Predicting Cross-Domain RAG Retrieval Quality using Von Neumann Graph Entropy

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We show that a single spectral quantity computed on a topological Mapper graph over a small unlabelled sample of an embedding corpus predicts retrieval quality without any labelled queries. We apply the Mapper algorithm from topological data analysis to the embedded corpus, constructing a multi-level graph in which nodes represent clusters of nearby documents and edges encode cluster overlap. As connectivity within each Mapper level increases from sparse to dense, we track the Von Neumann entropy of the normalised graph Laplacian of the largest connected component, producing a multi-scale entropy curve. The area under this curve (AUC) is negatively correlated with retrieval quality in two structurally dissimilar domains: financial document retrieval from SEC filings (Spearman $r = -0.68$, $p = 0.001$, $n = 20$ conditions) and tabular data retrieval over Wikipedia tables ($r = -0.47$, $p < 0.01$, $n = 20$ conditions). Both results are statistically significant. Low AUC indicates that the Mapper graph maintains concentrated, non-uniform spectral structure across all connectivity scales, suggesting a corpus with well-separated semantic clusters.

1. Introduction

Retrieval-augmented generation (RAG) systems embed a document corpus into a high-dimensional space and serve queries by nearest-neighbor search [3]. The quality of retrieval depends critically on how the corpus is represented: chunking strategy (for text) or representation depth (for structured data) can produce a $3\times$ difference in top-10 retrieval success rate across otherwise identical systems. Selecting the best configuration currently requires a labelled evaluation set, which is often unavailable at the point when design decisions must be made.

Our question. Can the topological structure of the embedding corpus, captured using the Mapper algorithm from topological data analysis (TDA) and measured without any retrieval labels, serve as a proxy for retrieval quality?

We propose a diagnostic that applies the Mapper algorithm to a small unlabelled sample of the embedded corpus, then measures the Von Neumann (VN) entropy of the normalised graph Laplacian of the resulting Mapper graph across a sweep of connectivity levels. The area under this entropy curve (AUC) is the key quantity. We show that this single scalar is significantly negatively correlated with retrieval quality in two structurally dissimilar domains, and interpret the finding in terms of how Mapper captures the multi-scale topology of high-dimensional embedding spaces.

2. The Spectral Entropy AUC

2.1. Graph construction via the Mapper algorithm

The central methodological contribution is the use of the **Mapper algorithm** [5] from topological data analysis to construct the graph. In a standard k-NN graph every individual document is a node and edges connect documents that are mutual nearest neighbors. The Mapper construction is qualitatively different:

1. **Nodes are clusters, not documents.** A family of overlapping cover elements is placed over the embedding space. Within each cover element, nearby documents are grouped by a local clustering step. Each resulting cluster becomes a single node in the Mapper graph, aggregating multiple documents into one topological unit.
2. **Edges represent set overlap, not pairwise distance.** Two nodes are connected by an edge if and only if their underlying document sets share at least one member. This overlap criterion encodes topological adjacency: it identifies which semantic regions of the corpus are contiguous, not merely which pairs of documents are close.

The result is a graph whose structure approximates the *shape* of the high-dimensional point cloud at a chosen resolution. Topological features such as connected components, branching structure, and loop closure are preserved in a way that k-NN graphs, which capture only local pairwise proximity, cannot.

We use the **Cobalt** Mapper implementation from BluelightAI [2] with $M{=}6$ cover levels and $K{=}6$ maximum neighbors in the internal clustering. Cobalt produces a graph whose nodes are groups of nearby embedded documents and whose edges record that two such groups share at least one document. The $M{=}6$ levels are not replications: each level represents a **different resolution** of the topological cover, yielding six Mapper graphs of varying granularity over the same embedding corpus. This multi-scale structure is the hallmark of TDA: topology is examined simultaneously at coarse and fine resolutions, revealing structure that single-scale methods miss.

2.2. The connectivity sweep within each Mapper level

Within each of the M Mapper levels, we sweep connectivity by progressively adding edges from that level's Mapper edge list, producing a sequence of graphs with increasing average node degree from $\overline{k} \approx 0$ to $\overline{k}_{\max} \approx 12$. We sample 10 edge-count snapshots per level, giving up to $M \times 10 = 60$ graph states per replicate. Three independent 500-item replicates are drawn per condition. The AUC is computed over all observed $(H^{(t)}, \overline{k}^{(t)})$ pairs across all levels and edge counts, and averaged over replicates.

2.3. Von Neumann entropy of the Mapper graph Laplacian

At each observed state (m, e) — Mapper level m, edge count e — let $\mathcal{C}^{(m,e)}$ be the largest connected component of the corresponding Mapper graph, and let $L^{(m,e)}$ be its combinatorial Laplacian. We define the *normalised Laplacian density matrix*:

$$\tilde{L}^{(m,e)} \;=\; \frac{L^{(m,e)}}{\mathrm{tr}\big(L^{(m,e)}\big)}. \tag{1}$$

Its Von Neumann entropy is:

$$H^{(m,e)} \;=\; -\mathrm{tr}\Big(\tilde{L}^{(m,e)} \log \tilde{L}^{(m,e)}\Big) \;=\; -\sum_{j} \lambda_j \log \lambda_j, \tag{2}$$

where $\{\lambda_j\}$ are the eigenvalues of $\tilde{L}^{(m,e)}$ [4]. $H^{(m,e)}$ measures how uniformly the Mapper graph's Laplacian spectrum is distributed. Because Mapper nodes are *clusters* of semantically related documents, the Laplacian here encodes how topological connectivity is distributed across semantic regions of the corpus.

2.4. The AUC metric

We define the *spectral entropy AUC* as the trapezoid integral of $H^{(m,e)}$ with respect to average node degree $\overline{k}^{(m,e)}$, pooled across all M Mapper levels and all edge-count snapshots within each:

$$\mathrm{AUC} \;=\; \sum_{m=1}^{M} \sum_{e} H^{(m,e)} \cdot \Delta \overline{k}^{(m,e)}. \tag{3}$$

Because it integrates across all M Mapper levels, it captures topological spectral complexity at *multiple scales* simultaneously, a property unique to Mapper-based analysis and not achievable with a single k-NN graph.

2.5. Why low AUC predicts better retrieval

A *low* AUC means the entropy curve stays suppressed throughout the sweep: even as edges are added, the Laplacian spectrum remains concentrated on a collection of dominant modes. This signals a Mapper graph with persistent structural non-uniformity, which in embedding-space terms corresponds to a corpus with stable, well-separated topical clusters, so that queries can be unambiguously routed to the correct cluster. A *high* AUC means the curve rises quickly and stays elevated: the spectrum becomes near-uniform, indicating a near-isotropic embedding space where many documents score nearly equally against any query, degrading retrieval.

3. Experimental Setup

We evaluate across two retrieval domains that differ in data modality, corpus format, query type, and variation axis (Table 1). We conjecture that if the metric holds in both, it is not an artifact of any one domain's specific properties.

Embedding models. Four models are used identically across both domains: e5-large-v2 and bge-large-en-v1.5 (1024-dim); e5-small-v2 and all-MiniLM-L6-v2 (384-dim). Sweep parameters are as in Section 2 (M=6, K=6, 10 connectivity steps, 500-item samples, 3 replicates; full settings in Appendix D).

Retrieval evaluation. For each (model, strategy) condition, retrieval quality is measured over the full query set using Hit@K, MRR@10, and NDCG@10. We report Spearman rank correlation between the AUC and Hit@10 across the 20 conditions within each domain.

Table 1: Experimental domains.
Each experiment has 4 embedding models $\times$ 5 corpus strategies $=$ 20 conditions.

Domain	Corpus and task	Variation axis
FinDER	38,801 SEC filing paragraphs; 5,703 analyst queries. *Task:* Retrieve the correct paragraph.	Chunk size: `fixed_256` $\rightarrow$ `fixed_1024`, natural paragraphs
Tabular	$\approx$16,500 Wikipedia tables; statements about table content. *Task:* Retrieve the correct table.	Context richness: headers only $\rightarrow$ full table with context

4. Results

4.1. Main result

Table 2: Spearman r between spectral entropy AUC and Hit@10. $n = 20$ conditions per domain (4 models $\times$ 5 strategies). Both results significant at $p < 0.01$.

Domain	Spearman r	p	Sig.
FinDER (financial text)	-0.68	0.001	$**$
Tabular (Wikipedia tables)	-0.47	$<$0.01	$**$

$** \; p < 0.01$; Spearman, two-tailed.

The spectral entropy AUC is negatively correlated with Hit@10 retrieval quality in both domains, statistically significant at $p < 0.01$. Figure 1 shows scatter plots of several Mapper-derived graph metrics against FinDER Hit@10; the Von Neumann entropy panel (lower-middle) shows the mean entropy across states, while the AUC variant — the trapezoid integral over all connectivity snapshots and Mapper levels — achieves the stronger $r{=}-0.68$ reported in Table 2.

4.2. What the ranking looks like in FinDER

Within the FinDER domain, the AUC's ability to rank corpus configurations is especially clear because the variation axis (chunk size) spans a wide performance range, from Hit@10 of 21% at 256-character chunks to 59% at 1024-character chunks. Averaging across the four embedding models, the AUC-based rank ordering of the five chunk strategies is identical to the retrieval quality rank ordering: `fixed_1024` (Hit@10 $=$ 59%) $>$ `para_min200` (56%) $>$ `para_min50` (43%) $>$ `fixed_512` (41%) $>$ `fixed_256` (21%). Measuring the spectral entropy AUC on a 500-item sample, without running any retrieval queries, correctly identifies which chunk strategy will perform best. The full ranking table is given in Appendix B.

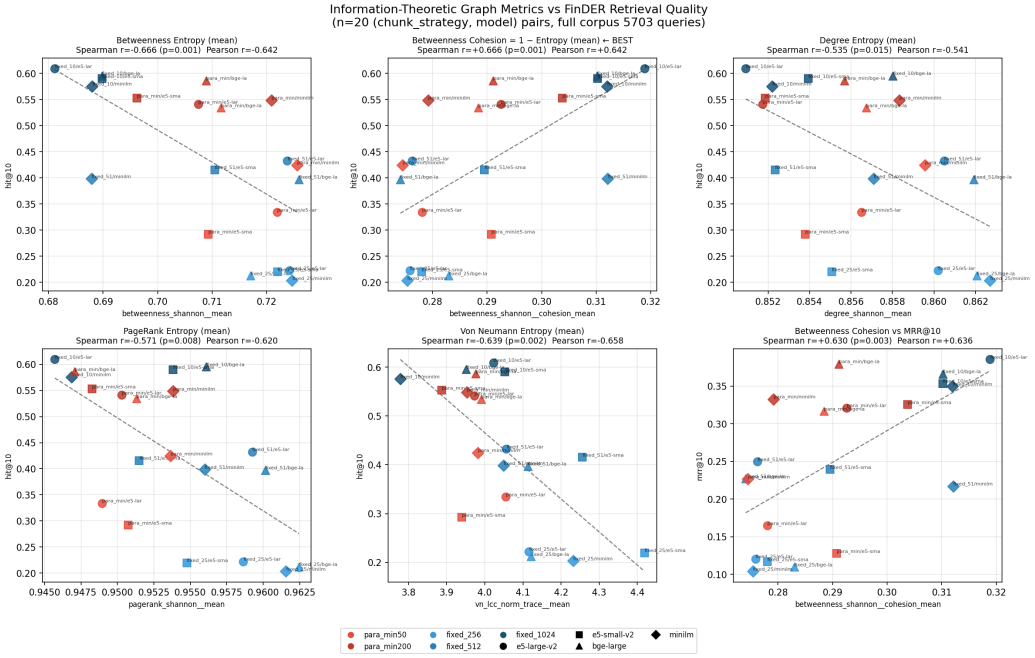

Figure 1: Spearman correlations between Mapper-derived graph metrics and FinDER Hit@10 (n=20 conditions: 4 embedding models $\times$ 5 chunk strategies). Each panel shows one metric with a dashed regression line and the Spearman r (top-right). The lower-middle panel shows Von Neumann entropy (mean across connectivity states); the AUC variant integrates this curve over all Mapper levels and achieves $r{=}-0.68$. Colours indicate chunk strategy; markers indicate embedding model.

4.3. Interpretation: what the entropy curve reveals

These results match the mechanism in Section 2. Large chunks (`fixed_1024`) and rich table representations cluster same-topic documents together: the Mapper graph keeps hub-and-spoke structure at every scale, the spectrum stays non-uniform, and the $H^{(t)}$ stays suppressed. Small chunks (`fixed_256`) and header-only tables scatter content across near-equidistant fragments: the Mapper graph approaches a regular topology, the spectrum flattens, and $H^{(t)}$ rises rapidly, corresponding to high AUC and poor retrieval. The two domains exhibit the same phenomenon on different axes (chunk size vs. representation richness).

5. Discussion and Conclusion

The spectral entropy AUC, on a Mapper graph of a 500-item unlabelled sample, is significantly negatively correlated with RAG retrieval quality in financial document retrieval ($r{=}-0.68$, $p{=}0.001$) and tabular retrieval ($r{=}-0.47$, $p{<}0.01$). Distance concentration in high-dimensional spaces [1] degrades dense retrieval; the AUC is a label-free indicator of semantic document separation as the Mapper graph density increases.

References

[1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer, 2001.

[2] Blue Light AI. Cobalt: High-dimensional data analysis and evaluation documentation. https://docs.cobalt.bluelightai.com/cobalt.html, 2026. Accessed: 2026-02-15.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[4] Filippo Passerini and Simone Severini. The von Neumann entropy of networks. *arXiv preprint arXiv:0812.2597*, 2008.

[5] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Eurographics Symposium on Point-Based Graphics*, pages 91–100, 2007.

Appendix A. Within-Model Scatter Plots

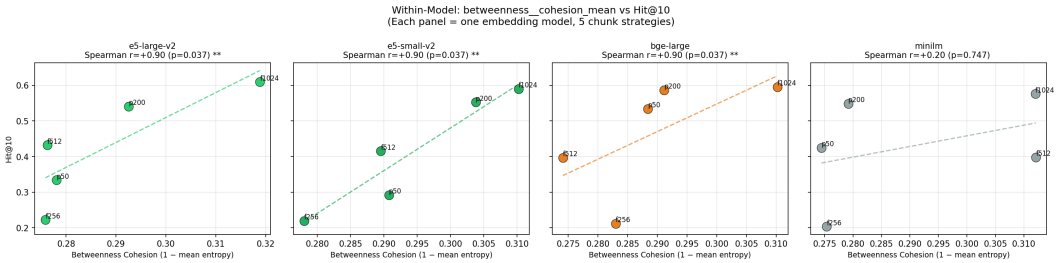

Figure 2: Within-model scatter: Mapper betweenness-cohesion mean (a related topological sweep statistic computed from the same Mapper graph) vs. Hit@10, one panel per embedding model (n=5 conditions each, one per chunk strategy). Three of four models achieve r=+0.90 (p=0.037); MiniLM is the exception (r=+0.20). The Von Neumann entropy AUC exhibits the same within-model pattern. Points are labelled by chunk strategy.

Appendix B. Full FinDER Retrieval Results

Table 3: Retrieval evaluation on 5,703 FinDER queries, all conditions, sorted by Hit@10. `fixed_2048` is included for reference but is outside the metric's domain of validity (coverage regime).

Model	Chunk strategy	Hit@1	Hit@5	Hit@10	MRR@10	NDCG@10
e5-large-v2	fixed_2048	.371	.613	.699	.474	.528
e5-small-v2	fixed_2048	.317	.574	.666	.425	.483
e5-large-v2	fixed_1024	.284	.522	.609	.385	.439
bge-large	fixed_1024	.269	.496	.596	.366	.421
e5-small-v2	fixed_1024	.251	.493	.590	.353	.410
MiniLM	fixed_1024	.254	.480	.576	.350	.404
bge-large	para_min200	.282	.505	.586	.379	.429
e5-small-v2	para_min200	.221	.467	.553	.325	.380
MiniLM	para_min200	.238	.459	.549	.333	.384
e5-large-v2	para_min200	.220	.463	.541	.321	.374
bge-large	para_min50	.219	.444	.534	.317	.369
MiniLM	para_min50	.146	.337	.425	.226	.273
e5-large-v2	para_min50	.101	.255	.334	.165	.205
e5-small-v2	para_min50	.069	.209	.292	.128	.166
e5-large-v2	fixed_512	.174	.354	.432	.250	.293
e5-small-v2	fixed_512	.166	.342	.416	.239	.281
MiniLM	fixed_512	.141	.322	.398	.217	.260
bge-large	fixed_512	.159	.318	.397	.227	.267
e5-large-v2	fixed_256	.082	.173	.222	.121	.145
e5-small-v2	fixed_256	.076	.172	.220	.117	.141
bge-large	fixed_256	.071	.162	.212	.110	.134
MiniLM	fixed_256	.066	.154	.203	.104	.127

Appendix C. Metric Definition Summary

`vn_lcc_norm_trace_auc` The Mapper algorithm (M=6 levels, K=6) is applied to a 500-item sample of the embedded corpus. Within each Mapper level, edges are added progressively (10 snapshots per level, $\leq$ 60 states per replicate). At each observed state, the largest connected component of the Mapper graph is identified — where nodes are document clusters and edges are cluster overlaps — and its normalised Laplacian density matrix $\tilde{L} = L/\mathrm{tr}(L)$ is formed. The Von

Neumann entropy $H = -\mathrm{tr}(\tilde{L}\log\tilde{L})$ is computed. The AUC is the trapezoid integral of all H values with respect to average node degree, pooled across all M levels and averaged over 3 replicates. **Lower AUC predicts better retrieval** (Spearman $r = -0.68$, $p = 0.001$ in FinDER; $r = -0.47$, $p < 0.01$ in Tabular).

Appendix D. Sweep Parameters

Parameter	**Value**
Mapper levels (M)	6
Max neighbours (K)	6
Connectivity steps per rep.	10
Sample size per replicate	500
Replicates per condition	3
Avg. degree range	$\approx$ 1–12
LCC	Largest connected component
Retrieval top-K	10
GPU	NVIDIA A100-SXM4-80GB