 1904): FA is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. It models the data $\mathbf{x}$ as $\mathbf{x} = \mathbf{Wz} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\mathbf{z}$ is a vector of d latent factors (typically assumed to be $\mathcal{N}(0, \mathbf{I})$), $\mathbf{W} \in \mathbb{R}^{D \times d}$ is the factor loading matrix, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\epsilon}$ is a vector of unique error terms, often assumed to be $\mathcal{N}(0, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ being a diagonal covariance matrix. FA is included to evaluate a generative linear model that explicitly accounts for measurement error, contrasting with PCA's variance-maximisation approach.

Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018): UMAP is a non-linear DR method based on manifold learning principles and topological data analysis. It constructs a high-dimensional graph representation of the data and then optimises a low-dimensional graph to be as structurally similar as possible. This is achieved by minimising the cross-entropy between fuzzy simplicial sets representing the neighbourhood structure in the high and low dimensions. UMAP is selected for its prowess in capturing complex global and local manifold structures, often outperforming linear methods and other non-linear methods like t-SNE (Liu et al., 2022) in preserving topological properties, which can be crucial for text embedding semantics.

Pairwise Controlled Manifold Approximation and Projection (PaCMAP) (Wang et al., 2021): PaCMAP is a more recent non-linear DR method designed to offer a better balance between local and global structure preservation than UMAP, while also being computationally efficient. It utilises a graph-based approach with a carefully designed loss function that incorporates mid-range pairwise distances and local neighbourhood preservation through graph degree. PaCMAP is included to benchmark a state-of-the-art manifold learner that aims to address some limitations of UMAP (e.g., sensitivity to initialisation), particularly concerning the overemphasis on local structure and the separation of global clusters.

B.2 Quantisation

In addition to dimensionality reduction, quantisation serves as an orthogonal and often complementary compression strategy. Quantisation reduces the numerical precision of the embedding values, thereby decreasing the memory footprint required to store each individual scalar component of an embedding vector.

We focus on 8-bit integer ('int8') quantisation, a widely adopted technique offering a trade-off between compression ratio and performance, with hardware support on modern CPUs and GPUs (Gholami et al., 2022). Given an embedding matrix $\mathbf{X} \in \mathbb{R}^{n \times D}$ (which could be the original embeddings or dimensionally reduced embeddings $\mathbf{Z}$), 'int8' quantisation maps the floating-point values ('float32') in $\mathbf{X}$ to 8-bit integers. For each scalar:

$$x_{quant} = \text{round}(\frac{x_{float}}{S} + ZP)$$

where x_{float} is the original floating-point value, x_{quant} is the quantised 8-bit integer, S is a floating-point scale factor, and ZP is an integer zero-point. The scale S and zero-point ZP are determined by the range of the floating-point values being quantised (e.g., min-max quantisation):

$$S = \frac{\max(x_{float}) - \min(x_{float})}{2^B - 1}$$

$$ZP = \text{round}(-\frac{\min(x_{float})}{S}) - 2^{B-1}$$

where $B = 8$ for 'int8' quantisation. The de-quantisation step to approximate the original floating-point value is:

$$x_{approx\_float} = S \cdot (x_{quant} - ZP)$$

It offers a direct 4x reduction in model size if converting from 'float32' (32 bits per value to 8 bits per value) without altering the embedding dimension. This can lead to substantial memory savings and faster data transfer, which are critical for on-device deployment and large-scale retrieval systems. Furthermore, operations on ‘int8’ data types can be significantly faster on compatible hardware accelerators (Jacob et al., 2018). We apply post-training quantisation to both original and reduced embeddings. We evaluate ‘int8’ as a stand-alone method and in combination with DR. We use ‘int8’ specifically over lower bit-depths (e.g., ‘int4’) due to its proven effectiveness with minimal performance loss across downstream tasks (Bhandare et al., 2019; Yao et al., 2022; Tao et al., 2022; Parsa Neshaei et al., 2024; Huerga-Pérez et al., 2025).

C Summary of Evaluation Metrics

The summary of evaluation metrics is given in Table 3.

Metric	Description
T_k	False neighbours introduced in $\mathbf{Z}$
C_k	Loss of true neighbours from $\mathbf{X}$ in $\mathbf{Z}$
$MRRE_k$	Change in neighbour ranks after compression
NP_k	Overlap of top-k neighbours before/after DR
LPro	Local geometric configuration (shapes/angles)
KS	Global distance distortion
SDC/PDC	Correlation of all pairwise distances
GPro	Global point cloud shape/orientation
EVR	Variance retained post-DR
PIP	Inner-product structure preservation
EOS	Alignment of dominant subspaces
EOS_k	Overlap after removing top-k PCs

Table 3: Summary of evaluation metrics.

D MTEB Benchmark Results

D.1 Datasets

We evaluate our framework on four tasks from the MTEB benchmark:

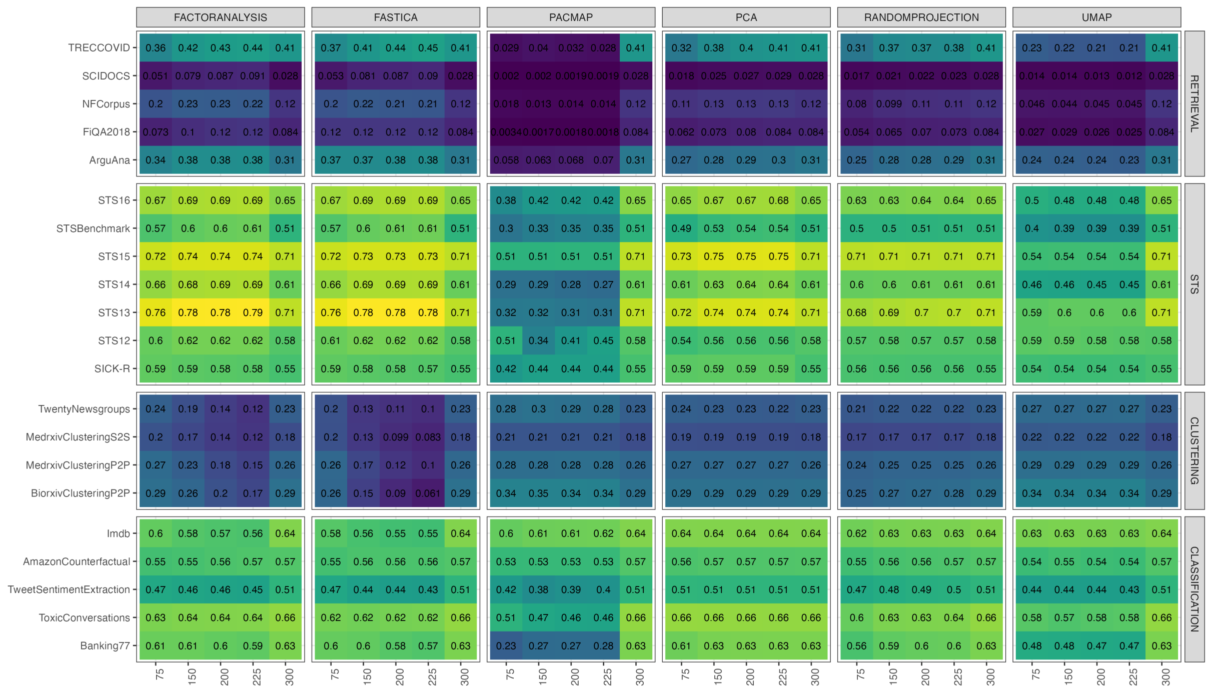
1. **Retrieval:** ArguAna, FiQA2018, NFCorpus, SCIDOCS, TRECCOVID
2. **Semantic Textual Similarity (STS):** SICKRSTS, STS12STS, STS13STS, STS14STS, STS15STS, STSBenchmarkSTS, STS16STS
3. **Clustering:** BiorxivClusteringP2P, MedrxivClusteringP2P, MedrxivClusteringS2S, TwentyNewsgroupsClustering
4. **Classification:** Banking77Classification, ToxicConversationsClassification, TweetSentimentExtractionClassification, AmazonCounterfactualClassification, ImdbClassification

D.2 Results

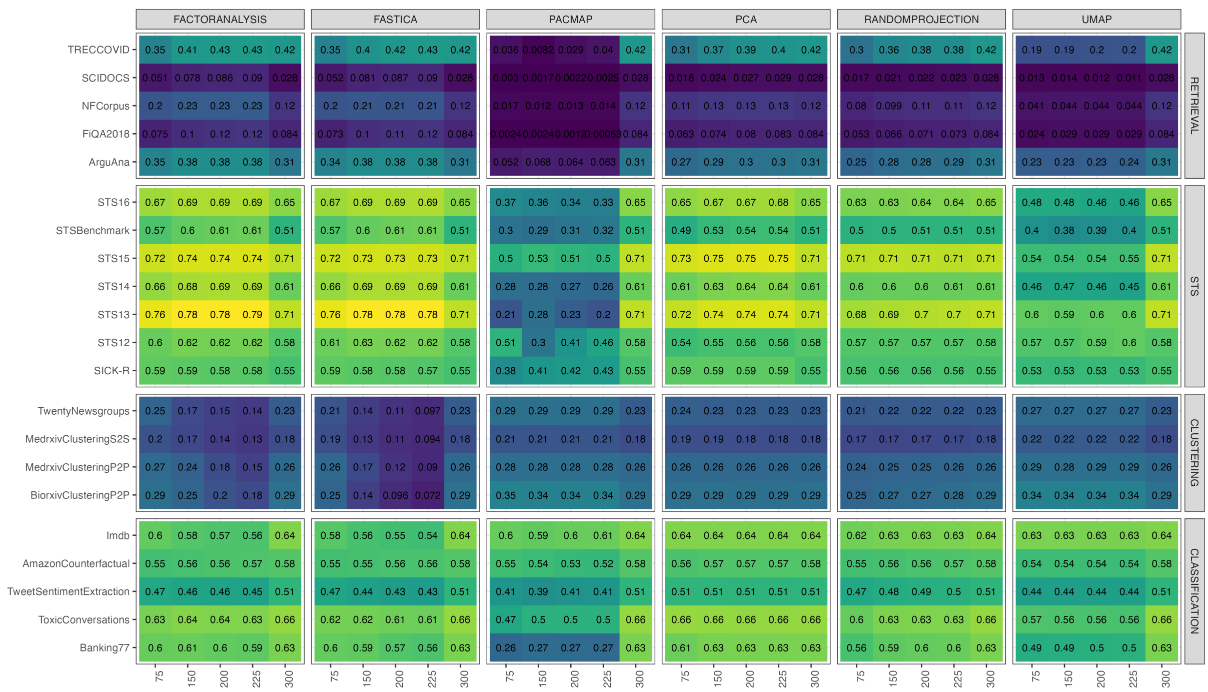
We evaluate compressed embeddings both with and without quantisation across downstream tasks. Results for GloVe are shown in Figure 4, for BERT in Figure 5, and for E5 embeddings in Figure 6.

E Evaluation of Compression Methods

This section presents the results of our extrinsic evaluation (Section 3) using radar plots to compare dimensionality reduction methods across embedding types. Figures 7, 8, and 9 show the averaged performance of GloVe, BERT, and E5 embeddings across intrinsic metrics under three different experimental settings.

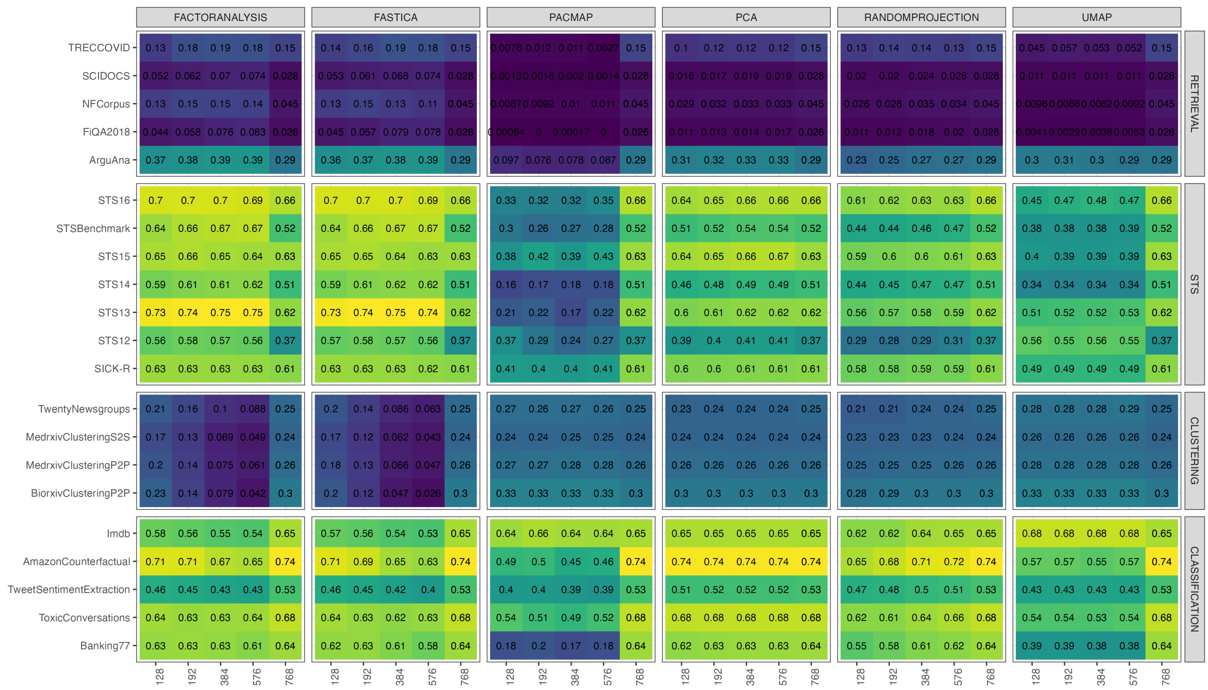


(a) Without quantisation

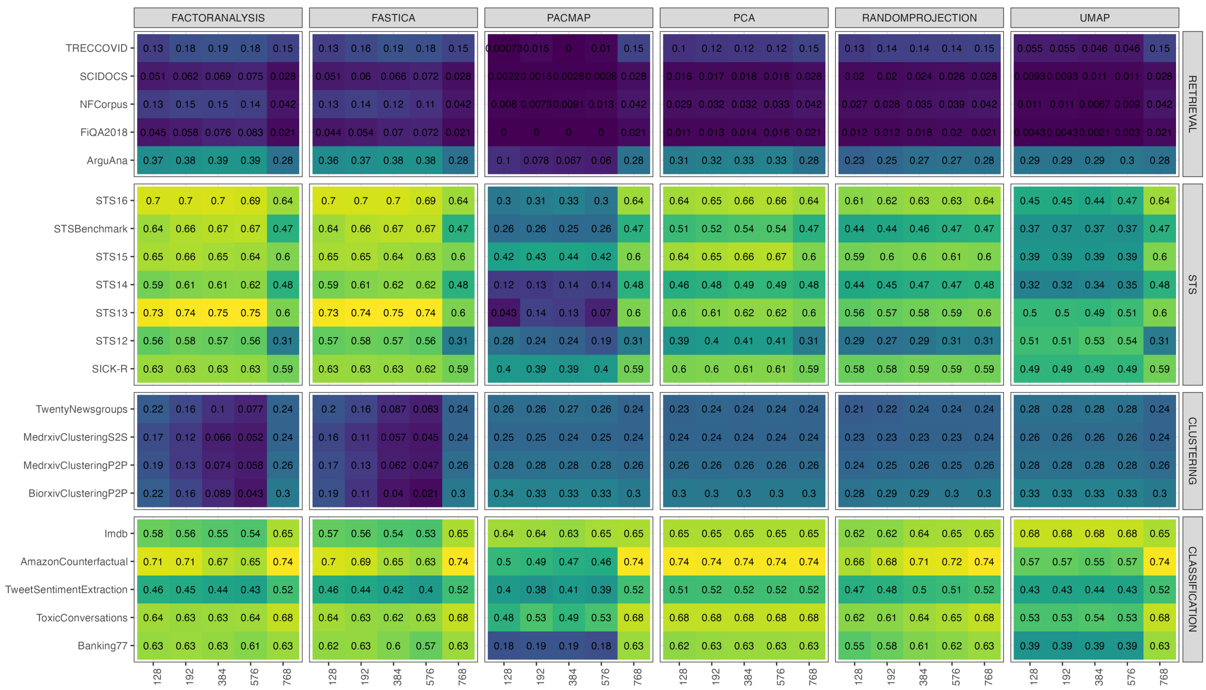


(b) With quantisation

Figure 4: Heatmap of primary evaluation metrics for 21 datasets for four tasks (retrieval, STS, clustering and classification) using GLoVe embeddings.

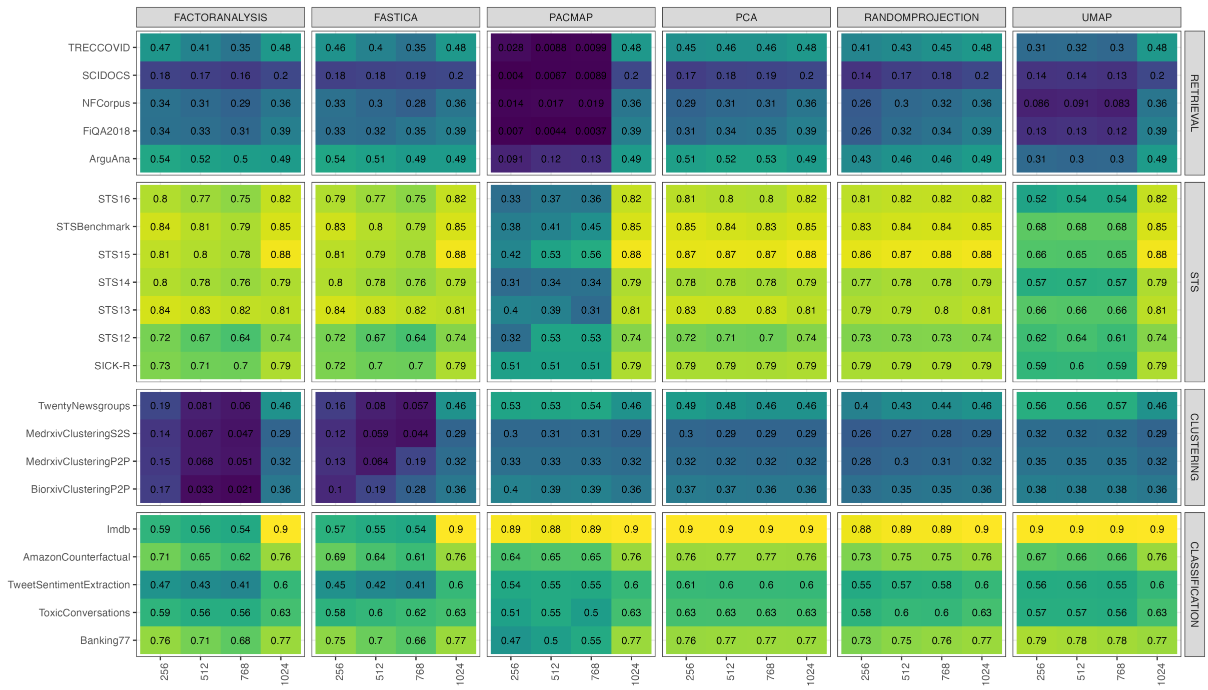


(a) Without quantisation

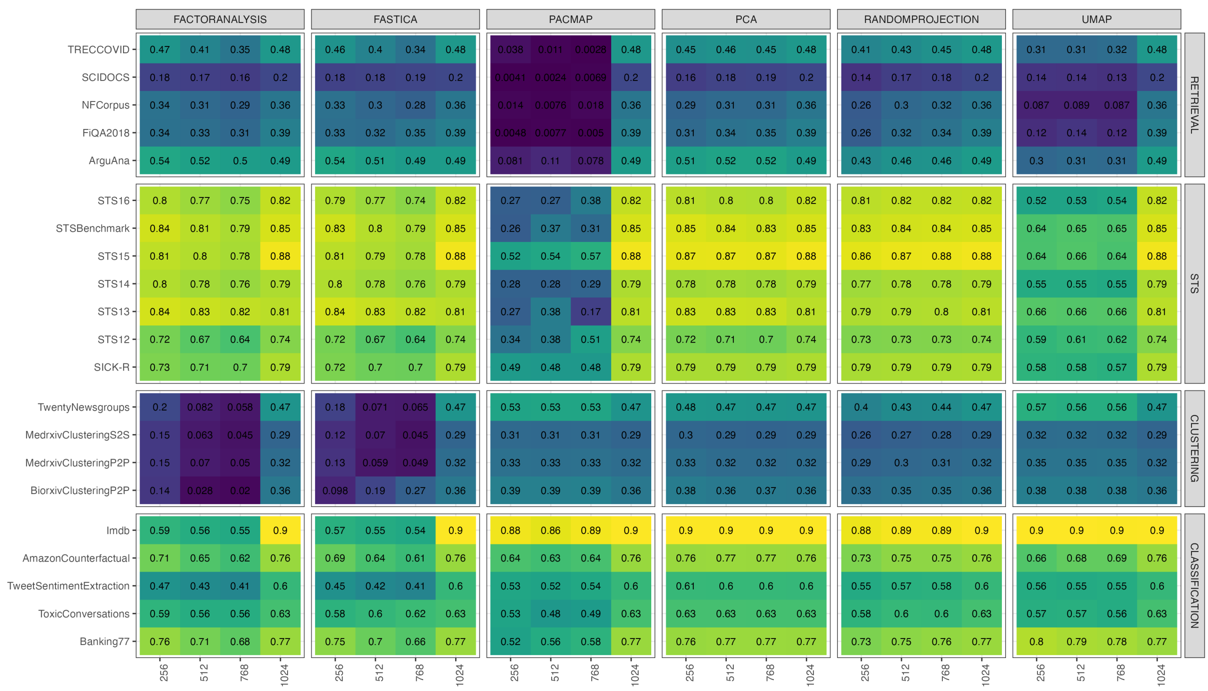


(b) With quantisation

Figure 5: Heatmap of primary evaluation metrics for 21 datasets for four tasks (retrieval, STS, clustering and classification) using BERT embeddings.



(a) Without quantisation



(b) With quantisation

Figure 6: Heatmap of primary evaluation metrics for 21 datasets for four tasks (retrieval, STS, clustering and classification) using E5 embeddings.

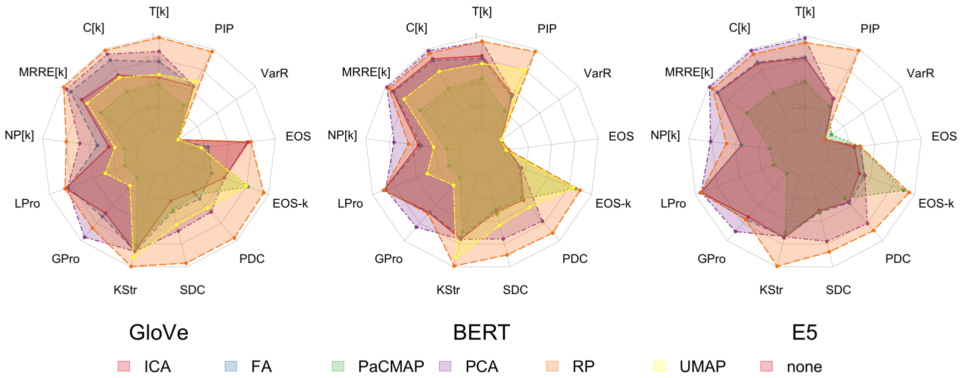


Figure 7: Comparative performance of dimensionality reduction across intrinsic metrics for GloVe, BERT, E5 sentence embeddings. Results shown are averaged over reduction ratios and utilise no quantisation and no preprocessing.

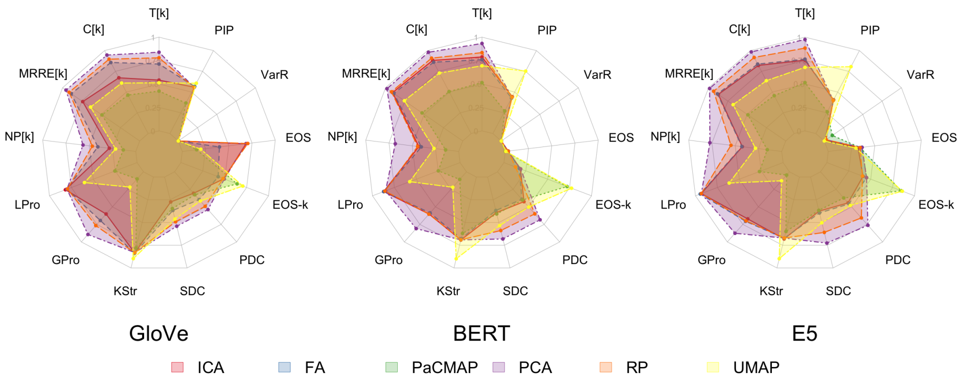


Figure 8: Comparative performance of dimensionality reduction across intrinsic metrics for GloVe, BERT, E5 sentence embeddings. Results shown are averaged over reduction ratios and utilise no quantisation and standardisation.

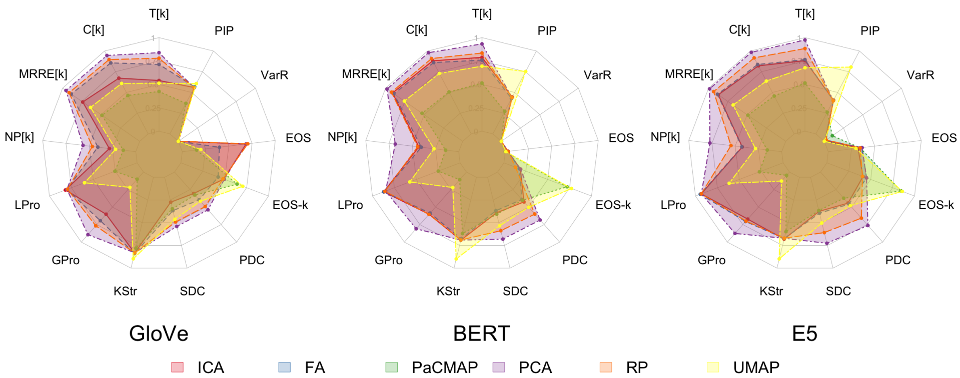


Figure 9: Comparative performance of dimensionality reduction across intrinsic metrics for GloVe, BERT, E5 sentence embeddings. Results shown are averaged over reduction ratios and utilise int8 quantisation and standardisation.