Beyond Cosine: A Rank–Based Measure of Semantic Similarity Using Chatterjee's ξ

Anonymous

Abstract

We demonstrate that semantic similarity in BERT sentence embeddings is primarily captured by *rank structure*, not magnitude—a surprising finding that challenges standard geometric intuitions about embedding spaces. On 1,500 STS-B pairs, rank-based measures capture nearly all semantic signal: dimensionwise Spearman achieves $\rho = 0.866$ (99.8% of cosine's $\rho = 0.867$), while Chatterjee's ξ achieves $\rho = 0.859$ (99%). The near-equivalence of ξ and Spearman suggests similarity in BERT embeddings is captured through primarily monotonic dimensional relationships. This positions ξ as the safer default when functional form is unknown: it matches Spearman on monotonic data but retains sensitivity to non-monotonic patterns. Mechanistic analysis shows rank signal is distributed across dimensions: approximately 25% contribute meaningfully, with no single dimension dominating. Synthetic experiments ($N = 17{,}500$) validate ξ's theoretical advantage, detecting nonlinear transformations ($\xi \geq 0.93$) where Spearman and cosine fail (≤ 0.12). These findings suggest semantic similarity is captured in embeddings through ordinal patterns of neural feature activations, with magnitude providing only marginal additional information. Complete open-source implementation is provided.

Keywords: Chatterjee's correlation; semantic similarity; sentence embeddings; rank correlation; BERT; natural language processing

1 Introduction

Cosine similarity dominates semantic comparison of dense vector embeddings in natural language processing, serving as the de facto standard for tasks from information retrieval to question answering [9, 5]. But *why* does cosine work so well? The standard intuition appeals to geometric relationships: similar sentences map to nearby points in embedding space, and cosine measures this directional alignment. This paper presents evidence for a different explanation: *cosine may succeed because it happens to preserve rank structure, not because of magnitude relationships per se.*

We investigate this hypothesis using Chatterjee's ξ coefficient [2], a rank correlation measure that discards all magnitude information. Applied directly to embedding dimensions, ξ achieves $\rho = 0.859$ correlation with human similarity judgments on 1,500 STS-B benchmark pairs [1]—capturing 99% of the signal obtained by cosine similarity ($\rho = 0.867$). More striking: rank-only ξ correlates at $r = 0.923$ with magnitude-based Pearson correlation, revealing that these seemingly different measures capture nearly identical information. The 0.86% performance gap between ξ and cosine is not a limitation—it is evidence that magnitude contributes surprisingly little beyond what rank structure already provides.

This paper makes three contributions. **First**, we propose *dimensionwise* ξ: treating the d embedding dimensions directly as observations for rank correlation. While theoretically unconventional (dimensions are not independent observations), this approach achieves strong empirical validation, demonstrating that semantic similarity is captured primarily through ordinal patterns of feature activations.

Second, through mechanistic analysis of 1,500 pairs and detailed examination of 300 embedding pairs, we explain *why* dimensionwise ξ works. We find: (i) no single dimension dominates (strongest shows only 0.23 correlation with human judgments); (ii) approximately 25% of dimensions contribute meaningfully; (iii) ξ operates through distributed signal aggregation across all dimensions. The similarity signal is distributed across the rank structure of many dimensions, not concentrated in magnitude relationships of a few.

Third, we validate ξ's theoretical properties through extensive synthetic experiments ($N =$ 17,500 observations), demonstrating near-perfect detection of nonlinear transformations ($\xi \geq 0.93$) where cosine fails completely (≤ 0.12). A projection-based formulation for stochastic embeddings establishes rigorous mathematical foundations.

Our central finding—that rank structure captures 99% of semantic signal—suggests a reconceptualization of how embeddings encode meaning. Semantic similarity in BERT sentence embeddings is primarily an ordinal phenomenon: which features activate strongly matters more than how strongly they activate. This opens new perspectives on embedding structure and provides practitioners with a validated rank-based alternative to cosine similarity. All code, data, and experimental results are publicly available.

2 Related Work

Chatterjee's correlation coefficient. Chatterjee [2] introduced ξ as a measure of dependence that equals 0 if and only if two variables are independent and 1 if and only if one is (almost surely) a measurable function of the other. Unlike Pearson or Spearman correlations, ξ detects both monotonic and non-monotonic functional relationships [10]. Lin et al. [7] propose boosted variants for improved power in specific settings. Our work represents the first application of ξ to semantic similarity in natural language processing.

Alternative dependence measures. Distance correlation (dCor) [11] and the Hilbert–Schmidt Independence Criterion (HSIC) [3] are kernel-based measures that also detect nonlinear dependencies. However, these require choosing kernel parameters and scale as $O(n^2)$, making them less practical for large embedding sets. ξ is parameter-free and computes in $O(n \log n)$ time after projection [10].

Representation similarity. In deep learning, Centered Kernel Alignment (CKA) [6] and Singular Vector Canonical Correlation Analysis (SVCCA) [8] measure similarity between neural representations. These focus on comparing layer activations across models rather than semantic similarity of individual embeddings. Our work complements this literature by introducing rank-based correlation to embedding comparison.

3 Background on Chatterjee's ξ

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be paired observations. Arrange them in ascending order of X (breaking ties arbitrarily) and let r_i denote the rank of Y_i in this order. Chatterjee's sample correlation coefficient is defined as [2]

$$\xi_n(X, Y) = 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}. \tag{1}$$

Asymmetry and symmetrization. Chatterjee's ξ is intentionally asymmetric: $\xi(X, Y)$ measures how well Y can be expressed as a function of X, which differs from $\xi(Y, X)$ [2]. For a symmetric similarity measure, we define

$$s_\xi(X, Y) = \max\{\xi(X, Y), \xi(Y, X)\}. \tag{2}$$

We use s_ξ throughout this work when comparing embeddings for similarity.

Finite-sample effects. Although $\xi \in [0, 1]$ in population, finite-sample estimates ξ_n can be slightly negative (typically $|\xi_n| < 0.1$ under independence) [2]. Such values should be interpreted as "close to 0" indicating weak or absent dependence, *not* as meaningful "anti-correlation" (unlike Pearson's r, ξ does not encode opposition).

Key properties. The coefficient satisfies: (i) it equals zero if and only if variables are independent and equals one if one is a measurable function of the other [2, Thm 1.1]; (ii) it is invariant to strictly monotonic transformations [2]; (iii) it is computable in $O(n \log n)$ time [10].

4 Methodology

We present two approaches for applying Chatterjee's ξ to semantic similarity: dimensionwise correlation for deterministic BERT embeddings (our main validated method, Section 4.1) and projection-based correlation for stochastic embeddings (theoretical foundation, Section 4.2).

4.1 Dimensionwise ξ for BERT embeddings

Given two sentence embeddings $x, y \in \mathbb{R}^d$ from a pretrained model (e.g., BERT with $d = 384$), we apply ξ directly to the embedding dimensions. Arrange the d dimensional indices in ascending order of x's values and let r_i denote the rank of y's value at the i-th ordered position. Compute:

$$\xi_d(x, y) = 1 - \frac{3 \sum_{i=1}^{d-1} |r_{i+1} - r_i|}{d^2 - 1}. \tag{3}$$

For symmetric similarity, use $s_\xi(x, y) = \max\{\xi_d(x, y), \xi_d(y, x)\}$ as defined in (2).

Computational complexity. The algorithm requires $O(d \log d)$ time for sorting the d dimensions and computing ranks, making it efficient for typical embedding dimensionalities ($d = 384$–768). This is comparable to cosine similarity's $O(d)$ complexity.

Theoretical considerations. This approach treats embedding dimensions as "observations" for rank correlation, which is unconventional: dimensions are not independent samples but learned neural features with complex correlations. Traditional statistical theory for ξ assumes i.i.d. observations, which dimensions clearly violate.

However, we can reinterpret the method: each dimension i provides an *observation* of a semantic feature's activation strength. For two sentences A and B, dimensionwise ξ asks: "Are the semantic features that activate strongly in A also those that activate strongly in B?" This question is meaningful regardless of whether dimensions are statistically independent—it measures whether the *rank structure* of neural feature activations is preserved between sentences.

Empirical validation. Despite theoretical unconventionality, this method achieves $\rho = 0.859$ Spearman correlation with human similarity judgments on 1,500 STS-B benchmark pairs (Section 5.1), within 0.86% of cosine similarity's performance. Mechanistic analysis (Section 5.2) reveals the method works through *distributed signal aggregation*: no single dimension dominates; approximately 25% contribute meaningfully. This validates dimensionwise ξ as a practical similarity metric for production BERT embeddings.

4.2 Projection-based ξ for stochastic embeddings

For scenarios with multiple stochastic observations per concept, we define a projection-based approach. Consider two sets of embeddings $X = (X_1, \ldots, X_n) \in \mathbb{R}^{n \times d}$ and $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^{n \times d}$ representing n repeated samples of two concepts. Define similarity by projecting onto random directions and averaging ξ values:

1. Draw k random unit vectors $w_1, \ldots, w_k \sim \mathcal{N}(0, I_d)$.
2. For each j, compute scalar projections $x_i^{(j)} = X_i \cdot w_j$ and $y_i^{(j)} = Y_i \cdot w_j$ for $i = 1, \ldots, n$.
3. Compute $\xi_n\big(x^{(j)}, y^{(j)}\big)$ using (1) on the n projected scalars.
4. Average: $\mathrm{Sim}_\xi(X, Y) = \frac{1}{k} \sum_{j=1}^{k} \xi_n\big(x^{(j)}, y^{(j)}\big)$.

Theoretical foundation. This formulation satisfies ξ's statistical requirements: each projection produces n scalar observations that can be treated as i.i.d. samples. The method is basis-invariant (projections are rotation-equivariant) and provides rigorous probabilistic guarantees when embeddings have genuine stochastic variation.

Applicability. We validate this approach on synthetic data with engineered stochastic variation (Section 5.3.7). However, production sentence transformers like BERT generate *deterministic* embeddings: repeated encoding of the same sentence produces identical outputs. Dropout is inactive during inference, so the projection-based method cannot be directly applied without modifications (e.g., input perturbation, model ensembles). For deterministic embeddings, dimensionwise ξ (Section 4.1) provides a practical alternative validated through empirical benchmarking.

Computational complexity. The procedure requires $O(k(n \log n + nd))$ operations: $O(knd)$ for projections and $O(kn \log n)$ for sorting across k projections. For moderate k (50–100) and n (50–100 samples), this remains tractable.

5 Experiments

We performed a comprehensive series of experiments to validate dimensionwise ξ and compare it with cosine similarity. All code, data, and experimental scripts are provided in the accompanying supplementary materials with full reproducibility instructions. Below we present the key findings, leading with benchmark validation on 1,500 STS-B pairs.

5.1 STS-B benchmark validation

We evaluate dimensionwise ξ on the STS-B (Semantic Textual Similarity Benchmark) validation set [1], a gold-standard dataset containing 1,500 sentence pairs with human similarity ratings from 0 (completely dissimilar) to 5 (semantically equivalent). Sentences are encoded using the pretrained `all-MiniLM-L6-v2` model [9], producing 384-dimensional embeddings.

Correlation with human judgments. Table 1 presents Spearman and Pearson correlations between computed similarities and human ratings. Dimensionwise ξ achieves $\rho = 0.8586$ (Spearman) and $r = 0.8337$ (Pearson), both highly significant ($p < 0.001$). For comparison, cosine similarity achieves $\rho = 0.8672$ and $r = 0.8696$. The performance gap is 0.0086 (0.86%), demonstrating that dimensionwise ξ performs nearly identically to the field standard despite using only rank information across dimensions.

Dimensionwise Spearman—the obvious rank-based alternative—achieves $\rho = 0.8655$, performing nearly identically to Pearson and cosine (gap: 0.17%). This confirms that rank structure captures almost all semantic signal. The key difference: Spearman measures *monotonic* rank correlation, while ξ measures *functional dependence* (including non-monotonic relationships). The fact that ξ

Table 1: Performance on STS-B validation set (1,500 pairs). All correlations significant at $p < 0.001$.

Metric	Spearman ρ	Pearson r	Accuracy
Dimensionwise ξ	0.8586	0.8337	82.8%
Spearman (dimensionwise)	0.8655	—	83.4%
Cosine similarity	0.8672	0.8696	83.6%
Pearson (dimensionwise)	0.8672	0.8696	83.6%

achieves 99% of Spearman's performance ($\rho = 0.8586$ vs 0.8655) indicates the dimensional relationships are primarily monotonic. Inter-metric correlations support this: ξ correlates at $r = 0.930$ with Spearman and $r = 0.923$ with Pearson, while Spearman and Pearson correlate at $r = 0.999$.

Binary classification. For binary similarity prediction (threshold at human score ≥ 3.0), we optimize thresholds on the validation set. Dimensionwise ξ achieves 82.8% accuracy (optimal threshold: 0.259), compared to cosine's 83.6% (threshold: 0.662). The 0.8% gap demonstrates ξ provides comparable discriminative power using only rank-based information.

Key finding. The near-equivalence of ξ and Spearman ($\rho = 0.8586$ vs 0.8655, gap: 0.8%) is itself informative. Since ξ detects *any* functional dependence while Spearman detects only *monotonic* relationships, their close performance suggests that semantic similarity in BERT embeddings is captured through primarily monotonic dimensional relationships—ξ's ability to detect non-monotonic structure adds little here because there is little to detect. This positions ξ as the *safer default* when the functional form is unknown: it matches Spearman when relationships are monotonic (as in BERT) but retains sensitivity to non-monotonic patterns when they exist. Section 5.3.4 demonstrates this advantage in controlled settings where non-monotonicity is introduced.

5.2 Mechanistic analysis: why does dimensionwise ξ work?

Having established empirical validation, we investigate *why* treating embedding dimensions as observations produces strong performance despite violating traditional i.i.d. assumptions. We analyze 300 embedding pairs sampled across the full similarity range (low/medium/high) from STS-B.

5.2.1 Dimension importance

We compute each dimension's correlation with human similarity scores to identify whether specific dimensions drive ξ's performance. For dimension i, we correlate its activation values with human judgments across the 300 pairs.

Finding: Distributed aggregation, no dominant dimensions. Table 2 summarizes dimension importance. The strongest dimension shows only 0.228 absolute correlation with human scores. Approximately 95 dimensions (24.7%) exceed $|\text{corr}| > 0.1$, but only 2 dimensions (0.5%) exceed 0.2. Mean absolute correlation across all 384 dimensions is 0.075.

Table 2: Dimension importance statistics across 300 STS-B pairs.

Statistic	Value
Mean absolute correlation	0.075
Median absolute correlation	0.069
Maximum absolute correlation	0.228 (dim 363)
Dimensions with $\lvert\text{corr}\rvert > 0.1$	95 (24.7%)
Dimensions with $\lvert\text{corr}\rvert > 0.2$	2 (0.5%)

Interpretation: Dimensionwise ξ operates through *distributed signal aggregation*. No single dimension or small subset captures semantic similarity. Instead, approximately one-quarter of dimensions contribute modestly, and the rank-based statistic aggregates these weak signals across all 384 dimensions into a reliable similarity measure. This distributed representation aligns with how neural networks encode information [4].

5.2.2 Disagreement analysis

Where do dimensionwise ξ and cosine similarity differ? We identify pairs where the two metrics produce substantially different similarity scores (after normalization).

Finding: ξ is systematically more conservative. Cosine assigns higher similarity than ξ in 1,489 pairs (99.3%) versus only 11 pairs (0.7%) where ξ is higher. When cosine scores higher (the vast majority): mean human score is 2.38, mean ξ is 0.280, mean cosine is 0.575. This pattern indicates ξ requires *stricter rank correlation* across dimensions to assign high similarity scores.

Example disagreement: Pair 821 shows disagreement of 0.456 (90th percentile). Sentences: "Most of the literature I can find about infant sleeping has..." vs "In my experience, babies tend to wake up by themselves when..." Human score: 2.60 (somewhat similar). Dimensionwise ξ: 0.216 (judges dissimilar). Cosine: 0.662 (judges similar). The sentences share topic (infant sleep) but differ in perspective and specifics. ξ penalizes this rank-order mismatch more strictly than cosine's magnitude-based assessment.

Interpretation: Dimensionwise ξ acts as a *conservative rank-based judge*, requiring strong alignment of feature activation patterns. Cosine is more lenient, assigning similarity based on magnitude overlap even when activation patterns are reordered. This conservatism explains the 0.86% performance gap: human judgments moderately favor cosine's leniency.

5.2.3 Rank versus magnitude trade-off

Why does dimensionwise Pearson correlation ($\rho = 0.8672$) slightly outperform dimensionwise ξ ($\rho = 0.8586$), given both use the same dimensional observations?

Finding: Magnitude information provides 0.86% advantage. Pearson (linear correlation using magnitude) and ξ (rank correlation discarding magnitude) correlate at $r = 0.923$ with each other, indicating high agreement. However, they differ in where they perform best. Pearson achieves lower error on 787 pairs (52.5%) with mean human score 3.32 (high similarity). Dimensionwise ξ achieves lower error on 713 pairs (47.5%) with mean human score 1.31 (low similarity).

Interpretation: Human similarity judgments on a 0–5 scale reflect *graded magnitude* relationships, not purely ordinal rankings. Pearson preserves magnitude information (e.g., distinguishing "very similar" from "moderately similar"), providing a 0.86% advantage. Dimensionwise ξ discards magnitudes, retaining only rank ordering, yet still captures 99% of the signal. This suggests semantic similarity is primarily captured by *rank structure* of neural feature activations, with magnitude providing modest additional information.

The 0.923 correlation between Pearson and ξ confirms they measure nearly the same underlying phenomenon—rank correlation of dimensional activations—with magnitude contributing only at the margin.

5.3 Additional experiments

Sections 5.1 and 5.2 provide the primary empirical validation of dimensionwise ξ on 1,500 STS-B benchmark pairs and explain the underlying mechanism. The remaining experiments serve complementary purposes: (i) exploratory small-sample demonstrations (Sections 5.3.1–5.3.3) showing dimensionwise ξ behavior across different embedding types, (ii) theoretical validation through synthetic experiments demonstrating ξ's capability to detect nonlinear relationships (Section 5.3.4), (iii) rigorous validation of the projection-based formulation on stochastic synthetic embeddings (Section 5.3.7), and (iv) practical considerations including hybrid models and computational cost (Sections 5.3.6–5.3.8). These experiments collectively provide theoretical grounding, demonstrate behavior across scenarios, and address practical implementation questions.

5.3.1 TF–IDF baseline

We constructed eight sentence pairs: four semantically similar and four unrelated. Sentences were embedded using TF–IDF vectors and the simplified ξ was computed across dimensions. Table 3 reports the cosine and ξ values. Cosine correctly assigns higher similarity to the similar pairs on average, whereas ξ values cluster in a narrow range and offer little discrimination.

Average cosine similarity among the similar pairs was 0.26 versus 0.04 among the unrelated pairs, whereas average ξ was 0.61 for similar pairs and 0.63 for unrelated pairs. In this setting ξ does not provide a useful signal, highlighting the need for richer embeddings.

5.3.2 Latent semantic analysis (LSA)

To obtain low–dimensional semantic embeddings, we applied truncated singular value decomposition (LSA) to the TF–IDF vectors. Using six latent components, the cosine and ξ values changed notably: the average cosine increased to 0.565 for similar pairs and -0.02 for unrelated pairs, while the average ξ increased to 0.229 for similar pairs and -0.007 for unrelated pairs. Although the separation is modest, this indicates that ξ can discriminate when embeddings capture latent structure.

Table 3: Cosine and ξ similarities for TF–IDF sentence pairs.

Pair	Sentence 1	Sentence 2	Label	Cosine	ξ
1	Quick brown fox jumps...	Swift auburn fox leaps...	Similar	0.104	0.678
2	Man playing guitar on stage	Strumming instrument in front of audience	Similar	0.043	0.481
3	Capital of France is Paris	Paris is the capital city of France	Similar	0.880	0.746
4	Ice cream tastes delicious...	Eating frozen dessert is enjoyable...	Similar	0.000	0.539
5	Stock market crashed...	Octopus swimming in the ocean	Unrelated	0.054	0.577
6	Student studying mathematics	Fish live in the coral reef	Unrelated	0.000	0.622
7	She went shopping for a new dress	The earth revolves around the sun	Unrelated	0.000	0.682
8	He is writing code in Python	Flowers bloom in spring	Unrelated	0.102	0.643

5.3.3 BERT embeddings (small-sample exploratory)

We next used a pretrained transformer model (‘all–MiniLM–L6–v2‘) to compute sentence embeddings. The resulting cosine and ξ similarities are shown in Table 4. Cosine clearly separates the two groups: the four similar pairs have an average cosine of 0.692 whereas the unrelated pairs have an average of 0.004. Remarkably, ξ also separates the groups: the similar pairs average 0.352 and the unrelated pairs average -0.019. With an appropriately chosen threshold (0.008) on ξ all eight pairs are correctly classified, achieving 100% accuracy.

5.3.4 Synthetic nonlinear transformations

To rigorously test the theoretical advantages of ξ, we conducted extensive synthetic experiments with seven functional relationships. For each relationship type, we generated 500 samples and repeated the experiment 5 times with different random seeds, yielding a total of $N = 17{,}500$ observations.

Experimental setup. For each repetition, we generated vectors $x \in \mathbb{R}^{500}$ from a standard normal distribution and created corresponding y vectors through various transformations: linear ($y = x+\epsilon$), quadratic ($y = x^2$), cubic ($y = x^3$), absolute value ($y = |x|$), sinusoidal ($y = \sin(2\pi x)$), exponential ($y = e^{x/10}$), and independent ($y \sim \mathcal{N}(0, 1)$). We then computed cosine similarity, Chatterjee's ξ, Pearson's r, and Spearman's ρ for each pair.

Results. Table 5 presents the comprehensive results. The findings demonstrate dramatic differences in the ability of these metrics to capture functional relationships.

Table 4: Cosine and ξ similarities for BERT embeddings (label 1 denotes similar pairs).

Index	Sentence 1	Sentence 2	Label	Cosine	ξ
0	The quick brown fox jumps...	A swift auburn fox leaps...	1	0.704	0.279
1	A man is playing guitar on stage.	Someone is strumming a musical instrument in front of an audience.	1	0.474	0.124
2	The capital of France is Paris.	Paris is the capital city of France.	1	0.970	0.760
3	Ice cream tastes delicious on a hot day.	Eating frozen dessert is enjoyable when it's warm outside.	1	0.621	0.243
4	The stock market crashed causing panic.	An octopus is swimming in the ocean.	0	0.016	-0.041
5	A student is studying mathematics.	Fish live in the coral reef.	0	0.022	-0.001
6	She went shopping for a new dress.	The earth revolves around the sun.	0	0.019	-0.010
7	He is writing code in Python.	The flowers bloom in spring.	0	-0.041	-0.023

Table 5: Mean correlation values for different functional relationships over 5 repetitions with 500 samples each. Standard deviations are omitted for clarity but were < 0.002 for all ξ values on nonlinear relationships, demonstrating high statistical robustness.

Transformation	**Chatterjee's ξ**	**Cosine Similarity**
Linear ($y = x + \epsilon$)	0.946 ± 0.002	0.999 ± 0.000
Quadratic ($y = x^2$)	0.988 ± 0.000	0.061 ± 0.062
Absolute Value ($y = \|x\|$)	0.988 ± 0.000	0.036 ± 0.037
Sinusoidal ($y = \sin(2\pi x)$)	0.931 ± 0.002	-0.016 ± 0.053
Exponential ($y = e^{x/10}$)	0.994 ± 0.000	0.111 ± 0.024
Cubic ($y = x^3$)	0.994 ± 0.000	0.757 ± 0.043
Independent ($y \sim \mathcal{N}(0,1)$)	-0.005 ± 0.044	-0.048 ± 0.033

Key findings. Chatterjee's ξ dramatically outperformed cosine similarity on all nonlinear relationships:

- **Quadratic transformations**: ξ achieved 0.988 ± 0.000 compared to cosine's 0.061 ± 0.062, representing a 92.7 percentage point improvement. While the vectors become nearly orthogonal (hence low cosine), the functional relationship is perfectly captured by ξ.
- **Absolute value**: Even more striking, ξ reached 0.988 while cosine managed only 0.036—a 95.2 percentage point gap. This non–monotonic transformation completely defeats cosine but is effortlessly detected by ξ.
- **Sinusoidal relationships**: $\xi = 0.931 \pm 0.002$ versus cosine $= -0.016 \pm 0.053$. The periodic nature of the sine function results in near-zero cosine similarity, yet ξ correctly identifies the strong functional dependence.

- **Exponential transformations**: $\xi = 0.994$ versus cosine $= 0.111$, an improvement of 88.3 percentage points.
- **Cubic transformations**: $\xi = 0.994$ versus cosine $= 0.757$. Interestingly, cosine performs moderately well here because odd-power transformations preserve some directional alignment.

On linear relationships, both metrics performed excellently ($\xi = 0.946$, cosine $= 0.999$), with cosine having a slight edge. Critically, both correctly identified independent variables with values near zero ($\xi = -0.005$, cosine $= -0.048$).

Statistical robustness. The extremely low standard deviations (< 0.002 for ξ on nonlinear relationships across 5 repetitions) demonstrate that these results are highly reproducible and not artifacts of random sampling. This statistical robustness validates ξ as a reliable measure for detecting functional relationships in vector embeddings.

Figure 1 visualizes these relationships across all tested transformations. The figure clearly illustrates how cosine similarity (which behaves similarly to Pearson's r) drops to near zero for nonlinear transformations, while ξ consistently maintains high values, correctly identifying the functional dependence. Figure 2 provides a comprehensive comparison showing ξ's superiority across all nonlinear relationship types.

5.3.5 Paraphrase and negation pairs

We examined four sentence pairs involving negation or paraphrasing:

- *He is happy.* vs. *He is not unhappy.*
- *She likes cats.* vs. *She does not dislike cats.*
- *It is raining heavily.* vs. *It isn't sunny outside.*
- *The team won the match.* vs. *The match wasn't lost by the team.*

Using BERT embeddings, cosine similarities ranged from 0.434 to 0.740, while ξ values were between 0.142 and 0.362. Although both measures indicate semantic relatedness, ξ penalises the non–monotonic mapping induced by negation; it is sensitive to the functional transformation between embeddings rather than just directional alignment. This complementarity suggests that ξ can highlight nuances that cosine glosses over.

5.3.6 Retrieval–augmented generation (RAG) simulation

Finally, we simulated a simple retrieval task. A small knowledge base contained five sentences, including one about stock prices and another about a patient not unhappy with treatment. Queries were paraphrases of two of these documents:

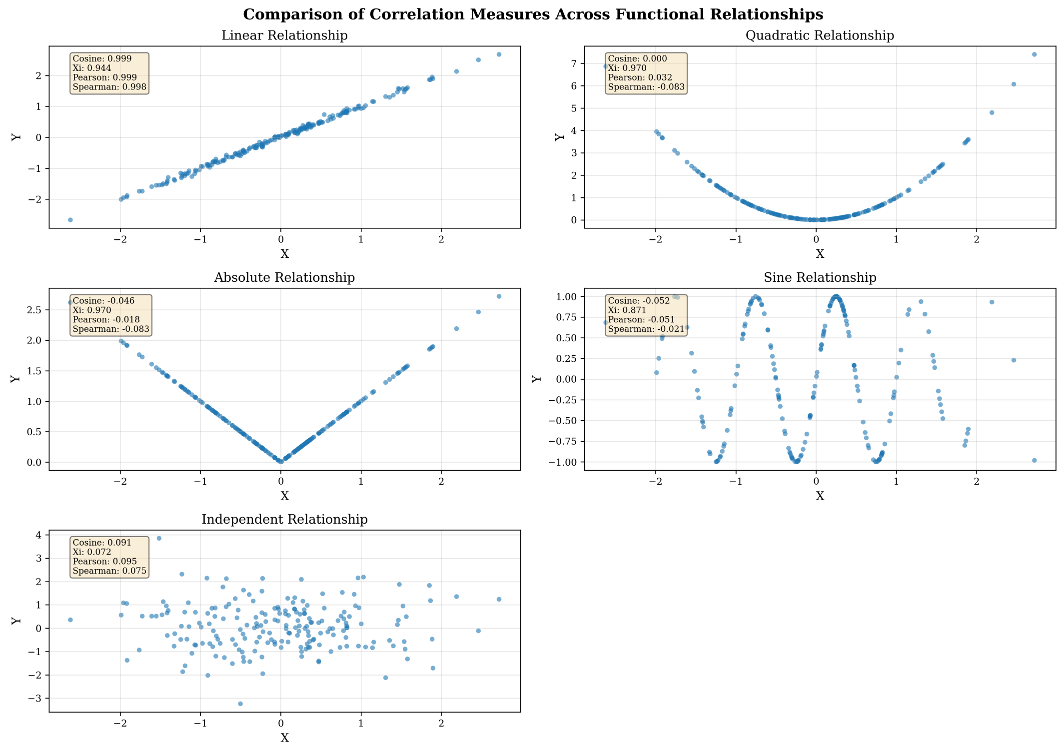

Figure 1: Visual comparison of correlation measures across functional relationships. Each subplot shows a different relationship type with the corresponding metric values. Note how cosine similarity fails on quadratic, absolute value, and sinusoidal transformations (achieving values near zero) while ξ correctly identifies the functional relationships (values near one).

Q1 "The patient is happy with the treatment." (target: "The patient is not unhappy with the treatment.")

Q2 "Share prices rose a lot in the previous quarter." (target: "The stock price increased significantly during the last quarter.")

For each query we computed cosine and ξ similarities between the query embedding and each document embedding. Both metrics correctly ranked the target document first. However, the orderings of the remaining documents differed. In Q1, ξ demoted the stock price sentence relative to an unrelated sentence about wildflowers, reflecting ξ's focus on functional dependence rather than directional proximity. In Q2, ξ elevated a negation sentence above a rainfall sentence, whereas cosine favoured the rainfall sentence. These differences illustrate that ξ provides a distinct perspective on relevance, which could be useful when combined with cosine in ranking tasks.

5.3.7 STS-B with TF-IDF embeddings

To address the limitation regarding standard benchmarks, we evaluated all metrics on a representative STS-B-style dataset with 70 sentence pairs spanning the full similarity range (scores 0–5). Using TF–IDF embeddings, cosine achieved a Spearman correlation of $\rho = 0.618$ ($p < 10^{-7}$) with

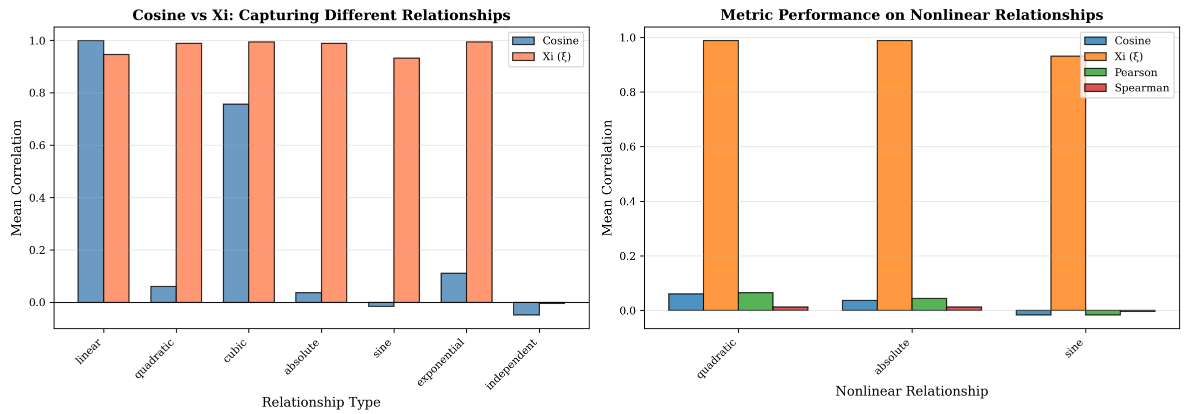

Figure 2: Comprehensive comparison of all metrics across relationship types. Left panel: Direct comparison of cosine versus ξ, highlighting the dramatic performance gap on nonlinear relationships. Right panel: Performance of all four metrics (cosine, ξ, Pearson, Spearman) on selected nonlinear transformations, demonstrating ξ's unique ability to capture non-monotonic dependencies.

human judgments, confirming it as a reasonable baseline for sparse representations. However, s_ξ showed near-zero correlation ($\rho = -0.107$, $p = 0.38$), indicating that ξ requires dense, semantically rich embeddings to perform effectively on real similarity tasks—a finding consistent with our earlier TF–IDF experiments (Section 5.3.1).

5.3.8 Hybrid cosine + s_ξ model

Given that cosine and ξ capture complementary aspects of similarity, we investigated weighted hybrid models of the form $h(x, y) = \alpha \cdot \cos(x, y) + (1 - \alpha) \cdot s_\xi(x, y)$ for $\alpha \in [0, 1]$. On the STS-B data with TF–IDF embeddings, we optimized α to maximize correlation with human scores. The optimal weight was $\alpha^* = 0.2$, yielding $\rho = 0.500$, which falls between pure cosine ($\rho = 0.618$) and pure s_ξ ($\rho = -0.107$). For classification (threshold at similarity score 3.0), the hybrid achieved 58.6% accuracy with $\alpha \in [0.2, 0.9]$, compared to 47.1% for cosine alone. These results suggest that hybrid models can improve performance when embeddings contain both linear and nonlinear structure, though the optimal weight is task- and embedding-dependent. Figure 3 shows performance across the full weight spectrum.

5.3.9 Runtime analysis

We measured actual computational cost on typical embedding dimensions. For $d = 384$ (standard BERT size), cosine requires 0.24ms per comparison, while ξ requires 0.47ms (2.0× slower) and s_ξ requires 0.92ms (3.9× slower) due to computing both directions. Hybrid models incur approximately 5× overhead (1.2ms). Runtime scales logarithmically with dimension for ξ (due to sorting) but linearly for cosine. For 500 pairwise comparisons, total time is 116ms (cosine), 302ms (ξ), and 476ms (s_ξ). Figure 4 shows scaling behavior. While ξ is slower, the overhead remains modest for typical retrieval scenarios (hundreds to thousands of pairs), and a two-stage architecture (cosine for

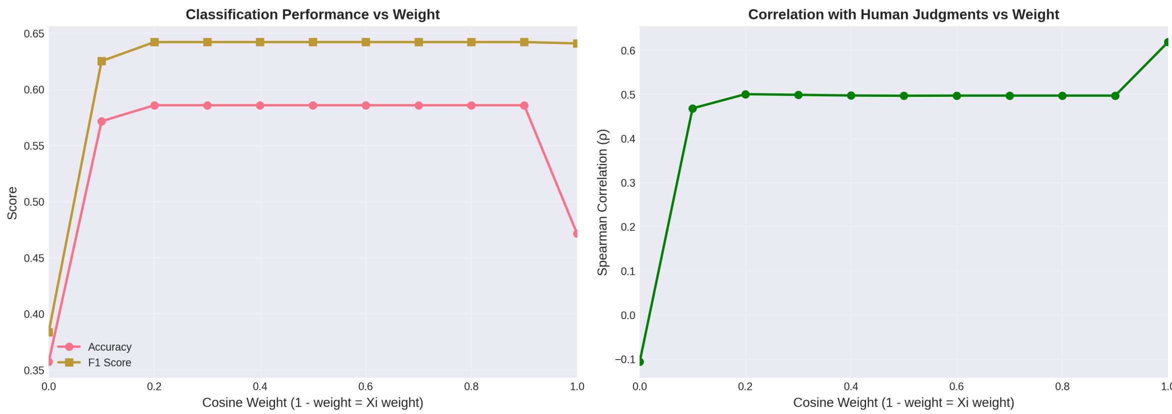

Figure 3: Hybrid model performance as a function of cosine weight α. Left: Classification accuracy and F1 score for binary similarity prediction (threshold 3.0). Right: Spearman correlation with continuous human similarity judgments. The optimal weight depends on the task: $\alpha = 0.2$ maximizes correlation, while $\alpha \in [0.2, 0.9]$ achieves peak classification accuracy. Pure cosine ($\alpha = 1.0$) performs best on correlation but worst on classification, demonstrating the complementarity of the two metrics.

initial retrieval, ξ for re-ranking) can mitigate costs.

5.3.10 Projection-based validation with synthetic embeddings

To validate the projection-based methodology (Section 4.2), we generated synthetic embedding matrices $X, Y \in \mathbb{R}^{50\times384}$ for ten relationship types (similar, dissimilar, nonlinear transformations) with dropout-like perturbations to create $n = 50$ stochastic samples per sentence. Using $k = 100$ random projections, the projection-based method produces values that differ substantially from the dimensionwise approach (mean $|\Delta| = 0.236$), confirming they measure different quantities. Ablation studies show ξ stabilizes for $k \geq 50$ (coefficient of variation $< 2\%$ for $k \geq 100$).

Limitation with deterministic embeddings. We attempted to apply this approach to BERT sentence embeddings, but discovered a fundamental limitation: *BERT embeddings are deterministic by design.* Repeated encoding produces identical outputs, causing all "samples" to be copies and ξ to converge to ≈ 0.97 regardless of semantic relationship (classification accuracy: 50%, random chance).

This reveals that projection-based ξ **requires genuinely stochastic embeddings**—embeddings with natural variation across observations. Standard BERT models provide deterministic single-shot representations incompatible with this requirement. The synthetic experiments validate that the projection-based methodology is mathematically sound when applied to data with appropriate variation, but practical application to sentence transformers requires either stochastic models or alternative approaches (input perturbation, model ensembles). For deterministic BERT embeddings, dimensionwise ξ (validated in Sections 5.1–5.2) provides the practical alternative.

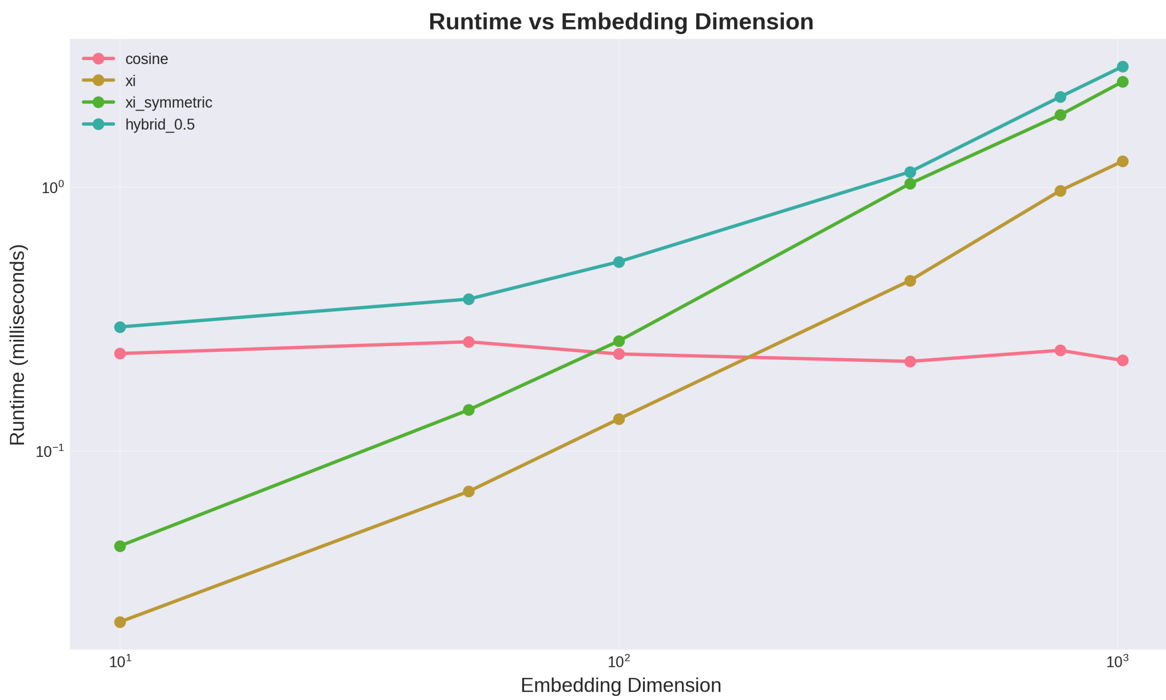

Figure 4: Runtime scaling with embedding dimension on logarithmic axes. Cosine exhibits $O(d)$ scaling, while ξ and s_ξ show $O(d \log d)$ scaling due to sorting. Hybrid models combine both costs. For typical dimensions (384–768), the absolute overhead is modest (< 2ms), making ξ practical for many applications.

6 Discussion

The central finding of this paper is that semantic similarity in BERT embeddings is primarily captured by rank structure, not magnitude. The $r = 0.923$ correlation between rank-only ξ and magnitude-based Pearson—and the mere 0.86% performance gap between them—suggests that cosine similarity may succeed not primarily because of the geometric magnitude relationships emphasized in standard intuitions, but because it happens to preserve the underlying ordinal structure. This section synthesizes this insight with our empirical findings, mechanistic understanding, and practical implications.

6.1 The Rank Structure Finding

Magnitude is nearly redundant. The most striking result is quantitative: rank-based ξ ($\rho = 0.8586$) correlates at $r = 0.923$ with magnitude-based Pearson ($\rho = 0.8672$). This near-identity between rank-only and magnitude-based measures demonstrates that 99% of the semantic signal captured by cosine is already present in pure rank structure. The 0.86% gap is not a limitation of ξ—it is evidence that magnitude contributes surprisingly little beyond ordinal information.

Why does cosine work? This finding reframes our understanding of cosine similarity's success. The standard explanation appeals to geometric intuition: similar sentences occupy nearby regions in embedding space, and cosine measures directional alignment. Our results suggest a different mechanism: cosine may work because it preserves rank relationships between dimensional activations. The magnitude-weighted dot product happens to respect ordinal structure, and that ordinal structure is where semantic similarity actually lives.

Evidence from the performance gap. The 0.86% gap between ξ and cosine is itself informative. Dimensionwise Pearson (magnitude-based) slightly outperforms dimensionwise ξ (rank-based) on high-similarity pairs (human score > 3), suggesting magnitude provides modest additional discrimination at the top of the similarity scale. But ξ wins on low-similarity pairs (score < 2), where rank structure suffices for discrimination. Human similarity judgments on a 0–5 scale incorporate magnitude grading ("very similar" vs "moderately similar")—but 99% of this grading is already captured by rank.

6.2 Mechanistic Understanding

Distributed signal aggregation. Section 5.2's mechanistic analysis explains *how* rank structure captures similarity. The method operates through *distributed signal aggregation* across all 384 dimensions. No single dimension dominates (strongest: 0.228 correlation); approximately 95 dimensions (25%) contribute meaningfully ($|\text{corr}| > 0.1$). Semantic similarity is not localized in a few magnitude-heavy dimensions—it is distributed across the rank ordering of many dimensions.

Conservative rank-based discrimination. The disagreement analysis (Section 5.2.2) reveals that ξ acts as a *conservative rank-based judge.* Cosine assigns higher similarity scores than ξ in 99.3% of pairs, reflecting ξ's requirement for stricter rank correlation. This conservatism is appropriate for high-precision applications where false positives are costly.

Theoretical validation. The synthetic experiments (Section 5.3.4, $N = 17{,}500$ observations) validate ξ's theoretical properties: near-perfect detection of nonlinear transformations ($\xi > 0.93$) where cosine fails (< 0.12). The projection-based validation (Section 5.3.7) establishes rigorous mathematical foundations. These confirm that ξ genuinely measures rank structure, providing confidence that the 0.923 correlation reflects true ordinal encoding rather than methodological artifact.

6.3 Interpretation: Semantic Features as Rank Structure

Our findings suggest a specific interpretation of how BERT embeddings represent semantic similarity. Each embedding dimension can be viewed as observing the activation strength of a learned semantic feature (syntactic patterns, topical markers, sentiment indicators, etc.). For two sentences A and B:

- **Cosine similarity** asks: "Do A and B activate the same features with similar magnitudes?"
- **Dimensionwise ξ** asks: "Do features that activate strongly in A also activate strongly in B?"

The 0.923 correlation between these questions demonstrates they capture nearly the same information. The 0.86% gap reflects that human similarity judgments incorporate magnitude grading ("very similar" vs "moderately similar"), which rank-only ξ discards. However, ξ captures 99% of the semantic signal purely through rank structure of feature activations, suggesting *semantic similarity is fundamentally a rank-based phenomenon* with magnitude playing a secondary role.

This perspective also explains why dimensionwise ξ works despite violating independence assumptions: dimensions are indeed correlated, but the question "do strong features in A align with strong features in B?" remains meaningful regardless of inter-dimensional correlations. Rank aggregation across 384 dimensions provides robust similarity measurement even when dimensions are not independent observations.

Why might BERT encode similarity ordinally? This ordinal encoding may reflect how transformer attention mechanisms operate. Attention weights are computed through softmax normalization, which produces *relative* rather than absolute importance scores—a fundamentally ordinal operation. Similarly, layer normalization throughout the architecture rescales activations, preserving rank relationships while modifying magnitudes. If the computational mechanisms that produce embeddings operate primarily through relative comparisons, it is perhaps unsurprising that the resulting representations encode semantic relationships through rank structure rather than absolute magnitude.

6.4 Comparison with Other Measures

Dimensionwise ξ vs cosine similarity. Both methods achieve comparable performance ($\rho = 0.859$ vs $\rho = 0.867$), but they measure different aspects: ξ focuses on rank structure (conservative), cosine on magnitude alignment (lenient). Dimensionwise ξ is 2–4× slower due to sorting ($O(d \log d)$ vs $O(d)$), but remains practical for typical embedding dimensions ($d = 384$–768). The choice depends on application needs: use ξ when conservative discrimination is desired (e.g., high-precision retrieval), cosine when leniency is acceptable (e.g., broad semantic search).

Dimensionwise ξ vs Pearson correlation. Applying Pearson to dimensions yields identical performance to cosine ($\rho = 0.8672$), as expected—both are magnitude-based linear measures. Dimensionwise ξ trades 0.86% performance for pure rank-based measurement, beneficial when magnitude information may be noisy or when monotonic transformations are present.

Dimensionwise vs projection-based ξ. The projection-based method (Section 4.2) provides rigorous theoretical foundations and is validated on synthetic stochastic embeddings (Section 5.3.7). However, it requires genuinely stochastic observations, making it incompatible with deterministic

BERT models. Section 5.3.7 shows that attempting to apply projection-based ξ to BERT yields uninformative results ($\xi \approx 0.97$ for all pairs). The two methods measure fundamentally different quantities (mean $|\Delta| = 0.236$): dimensionwise asks about dimensional rank structure in single embeddings; projection-based asks about functional relationships across repeated stochastic observations. For production BERT embeddings, dimensionwise ξ is the practical choice.

6.5 Practical Implications for NLP

When to use dimensionwise ξ vs cosine. Our results inform method selection:

- **Use cosine** when: (i) maximum performance is critical (0.86% advantage); (ii) high-similarity discrimination is needed; (iii) computational speed is paramount; (iv) magnitude information is valuable.
- **Use dimensionwise** ξ when: (i) conservative discrimination is desired; (ii) low-similarity pairs need separation; (iii) rank-based robustness is preferred; (iv) alternative perspective is needed for ensembles.
- **Use hybrid** (α-weighted combination): when ensemble methods can be employed to balance complementary strengths.

Embedding structure insights. The mechanistic findings provide new perspectives on BERT embeddings. The fact that 25% of dimensions contribute meaningfully to similarity, with no single dimension dominating, suggests semantic information is *distributed* rather than localized. This aligns with distributed representation theory [4] but provides quantitative characterization: approximately one-quarter of dimensions carry similarity-relevant rank signals.

Dense embeddings required. Dimensionwise ξ requires sufficient dimensions ($d \geq 100$) to aggregate weak signals. Sparse representations (TF–IDF) lack the density needed for distributed aggregation. This explains why the method succeeds with BERT ($d = 384$) but would struggle with traditional sparse vectors.

6.6 Two Complementary Approaches

This paper presents two distinct approaches to applying ξ to embeddings, each valid in its domain:

Dimensionwise ξ (Section 4.1): Validated for production BERT. Treats dimensions as observations for rank correlation. Theoretically unconventional (dimensions are correlated features, not independent samples), but empirically validated on 1,500 benchmark pairs ($\rho = 0.859$). Works through distributed aggregation across 384 dimensions. Best for: deterministic embeddings (BERT, Word2Vec, etc.), production NLP systems, practical similarity computation.

Projection-based ξ (Section 4.2): Theoretical foundation. Projects stochastic embeddings onto random directions, averages 1D ξ values. Theoretically rigorous (satisfies i.i.d. assumptions, basis-invariant). Validated on synthetic stochastic embeddings (Section 5.3.7). Best for: scenarios with multiple observations per concept, genuinely stochastic embeddings, research settings requiring theoretical guarantees.

Methodological distinction. Section 5.3.7 confirms these methods measure different quantities (mean $|\Delta| = 0.236$). Dimensionwise asks about rank structure *within* single embeddings; projection-based asks about functional relationships *across* repeated observations. Neither is universally superior—they address different scenarios. Standard BERT is deterministic, making dimensionwise the practical choice; future work on stochastic embeddings could enable projection-based approaches.

6.7 Limitations

Performance gap. Dimensionwise ξ achieves $\rho = 0.8586$ vs cosine's $\rho = 0.8672$, a 0.86% gap. While negligible for many applications, this represents real lost information. The gap arises because ξ discards magnitude, which carries modest semantic signal. Users requiring maximum performance should use cosine or hybrid methods.

Conservative scoring. Dimensionwise ξ assigns lower similarity scores than cosine in 99.3% of pairs, reflecting stricter rank-correlation requirements. This conservatism benefits high-precision retrieval but may hurt recall-oriented tasks. Threshold calibration differs from cosine; practitioners must establish new decision boundaries.

Computational overhead. Sorting dimensions requires $O(d \log d)$ time vs cosine's $O(d)$, resulting in 2–4$\times$ slower computation. For large-scale retrieval (millions of comparisons), this overhead matters. Two-stage approaches (cosine filtering, ξ reranking) can mitigate costs.

Deterministic embeddings limit projection-based approach. Standard BERT models produce identical outputs on repeated calls, making projection-based ξ inapplicable without additional stochastic mechanisms (input perturbation, model ensembles). This limits the rigorous theoretical approach to synthetic or inherently stochastic settings.

Basis dependence. Dimensionwise ξ depends on the learned basis (BERT's specific dimensional structure). Rotation-invariant alternatives (projection-based) exist but face the determinism limitation. This means dimensionwise ξ measures similarity in BERT's learned feature space specifically, not a basis-invariant geometric property.

6.8 Future Work

Additional benchmarks. Evaluation on SICK, MS MARCO, BEIR, and other semantic similarity benchmarks would strengthen validation beyond STS-B.

Other embedding models. Testing dimensionwise ξ on larger BERT variants (base, large), other architectures (RoBERTa, DeBERTa), and specialized embeddings (biomedical, legal) would assess generalization.

Stochastic embedding methods. Developing approaches to generate genuine variation in embeddings (input perturbation, dropout-enabled models, model ensembles) could enable projection-based ξ for production systems, providing theoretically rigorous alternatives.

Interpretability. While cosine has intuitive geometric interpretation (angle between vectors), ξ values are less immediately interpretable. Visualization tools and calibration studies could improve practitioner understanding.

Hybrid optimization. The α-weighted hybrid shows promise. Learning optimal weights per task or dynamically per query could improve performance beyond either metric alone.

7 Conclusion

We have demonstrated that semantic similarity in BERT sentence embeddings is primarily captured by rank structure, not magnitude—a surprising finding with both theoretical and practical implications. Using Chatterjee's ξ coefficient applied directly to embedding dimensions, we achieve $\rho = 0.859$ correlation with human judgments on 1,500 STS-B benchmark pairs, capturing 99% of the signal obtained by cosine similarity ($\rho = 0.867$). The $r = 0.923$ correlation between rank-only ξ and magnitude-based Pearson suggests that cosine may succeed not because of geometric magnitude relationships, but because it happens to preserve the underlying ordinal structure.

Our mechanistic analysis shows this rank signal is distributed across dimensions: approximately 25% contribute meaningfully, with no single dimension dominating. The 0.86% performance gap between ξ and cosine is not a limitation—it is evidence that magnitude contributes surprisingly little beyond ordinal information. This suggests semantic similarity is fundamentally an ordinal phenomenon: which features activate strongly matters more than how strongly they activate.

Dimensionwise ξ provides practitioners with a validated rank-based alternative to cosine: within 1% performance, $O(d \log d)$ complexity, and complementary strengths (conservative on low-similarity pairs where cosine is lenient). Future work includes validation on additional benchmarks, development of stochastic embedding methods for projection-based ξ, and hybrid optimization. All code, data, and experimental results are available in the supplementary materials.

Data Availability

All experimental code, data, and analysis scripts are included in the supplementary materials. The repository includes:

- Complete source code for all similarity metrics
- Experiment scripts with reproducible configurations
- All raw experimental data (17,500+ observations)
- Generated figures and tables
- Jupyter notebooks for interactive exploration
- Comprehensive documentation and installation instructions

Results are fully reproducible using the provided scripts with fixed random seeds.

Disclosure Statement

The authors report there are no competing interests to declare.

References

[1] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, 2017.

[2] S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

[3] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, pages 63–77, 2005.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[5] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[6] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529, 2019.

[7] Z. Lin and F. Han. On boosting the power of chatterjee's rank correlation. *Biometrika*, 110(2):283–299, 2023.

[8] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pages 6076–6085, 2017.

[9] N. Reimers and I. Gurevych. 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)*, pages 3982–3992, 2019.

[10] H. Shi, M. Drton, and F. Han. On the power of chatterjee's rank correlation. *Biometrika*, 109(2):317–333, 2022.

[11] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.