

Spectral Tempering for Embedding Compression in Dense Passage Retrieval

Yongkang Li

y.li7@uva.nl

University of Amsterdam
Amsterdam, The Netherlands

Panagiotis Eustratiadis

p.efstratiadis@uva.nl

University of Amsterdam
Amsterdam, The Netherlands

Evangelos Kanoulas

e.kanoulas@uva.nl

University of Amsterdam
Amsterdam, The Netherlands

Abstract

Dimensionality reduction is critical for deploying dense retrieval systems at scale, yet mainstream post-hoc methods face a fundamental trade-off: principal component analysis (PCA) preserves dominant variance but underutilizes representational capacity, while whitening enforces isotropy at the cost of amplifying noise in the heavy-tailed eigenspectrum of retrieval embeddings. Intermediate spectral scaling methods unify these extremes by reweighting dimensions with a power coefficient γ , but treat γ as a fixed hyperparameter that requires task-specific tuning. We show that the optimal scaling strength γ is not a global constant: it varies systematically with target dimensionality k and is governed by the signal-to-noise ratio (SNR) of the retained subspace. Based on this insight, we propose Spectral Tempering (**SpecTemp**), a learning-free method that derives an adaptive $\gamma(k)$ directly from the corpus eigenspectrum using local SNR analysis and knee-point normalization, requiring no labeled data or validation-based search. Extensive experiments demonstrate that Spectral Tempering consistently achieves near-oracle performance relative to grid-searched $\gamma^*(k)$ while remaining fully learning-free and model-agnostic. Our code is publicly available at <https://anonymous.4open.science/r/SpecTemp-0D37>.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; • **Computing methodologies** → **Natural language processing**.

Keywords

Dense Retrieval, Embedding Compression, Principal Component Analysis

ACM Reference Format:

Yongkang Li, Panagiotis Eustratiadis, and Evangelos Kanoulas. 2026. Spectral Tempering for Embedding Compression in Dense Passage Retrieval. In . ACM, New York, NY, USA, 6 pages.

1 Introduction

Dense retrieval has become the dominant paradigm for first-stage retrieval in modern search systems [10, 25, 33], where queries and documents are encoded as high-dimensional embeddings and relevance is computed via similarity functions such as cosine similarity. While recent encoders based on Large Language Models (LLMs) [14, 17, 36] achieve state-of-the-art (SOTA) performance, they routinely

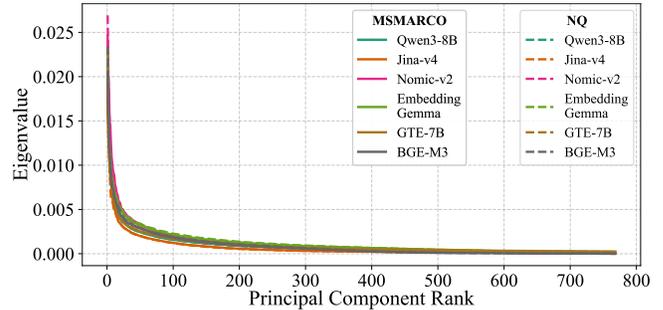


Figure 1: Consistent spectral structure of dense retrieval embeddings. Eigenvalue distributions from 1M sampled embeddings on MS MARCO and NQ exhibit consistent heavy-tailed decay across diverse retrievers, revealing a head–tail signal-to-noise ratio (SNR) gradient—leading components are signal-dominant while tail dimensions grow noise-prone—motivating dimensionality-adaptive tempering.

produce high-dimensional embeddings (e.g., 1024–4096), increasing the memory footprint of vector indexes and the cost of similarity computation in large-scale deployment.

To mitigate these costs, training-based approaches such as learned projections [35], conditional autoencoders [16], and knowledge distillation [15] have been explored, but require retraining infrastructure tied to specific encoders. Consequently, post-hoc compression—reducing dimensionality without parameter updates—offers a more practical alternative, yet its dominant baselines occupy flawed extremes. Principal Component Analysis (PCA) retains maximal variance [34] but leaves the energy distribution highly skewed, allowing head dimensions to overshadow complementary discriminative signals. Conversely, standard whitening [28] enforces isotropy by normalizing all dimensions to unit variance; yet the eigenspectrum of retrieval embeddings is heavily tailed (Figure 1), and this normalization substantially amplifies noise. Intermediate spectral scaling methods attempt to resolve this dilemma by weighting dimensions with a fractional power $\lambda_i^{-\gamma/2}$ ($\gamma \in [0, 1]$) [27]. However, prior work treats γ as a static hyperparameter that requires per-task tuning, overlooking that optimal tempering varies systematically with the target dimensionality k . For instance, aggressive whitening ($\gamma \approx 1$) benefits compact subspaces ($k = 64$) but degrades quality at large k by amplifying low SNR tail components.

In this work, we formalize this dimensionality-dependent behavior through a local SNR analysis of the corpus eigenspectrum. By estimating a spectral noise floor, we obtain an SNR profile that



This work is licensed under a Creative Commons Attribution 4.0 International License. SIGIR '26, July 20–24, 2026, © 2026 Copyright held by the owner/author(s).

reveals a smooth head–tail transition from signal-dominant to noise-prone components—explaining why optimal tempering strength should decrease as target dimensionality k grows to include low-SNR tail directions. Building on this insight, we propose Spectral Tempering (**SpecTemp**), a learning-free method that analytically derives an adaptive $\gamma(k)$ directly from the SNR profile, automatically interpolating between variance preservation (PCA) and isotropy (whitening). The resulting linear transform is computed offline from corpus embeddings and applied identically to queries at inference time, requiring no labeled data or validation tuning.

Our contributions are three-fold:

- We characterize the *dimensionality-dependent* optimality of spectral scaling, demonstrating that the ideal γ is intrinsically governed by the subspace SNR rather than being a fixed constant.
- We propose **SpecTemp**, a learning-free method that analytically derives an adaptive $\gamma(k)$ from the corpus eigenspectrum, requiring no labeled data or validation-based tuning.
- We conduct extensive experiments across multiple LLM-based embedding models and diverse retrieval datasets, demonstrating that SpecTemp consistently achieves near-oracle performance relative to grid-searched $\gamma^*(k)$.

2 Related Work

Dense Retrieval. Dense retrieval has evolved from BERT-based bi-encoders [3, 6, 10, 33] with compact 768d representations to massive LLM-based architectures. To capture complex semantics, recent SOTA models like RepLLaMA [19], E5-Mistral [32], and Qwen3-Embedding [36] employ billion-scale, often decoder-only backbones. While yielding superior generalization, this shift often produces high-dimensional embeddings (e.g., 4096d), creating the storage bottlenecks that motivate our study.

Embedding Compression. Strategies to mitigate these overheads fall into two broad categories: training-based and post-hoc.

Training-based methods optimize compression objectives during or after training-time. Matryoshka Representation Learning (MRL) [11] has gained widespread adoption for enabling flexible truncation by nesting information in prefix dimensions. Other approaches employ knowledge distillation to transfer capabilities to smaller students [15], or optimize conditional autoencoders to compress fixed embeddings into latent codes [16]. While effective, these strategies require additional training data and incur high computational costs for retraining, rendering them impractical for off-the-shelf or API-only models.

Post-hoc methods, in contrast, transform pretrained embeddings without parameter updates. Spectral projections dominate this landscape, scaling dimensions based on their eigenvalues. PCA ($\gamma = 0$) maximizes variance but leaves the space anisotropic [18, 34, 37], while Standard Whitening ($\gamma = 1$) enforces isotropy but risks amplifying tail noise [7, 28]. Intermediate strategies employ a fractional exponent $\gamma \in [0, 1]$ to interpolate between these extremes [27], yet they rely on a static hyperparameter requiring per-task tuning. Alternatively, Random Projection offers dimension-agnostic compression via the Johnson–Lindenstrauss lemma [9] but ignores the learned manifold structure.

A separate line of work targets isotropy via post-processing, such as removing dominant directions [21, 23, 24] or mapping to uniform distributions [13], though these focus on quality rather than dimensionality reduction. Similarly, Product Quantization (PQ) [8] and its variants achieve index-level compression via codebooks [4]; being a downstream operation, this approach is orthogonal to and composable with linear projections like ours.

SpecTemp occupies a distinct position in this landscape: it is a *post-hoc, learning-free* linear projection that derives a dimensionality-adaptive tempering strength $\gamma(k)$ from the local SNR of the retained subspace, requiring no labeled data, retraining, validation-based tuning, or index-level modifications.

3 Methodology

We now describe Spectral Tempering (**SpecTemp**), a post-hoc compression method that derives a dimensionality-adaptive tempering exponent $\gamma(k)$ directly from the eigenspectrum of corpus embeddings. The method proceeds in three stages: spectral decomposition, SNR-guided exponent derivation, and embedding transformation.

3.1 Spectral Decomposition

Given a corpus embedding matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, we first center it by subtracting the column-wise mean $\boldsymbol{\mu}$:

$$\bar{\mathbf{X}} = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top \quad (1)$$

Centering reduces the influence of a global offset direction and yields a more stable covariance spectrum; we apply the same corpus-derived centering to both queries and documents to preserve geometric consistency. We then compute the eigendecomposition of the covariance matrix:

$$\mathbf{C} = \frac{1}{n-1} \bar{\mathbf{X}}^\top \bar{\mathbf{X}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top \quad (2)$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$, and $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_d]$ are the corresponding eigenvectors.

3.2 SNR-Guided Exponent Derivation

The core insight of Spectral Tempering is that the appropriate tempering strength should be governed by the signal quality of the retained subspace. We formalize this through a local SNR analysis.

Noise Floor Estimation. We estimate the noise floor σ_{noise}^2 as the mean eigenvalue of the spectral tail:

$$\sigma_{\text{noise}}^2 = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \lambda_i \quad (3)$$

where \mathcal{T} denotes the last 10% of eigenvalue indices. As shown in Figure 1, diverse retrieval encoders exhibit a consistently heavy-tailed eigenspectrum whose tail consistently plateaus into a stable noise floor, making this region a reliable, model-agnostic anchor for noise estimation. We verify in Section 4.2.4 that SpecTemp is insensitive to the exact percentile choice, confirming that this default requires no per-task tuning.

Local SNR Computation. The local SNR at rank i measures the excess energy above the noise floor:

$$\text{SNR}(i) = \max\left(0, \frac{\lambda_i - \sigma_{\text{noise}}^2}{\sigma_{\text{noise}}^2}\right) \quad (4)$$

We note that this quantity is not intended as a generative statistical estimate in the sense of spiked covariance models, but as a monotonic, spectrum-level proxy for relative signal dominance—sufficient for calibrating the tempering exponent. This quantity is large for head components where the signal dominates, and vanishes in the tail where eigenvalues converge to the noise floor.

Anchor Point and Adaptive $\gamma(k)$. To derive $\gamma(k)$ without task-specific tuning, we need a reference point that separates the high-confidence signal regime from the transitional regime. We identify this anchor as the *knee point* of the SNR curve—the rank at which SNR transitions from rapid to gradual decay—detected via the Knee-dle algorithm [26]. Let k_{knee} denote this rank and $S_{\text{ref}} = \text{SNR}(k_{\text{knee}})$ the corresponding SNR value.

Since the k -th component defines the noise bottleneck of the retained subspace, we use its SNR as a conservative proxy for subspace signal quality. This ensures that the tempering strength is constrained by the worst-case noise exposure rather than being overly influenced by optimistic, high-variance directions. The adaptive exponent for target dimensionality k is then:

$$\gamma(k) = \min\left(1, \frac{\text{SNR}(k)}{S_{\text{ref}}}\right) \quad (5)$$

Normalizing by S_{ref} ensures that all target dimensionalities within the high-SNR regime ($k \leq k_{\text{knee}}$) receive full whitening ($\gamma = 1$), while dimensions beyond the knee are progressively tempered. We adopt a linear mapping between SNR and γ following the principle of parsimony, as this simple formulation avoids introducing additional degrees of freedom and is empirically sufficient and robust. The resulting behavior is as desired: small k yields $\gamma(k) \approx 1$ (near-whitening); as k grows and incorporates noisier components, $\gamma(k)$ monotonically decreases toward 0 (near-PCA).

3.3 Transformation

Given target dimensionality k , we construct the transformation matrix by combining the top- k eigenvectors with the derived exponent:

$$\mathbf{W}_k = \mathbf{U}_k \cdot \text{diag}\left(\lambda_1^{-\gamma(k)/2}, \dots, \lambda_k^{-\gamma(k)/2}\right) \quad (6)$$

where $\mathbf{U}_k = [\mathbf{u}_1, \dots, \mathbf{u}_k] \in \mathbb{R}^{d \times k}$. The compressed embedding for any input \mathbf{x} (query or document) is:

$$\mathbf{y} = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{W}_k \in \mathbb{R}^k \quad (7)$$

The eigendecomposition is computed once on a corpus sample; the resulting $\boldsymbol{\mu}$ and \mathbf{W}_k are then applied identically to documents (offline) and queries (online), ensuring compatibility with standard ANN indexing. When the downstream similarity metric is cosine similarity, the transformed vectors are additionally L2-normalized.

4 Experiments

In this section, we present empirical evaluations to validate the effectiveness of **SpecTemp** across diverse retrieval datasets.

4.1 Experiment Setup

4.1.1 Datasets. We evaluate on four retrieval datasets: **MS MARCO** Passage Ranking [1] for web search, **Natural Questions (NQ)** [12] for open-domain QA, **FEVER** [30] for evidence retrieval in fact

Table 1: Retrieval model statistics (Params: total parameters; MRL: Matryoshka Representation Learning support).

Model	Params	Dim	Length	MRL	Release
Qwen3-8B	8.0B	4096	32k	✓	Jun 2025
Jina-v4	3.8B	2048	32k	✓	Jun 2025
Nomic-v2	475M	768	512	✓	Feb 2025
EmbeddingGemma	308M	768	2048	✓	Sep 2025
GTE-7B	7.0B	3584	32k	✗	Jun 2024
BGE-M3	560M	1024	8192	✗	Feb 2024

verification, and **FiQA** [20] for domain-specific financial retrieval, covering diverse domains and scales.

4.1.2 Retrieval Models. We experiment on six widely used open-source dense retrievers spanning different scales and embedding dimensions, as summarized in Table 1. Four models (**Qwen3-8B** [36]¹, **Jina-v4** [5]², **Nomic-v2** [22]³, **EmbeddingGemma** [31]⁴) support Matryoshka Representation Learning, providing strong truncation baselines. Two models (**GTE-7B** [14]⁵, **BGE-M3** [2]⁶) lack native MRL support, testing the generality of post-hoc compression.

4.1.3 Baselines. We compare against representative learning-free post-hoc methods that require no labeled data or model fine-tuning. **Prefix Truncation** retains the first k dimensions (standard for MRL-compatible models). **Random Truncation** subsamples k dimensions, a simple strategy shown to be surprisingly competitive in recent work [29]. **Random Projection** compresses via a Gaussian random matrix as a theoretical baseline. For spectral methods, we evaluate **PCA** ($\gamma = 0$), **Standard Whitening** ($\gamma = 1$), and γ -**Whitening** with a fixed $\gamma = 0.5$ to represent static power normalization. All spectral transformations are derived from the corpus.

4.1.4 Evaluation Protocol. We evaluate all models at target dimensions $k \in \{768, 512, 256, 128, 64\}$. For models with a native dimension of 768, the $k = 768$ case coincides with no dimensionality reduction. We report **MRR@10** for MS MARCO and **nDCG@10** for the remaining datasets.

4.1.5 Implementation Details. All spectral decompositions and transformations are implemented in NumPy. Embeddings are generated using the original model checkpoints with default configurations. The covariance matrix and noise-floor statistics are estimated from the document corpus of each dataset, using up to 1M randomly sampled documents or the full corpus when fewer are available. Experiments are conducted on a cluster with 4×NVIDIA H100 GPUs.

4.2 Experiment Results

4.2.1 Main Results. We focus our main analysis on three representative models (Qwen3-8B, Jina-v4, GTE-7B) covering diverse architectures and scales. As shown in Table 2, our SpecTemp method achieves the best or tied-best performance among spectral methods in the majority of configurations without any tuning. PCA

¹<https://huggingface.co/Qwen/Qwen3-Embedding-8B>

²<https://huggingface.co/jinaai/jina-embeddings-v4>

³<https://huggingface.co/nomic-ai/nomic-embed-text-v2-moe>

⁴<https://huggingface.co/google/embeddinggemma-300m>

⁵<https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct>

⁶<https://huggingface.co/BAAl/bge-m3>

Table 2: Retrieval performance on four datasets at target dimensions k . Bold denotes the best per column within each model. All results are averaged over three random seeds (1999, 5, 2026). Superscript ^{ns} indicates that, for all three runs, the difference from *Full Dimension* is not significant (two-sided paired t -test, $p < 0.05$). Absence of ^{ns} indicates significance in at least one run.

Model	Method ↓	$k \rightarrow$	MS MARCO				NQ				FEVER				FiQA						
			768	512	256	128	64	768	512	256	128	64	768	512	256	128	64	768	512	256	128
Qwen3-8B	Full Dimension		36.8				64.9				91.8				64.7						
	Prefix Truncation	36.4	35.7	34.5	32.4	28.2	63.7	63.2	61.1	57.3	49.1	91.8^{ns}	91.6	91.2	90.1	85.5	63.9	63.5	61.6	57.4	51.3
	Random Truncation	35.9	35.5	34.3	31.5	24.8	63.5	62.5	59.3	53.2	40.0	91.6^{ns}	91.3	90.8	89.0	79.6	63.0	61.6	59.4	53.4	41.7
	Random Projection	36.2	35.7	34.3	32.0	26.4	63.6	62.5	60.4	55.5	44.2	91.5	91.3	90.6	89.4	82.3	63.0	62.6	59.8	55.5	45.5
	PCA	36.0	35.5	34.1	31.3	25.0	64.6 ^{ns}	63.8	62.1	57.2	47.0	91.0	90.7	89.6	87.4	83.1	63.7	63.7	62.1	59.2	53.2
	Whitening	34.9	35.2	34.1	32.3	26.9	61.9	62.3	61.8	58.3	49.4	91.0	90.8	89.8	87.1	83.5	58.4	59.8	60.5	59.0	52.8
	γ -Whitening	35.9	35.5	34.6	32.3	26.4	64.1	63.9	62.5	58.8	49.1	91.1	91.0	89.9	87.6	83.6	62.6	62.4	62.5	59.8	53.8
	SpecTemp	36.1	35.6	34.6	32.4	26.8	64.9^{ns}	64.1	62.6	58.7	49.4	91.2	90.9	89.9	87.2	83.5	64.0	63.7	62.8	59.7	53.1
Jina-v4	Full Dimension		32.1				61.6				87.8				47.7						
	Prefix Truncation	31.4	31.1	30.2	28.1	21.9	60.8	60.3	57.8	53.3	41.5	87.4	87.3	85.8	82.2	67.0	47.0 ^{ns}	46.8 ^{ns}	44.3	40.6	31.1
	Random Truncation	31.4	30.9	29.4	26.7	20.1	60.4	59.4	56.1	50.1	35.9	87.0	86.4	84.1	78.4	60.9	46.4	45.2	42.6	37.2	27.7
	Random Projection	31.1	30.7	29.2	26.5	20.8	59.9	59.4	56.3	50.7	38.1	86.8	86.1	84.7	78.4	63.4	46.0	45.5	43.4	38.0	28.3
	PCA	31.9 ^{ns}	31.5	30.4	27.5	18.6	61.9 ^{ns}	61.6 ^{ns}	60.1	55.8	43.0	87.4	87.0	85.4	81.5	69.0	46.8	46.7	45.6	43.1	37.7
	Whitening	29.0	30.0	30.7	29.8	24.2	56.2	57.1	57.8	56.4	49.6	84.9	85.1	84.7	82.0	71.4	41.0	41.6	42.4	42.1	36.8
	γ -Whitening	31.3	31.7 ^{ns}	31.4	29.5	22.7	60.3	60.6	60.5	58.2	49.3	87.1	87.0	86.0	82.6	71.7	45.7	45.7	45.2	43.9	37.8
	SpecTemp	31.9 ^{ns}	31.8	31.2	29.3	23.7	62.1	61.7 ^{ns}	61.0	58.3	49.8	87.5	87.1	85.8	82.6	71.7	47.2 ^{ns}	47.0	45.6	43.9	37.2
GTE-7B	Full Dimension		39.1				66.8				95.2				61.8						
	Prefix Truncation	38.4	38.0	36.9	34.2	28.7	65.3	64.7	62.0	57.0	47.2	95.0	95.0	94.5	93.8	91.2	59.8	58.5	53.6	48.6	40.3
	Random Truncation	38.4	37.9	36.5	34.1	28.3	65.3	64.4	61.4	56.9	45.1	94.9 ^{ns}	94.6	94.3	93.4	89.3	60.3	59.4	56.6	50.3	39.4
	Random Projection	38.3	37.8	36.7	34.4	29.1	65.3	64.4	62.5	57.3	47.6	94.8	94.8	94.4	93.5	91.1	60.6	59.7	56.7	52.1	41.7
	PCA	38.8	38.3	36.9	34.7	29.9	67.1 ^{ns}	66.4	64.3	59.7	50.4	95.2 ^{ns}	95.1 ^{ns}	94.7	93.8	90.6	62.3 ^{ns}	61.6 ^{ns}	58.8	55.9	48.7
	Whitening	37.2	37.4	36.6	35.0	31.0	64.4	65.2	64.5	61.1	52.9	95.5	95.3	95.0 ^{ns}	94.3	92.4	58.5	59.6	59.3	57.0	51.3
	γ -Whitening	38.2	38.3	37.0	35.2	30.7	66.4	66.6 ^{ns}	65.1	61.3	52.4	95.5	95.4	95.0^{ns}	94.3	92.0	62.1 ^{ns}	62.0 ^{ns}	60.5	57.1	50.7
	SpecTemp	38.9	38.4	37.0	35.1	31.0	67.2	66.8^{ns}	65.1	61.3	52.9	95.3	95.3 ^{ns}	95.0	94.3	92.4	62.5	62.3^{ns}	60.4	57.4	51.3

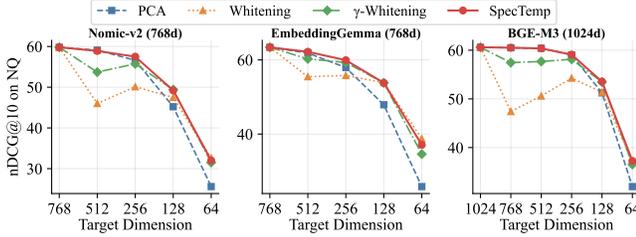


Figure 2: Performance consistency across additional models.

performs well at high dimensions but degrades under aggressive compression, while Whitening shows the opposite pattern; the fixed γ -Whitening offers a compromise but cannot adapt across compression regimes. SpecTemp automatically adjusts its tempering exponent and consistently matches or outperforms all fixed- γ alternatives. Prefix Truncation is competitive on FEVER for MRL-trained models but falls behind on non-MRL models and on tasks with complex query semantics (e.g., FiQA), as it is restricted to the first k training-time coordinates. Spectral methods—and SpecTemp in particular—consistently lead in these settings by projecting onto corpus-adaptive eigenvectors with richer expressivity.

4.2.2 Consistency across Retrieval Models. To verify that our findings generalize beyond the main evaluation, we test on three additional models on the NQ dataset: Nomic-v2, EmbeddingGemma, and BGE-M3 (Figure 2). We compare spectral methods only, as they share the same eigendecomposition backbone and isolate the effect of tempering strategy. The results reveal a consistent pattern: PCA degrades sharply at low dimensions where its skewed energy distribution fails to preserve fine distinctions, while Whitening suffers

Table 3: Comparison between oracle $\gamma^*(k)$ obtained via grid search and the theoretically predicted $\gamma(k)$ on NQ by GTE-7B.

Target Dimension $k \rightarrow$	768	512	256	128	64
Oracle γ^*_{grid} (Empirical)	0.15	0.25	0.45	0.55	0.95
Predicted $\gamma(k)$ (SpecTemp)	0.15	0.24	0.49	0.96	1.00
$ \Delta $ nDCG@10 (0–100 scale)	0.02	0.01	0.06	0.11	0.05

at high dimensions where it amplifies spectral noise. SpecTemp remains on the Pareto frontier across all dimensions, confirming that the adaptive $\gamma(k)$ mechanism robustly balances signal preservation and noise suppression across diverse architectures and scales.

4.2.3 Alignment with Empirical Optima. To validate that our predicted $\gamma(k)$ tracks the true optimum, we perform a grid search over $\gamma \in \{0, 0.05, \dots, 1.0\}$ on GTE-7B (NQ), selecting the best-performing γ at each target dimension. As shown in Table 3, the predicted $\gamma(k)$ closely matches the oracle at most dimensions. At $k=128$, despite the divergence in parameter space (0.55 vs. 0.96), the resulting performance penalty is minimal ($|\Delta|$ nDCG@10 = 0.11 points on a 0–100 scale). This indicates a flat optimization landscape where SpecTemp successfully locates a robust operating point within the near-optimal basin, achieving near-oracle performance without expensive validation, with an average $|\Delta|$ of just 0.05 points.

4.2.4 Sensitivity Analysis of \mathcal{T} . We test sensitivity to the tail set \mathcal{T} in Eq. 3 by varying its size from 5% to 20%. On GTE-7B \rightarrow NQ, nDCG@10 varies by at most 0.03 on a 0–100 scale. Given this robustness, we set \mathcal{T} to the last 10% of eigenvalue indices for all experiments without per-task tuning. This confirms that the noise floor estimate is stable across percentiles and requires no calibration.

5 Conclusion

We proposed **SpecTemp**, a learning-free post-hoc compression method for dense retrieval embeddings. By deriving a dimensionality-adaptive tempering exponent $\gamma(k)$ from the local SNR profile of the eigenspectrum, our method effectively bridges the trade-off between variance preservation (PCA) and isotropy (Whitening). Extensive experiments across six diverse models show that SpecTemp closely matches grid-searched oracle $\gamma^*(k)$ performance without any hyperparameter tuning. We hope this work serves as a practical baseline for learning-free embedding compression.

References

- [1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).
- [2] Jianyu 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*. Association for Computational Linguistics, Bangkok, Thailand, 2318–2335. <https://doi.org/10.18653/v1/2024.findings-acl.137>
- [3] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/V1/N19-1423>
- [4] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). [arXiv:2401.08281](https://arxiv.org/abs/2401.08281) [cs.LG]
- [5] Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Esлами, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval. [arXiv:2506.18902](https://arxiv.org/abs/2506.18902) [cs.AI] <https://arxiv.org/abs/2506.18902>
- [6] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 113–122. <https://doi.org/10.1145/3404835.3462891>
- [7] Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An Easy Unsupervised Sentence Embedding Approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 238–244. <https://aclanthology.org/2021.findings-emnlp.23/>
- [8] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1 (2011), 117–128. <https://doi.org/10.1109/TPAMI.2010.57>
- [9] William B Johnson, Joram Lindenstrauss, et al. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics* 26, 189–206 (1984), 1. <https://api.semanticscholar.org/CorpusID:117819162>
- [10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020*. Association for Computational Linguistics, 6769–6781. <https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550>
- [11] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham M. Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html
- [12] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/TACL_A_00276
- [13] 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)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 9119–9130. <https://aclanthology.org/2020.emnlp-main.733/>
- [14] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. [arXiv:2308.03281](https://arxiv.org/abs/2308.03281) [cs.CL] <https://arxiv.org/abs/2308.03281>
- [15] Vasileios Lioutas, Ahmad Rashid, Krtin Kumar, Md. Akmal Haidar, and Mehdi Rezagholizadeh. 2020. Improving Word Embedding Factorization for Compression Using Distilled Nonlinear Neural Decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2774–2784. <https://doi.org/10.18653/v1/2020.findings-emnlp.250>
- [16] Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. Dimension Reduction for Efficient Dense Retrieval via Conditional Autoencoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5692–5698. <https://doi.org/10.18653/v1/2022.emnlp-main.384>
- [17] Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. 2025. DIVER: A Multi-Stage Approach for Reasoning-intensive Information Retrieval. [arXiv:2508.07995](https://arxiv.org/abs/2508.07995) [cs.IR] <https://arxiv.org/abs/2508.07995>
- [18] Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. 2021. Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2854–2859. <https://aclanthology.org/2021.emnlp-main.227/>
- [19] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 2421–2425. <https://doi.org/10.1145/3626772.3657951>
- [20] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. In *Companion of The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 1941–1942. <https://doi.org/10.1145/3184558.3192301>
- [21] Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HkuGJ3kCb>
- [22] Zach Nussbaum and Brandon Duderstadt. 2025. Training Sparse Mixture Of Experts Text Embedding Models. [arXiv:2502.07972](https://arxiv.org/abs/2502.07972) [cs.CL] <https://arxiv.org/abs/2502.07972>
- [23] Sara Rajaei and Mohammad Taher Pilehvar. 2021. A Cluster-based Approach for Improving Isotropy in Contextual Embedding Space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 575–584. <https://aclanthology.org/2021.acl-short.73/>
- [24] Vikas Raunak, Vivek Gupta, and Florian Metzke. 2019. Effective Dimensionality Reduction for Word Embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (Repl4NLP-2019)*, Isabelle Augenstein, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei (Eds.). Association for Computational Linguistics, Florence, Italy, 235–243. <https://aclanthology.org/W19-4328/>
- [25] 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. <https://doi.org/10.18653/V1/D19-1410>
- [26] Ville Satopaa, Jeannie R. Albrecht, David E. Irwin, and Barath Raghavan. 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *31st IEEE International Conference on Distributed Computing Systems Workshops (ICDCS 2011 Workshops)*, 20-24 June 2011, Minneapolis, Minnesota, USA. IEEE Computer Society, 166–171. <https://doi.org/10.1109/ICDCSW.2011.20>
- [27] Jianlin Su. 2022. When BERT Whitening Introduces Hyperparameters: There Is Always One That Suits You. <https://kexue.fm/archives/9079> Chinese blog post.

- [28] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whiten- ing Sentence Representations for Better Semantics and Faster Retrieval. arXiv:2103.15316 [cs.CL] <https://arxiv.org/abs/2103.15316>
- [29] 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*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 27705–27726. <https://doi.org/10.18653/v1/2025.emnlp-main.1410>
- [30] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 809–819. <https://doi.org/10.18653/V1/N18-1074>
- [31] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesh Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. arXiv:2509.20354 [cs.CL] <https://arxiv.org/abs/2509.20354>
- [32] 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. *CoRR* abs/2212.03533 (2022). <https://doi.org/10.48550/ARXIV.2212.03533> arXiv:2212.03533
- [33] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. <https://openreview.net/forum?id=zeFrfgYzIn>
- [34] 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)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 6530–6543. <https://aclanthology.org/2024.lrec-main.579/>
- [35] Gaifan Zhang, Yi Zhou, and Danushka Bollegala. 2026. CASE – Condition-Aware Sentence Embeddings for Conditional Semantic Textual Similarity Measurement. arXiv:2503.17279 [cs.CL] <https://arxiv.org/abs/2503.17279>
- [36] 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. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176 [cs.CL] <https://arxiv.org/abs/2506.05176>
- [37] Chunsheng Zuo and Daniel Khashabi. 2026. More Than Efficiency: Embedding Compression Improves Domain Adaptation in Dense Retrieval. arXiv:2601.13525 [cs.IR] <https://arxiv.org/abs/2601.13525>