Semantic Shift: the Fundamental Challenge in Text Embedding and Retrieval

Anonymous ACL submission

Abstract

Transformer-based embedding models rely on pooling to map variable-length text into a single vector, enabling efficient similarity search but also inducing well-known geometric pathologies such as anisotropy and length-induced embedding collapse. Existing accounts largely describe *what* these pathologies look like, yet provide limited insight into *when* and *why* they harm downstream retrieval. In this work, we argue that the missing causal factor is *semantic shift*: the intrinsic, structured evolution and dispersion of semantics within a text.

We first present a theoretical analysis of *semantic smoothing* in Transformer embeddings: as the semantic diversity among constituent sentences increases, the pooled representation necessarily shifts away from every individual sentence embedding, yielding a smoothed and less discriminative vector. Building on this foundation, we formalize semantic shift as a computable measure integrating local semantic evolution and global semantic dispersion. Through controlled experiments across corpora and multiple embedding models, we show that semantic shift aligns closely with the severity of embedding concentration and predicts retrieval degradation, whereas text length alone does not. Overall, semantic shift offers a unified and actionable lens for understanding embedding collapse and for diagnosing when anisotropy becomes harmful.

1 Introduction

Text embeddings have become indispensable for retrieval, question answering, clustering, and a wide range of semantic processing tasks. Classic distributional methods (e.g. Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText(Bojanowski et al., 2017)) have been largely superseded by Transformer-based Pretrained Language Models (PLM) such as BERT (Devlin et al., 2019) and its variants (Liu et al., 2019), as well as GPT-style models (Radford et al., 2019), which produce context-sensitive representations that substantially improve semantic matching.

Despite their empirical success, a growing body of work has revealed that embedding spaces exhibit non-trivial geometric *pathologies*. A widely discussed phenomenon is *anisotropy*, where embeddings concentrate into a narrow cone rather than being uniformly distributed (Gao et al., 2019; Ethayarajh, 2019). A series of post-processing and normalization techniques have been proposed to mitigate such issues, e.g. removing dominant directions (Mu and Viswanath, 2018; Arora et al., 2017; Raunak et al., 2019), whitening (Su et al., 2021; Huang et al., 2021), or flow-based transformations (Li et al., 2020). However, recent analyzes suggest that global concentration metrics can be misleading and do not reliably predict semantic quality or downstream performance (Timkey and van Schijndel, 2021; Fuster-Baggetto and Fresno, 2022; Ait-Saada and Nadif, 2023).

Recent work has identified and formalized length-induced embedding collapse, where embeddings of longer texts exhibit reduced variance and become increasingly difficult to distinguish (Zhou et al., 2025). They attribute this effect to the attention mechanism: as input length grows, the attention matrix exhibits a stronger low-pass filtering behavior, accelerating the suppression of high-frequency semantic variations and consequently driving long-text embeddings toward increasingly similar representations.

These observations are important but incomplete: they characterize *what* embedding spaces look like, not *why* such structures arise or when they actually harm downstream performance. A striking paradox illustrates this gap. When we embed the same corpus using different models, the resulting Mean Pairwise Distance (MPD) (Ethayarajh, 2019; Ait-Saada and Nadif, 2023), a common measure of concentration/anisotropy, can vary dramatically.

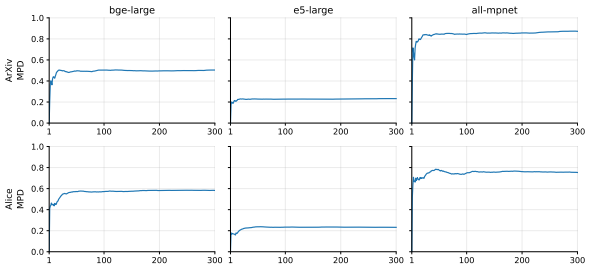

Figure 1: Mean Pairwise Distance (MPD) curves for three embedding models across two corpora. The x-axis is the number of sentences; the y-axis is MPD.

Figure 1 illustrates how the MPD of sentence embeddings evolves on two corpora, ArXiv (Common Pile and arXiv.org, 2023) and Alice's Adventures in Wonderland (Project Gutenberg; Carroll), under several widely used embedding models bge-large (Xiao et al., 2024), e5-large (Wang et al., 2022), and all-mpnet (Song et al., 2020). In this experiment, texts are segmented into sentences, all sentences are embedded, and the MPD is computed incrementally over the first 1, 2, ..., n sentences. As shown in Figure 1, the MPD converges to a stable value once n becomes sufficiently large. However, the converged MPD differs drastically across models: bge-large stabilizes around 0.6, e5-large around 0.2, and all-mpnet around 0.8.

Despite these large discrepancies in embedding concentration, these models exhibit broadly comparable performance in practical downstream tasks. This observation makes it difficult to attribute degraded performance solely to anisotropy (i.e., embedding concentration), and it cannot be explained by length-induced embedding collapse either, since all models embed texts of identical length. Together, these findings raise a central question.

> **If neither embedding concentration, anisotropy, nor length collapse can account for the behavior observed in Figure 1, what factors—beyond model-specific effects—fundamentally drive embedding concentration and, more importantly, lead to difficulties in embedding-based retrieval?**

In this paper, we argue that the fundamental factor is **semantic shift**: the intrinsic, structured evolution of semantics within a text. Natural language exhibits strong local coherence – adjacent sentences tend to be semantically similar, but this similarity naturally decays as one moves through the text. Over long ranges, the accumulated change can be substantial. This process resembles a "telephone game": each step preserves local information, yet the meaning at the end can differ markedly from the beginning.

To substantiate this claim, we first provide a theoretical explanation for *semantic smoothing* in Transformer embeddings. We show that because Transformer encoders necessarily aggregate token-level representations through mean pooling or attention pooling, the resulting text embedding is effectively a convex combination of its constituent token/sentence embeddings. We further prove that as the pairwise semantic diversity among tokens/sentences increases, the aggregated embedding inevitably moves farther away from every individual token/sentence. This mathematically explains why embeddings for multi-sentence texts tend to under-represent any specific semantic component, shifting toward a compromise direction. This smoothing effect directly connects semantic diversity to length collapse and anisotropy.

Building on this theoretical foundation, we introduce a formal definition of *semantic shift* that captures both local semantic evolution and global semantic dispersion. Using controlled experiments on synthetic concatenation patterns, we show that semantic shift, not text length, predicts the severity of embedding concentration. Furthermore, in the retrieval experiments, we observe that anisotropy becomes much more harmful when induced by strong semantic shifts, whereas the harm caused by anisotropy solely based on length is much less significant.

2 Semantic Smoothing in Transformer-Based Embedding Models

2.1 Token-Level Pooling and Its Sentence-Level Interpretation

Transformer encoders construct text embeddings by aggregating contextualized token representations through a fixed pooling mechanism. In this section, we show that any pooling-based embedding model inevitably smooths and dilutes the semantics of a multi-sentence text, and that the extent of this dilution grows monotonically with the semantic diversity of the constituent sentences. This provides a theoretical foundation for understanding anisotropy, length collapse, and their connection to semantic shift.

Let an input text be tokenized as $(x_1, \ldots, x_N)$,

and let a Transformer encoder produce contextualized token embeddings

$$h_1, h_2, \ldots, h_N \in \mathbb{R}^d. \tag{1}$$

To obtain a fixed-length text embedding, all widely used models apply a pooling operator:

$$z = \mathrm{Pool}(h_1, \ldots, h_N). \tag{2}$$

Two pooling mechanisms dominate practice:

Mean pooling. Used in Sentence-BERT (Reimers and Gurevych, 2019), SimCSE (Gao et al., 2021), E5 (Wang et al., 2022), BGE (Xiao et al., 2024), and many other embedding models:

$$z = \frac{1}{N} \sum_{t=1}^{N} h_t. \tag{3}$$

Attention-weighted pooling. Used in vanilla BERT (Devlin et al., 2019; Clark et al., 2019; Ethayarajh, 2019), where the `[CLS]` token representation after self-attention can be expressed as:

$$h_{\mathrm{CLS}} = \sum_{t=1}^{N} \alpha_t \, v_t, \tag{4}$$

with attention-derived weights $\alpha_t \geq 0$ and $\sum_t \alpha_t = 1$. Thus, CLS pooling is also a convex combination of token embeddings.

Since tokens naturally organize into sentences and attention layers allow information exchange within each sentence, the aggregated embedding can be rewritten as a weighted sum over sentence-level embeddings:

$$z = \sum_{i=1}^{k} w_i \, e_i, \tag{5}$$

where e_i is the (averaged) embedding of the i-th sentence and the weights w_i correspond to token proportions or attention distributions. This equivalence justifies analyzing the behavior of text embeddings by treating a text as a set of sentence embeddings being pooled into a single vector.

2.2 Semantic Diversity Forces Semantic Dilution

Pooling imposes strong geometric constraints: the resulting text embedding must lie in the convex hull of its constituent sentence embeddings. When these sentences are semantically homogeneous, pooling preserves their direction. When they are diverse, pooling forces them into a compromised direction, diluting each sentence's individual meaning.

To make this precise, suppose a text contains k unit-normalized sentence embeddings:

$$e_1, \ldots, e_k \in \mathbb{R}^d, \qquad \|e_i\| = 1, \tag{6}$$

and let the pooled embedding be

$$\mu = \frac{1}{k} \sum_{i=1}^{k} e_i, \qquad \hat{\mu} = \frac{\mu}{\|\mu\|}. \tag{7}$$

We quantify sentence-level semantic diversity by the mean pairwise cosine distance:

$$C_{\mathrm{pair}} = \frac{2}{k(k-1)} \sum_{i<j} \left(1 - e_i^\top e_j\right), \tag{8}$$

and measure how "unlike" the aggregated embedding is relative to the original sentences by

$$C_{\mathrm{mean}} = \frac{1}{k} \sum_{i=1}^{k} \left(1 - e_i^\top \hat{\mu}\right). \tag{9}$$

Theorem 1 (Semantic Dilution)**.** *For any set of unit-normalized sentence embeddings, the discrepancy between the pooled text embedding $\hat{\mu}$ and the constituent sentences satisfies the following:*

$$C_{\mathrm{mean}} = 1 - \sqrt{1 - \frac{k-1}{k} C_{\mathrm{pair}}}. \tag{10}$$

Consequently, C_{mean} is a strictly increasing function of C_{pair} for all $k \geq 2$.

Proof. We first compute

$$C_{\mathrm{mean}} = 1 - \frac{1}{k} \sum_{i=1}^{k} e_i^\top \hat{\mu} = 1 - \frac{1}{k\|\mu\|} \sum_{i=1}^{k} e_i^\top \mu. \tag{11}$$

Since $\sum_{i=1}^{k} e_i = k\mu$, we obtain

$$C_{\mathrm{mean}} = 1 - \|\mu\|. \tag{12}$$

Next, we expand the squared norm of μ:

$$\|\mu\|^2 = \left\| \frac{1}{k} \sum_{i=1}^{k} e_i \right\|^2 = \frac{1}{k^2} \left(k + 2 \sum_{1 \leq i < j \leq k} e_i^\top e_j \right) \tag{13}$$

Expanding Equation 8

$$C_{\mathrm{pair}} = 1 - \frac{2}{k(k-1)} \sum_{i<j} e_i^\top e_j, \tag{14}$$

we find

$$\sum_{i<j} e_i^\top e_j = \frac{k(k-1)}{2}\left(1 - C_{\text{pair}}\right), \tag{15}$$

and therefore

$$\|\mu\|^2 = 1 - \frac{k-1}{k} C_{\text{pair}}. \tag{16}$$

Combining (12) and (16) yields

$$C_{\text{mean}} = 1 - \sqrt{1 - \frac{k-1}{k} C_{\text{pair}}}\,. \tag{17}$$

The expression is strictly increasing in C_{pair} because its derivative

$$\frac{dC_{\text{mean}}}{dC_{\text{pair}}} = \frac{k-1}{2k} \cdot \frac{1}{\sqrt{1 - \frac{k-1}{k} C_{\text{pair}}}} \tag{18}$$

is strictly positive for $k \geq 2$. □

This theorem states that *the more diverse the sentences that make up a text, the greater the average difference between the overall semantics of the text and the semantics of each individual sentence*.

2.3 Empirical Validation of Theorem 1

Theorem 1 establishes a strict monotonic relationship between sentence-level semantic diversity and text–sentence discrepancy under an idealized pooling assumption. We now empirically verify that this relationship also holds in practice for real Transformer-based embedding models, **where text embeddings are produced by encoding the concatenated text directly rather than by explicit sentence averaging.**

Using the ArXiv (Common Pile and arXiv.org, 2023) corpus and the bge-large model (Xiao et al., 2024), we construct sentence groups of size $k = 10$ under three sampling regimes: local (consecutive sentences), medium (non-adjacent sentences) and high (uniformly random sentences from the corpus), repeating each regime 200 times. In each trial, we select 10 sentences according to the corresponding regime, concatenate them into a single text, encode the text once, and obtain the embeddings of the constituent sentences by encoding each sentence separately. We then compute the mean pairwise cosine distance between sentence embeddings, C_{pair}, and the mean cosine distance between the text embedding and its constituent sentence embeddings, C_{mean}. Additional results across different models and corpora are provided in Appendix D.

Correlation between C_{pair} and C_{mean}. Figure 2 reports the scatter plot of C_{mean} versus C_{pair} under three sampling regimes. We observe a strong monotonic association: Spearman's rank correlation is $\rho = 0.8838$ and Kendall's $\tau = 0.7074$ (both highly significant with $p \ll 10^{-100}$). **This empirical result supports Theorem 1 in practice: as sentence-level semantic diversity increases (larger C_{pair}), the discrepancy between the concatenated-text embedding and its constituent sentence embeddings also increases (larger C_{mean}).**

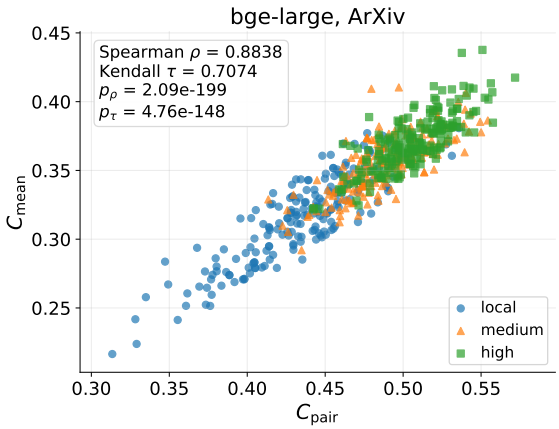

Figure 2: Scatter plot of C_{mean} vs. C_{pair} on ArXiv using bge-large model.

These results provide strong empirical support for Theorem 1. Despite the fact that real embedding models employ complex self-attention and normalization mechanisms rather than explicit sentence averaging, increased sentence-level diversity reliably leads to a text embedding that is farther, on average, from every individual sentence embedding. This confirms that semantic dilution is not merely a theoretical artifact of mean pooling, but a robust phenomenon in practical embedding systems.

2.4 Implications for Length Collapse and Anisotropy

Theorem 1 has direct implications for the geometry of Transformer-based embedding spaces.

Length-induced collapse. Longer texts tend to contain more diverse semantics. As C_{pair} increases with text length, Theorem 1 implies that $\|\mu\|$ decreases (Equation 12), pushing pooled embeddings to collapse to the origin. After normalization, these embeddings become more concentrated in direction, producing the length-induced collapse.

Anisotropy as a consequence of pooling. If many texts contain diverse semantics, their pooled embeddings cluster around a small region of the unit sphere, increasing global anisotropy. Crucially, anisotropy is therefore not an inherent defect of embedding representation, but a geometric consequence of semantic diversity combined with pooling.

Semantic shift as the missing causal factor. Pooling itself does not harm retrieval: if all sentences are similar (C_{pair} small), then $\hat{\mu}$ remains faithful to each sentence. Problems arise only when semantic diversity is large, causing embeddings to blend multiple divergent meanings into a single vector. This observation motivates our formalization of semantic shift in the next section.

3 Semantic Shift: Formalization and Properties

The theoretical analysis in Section 2 shows that semantic diversity among sentences causes semantic dilution: the pooled text embedding becomes increasingly distant from each sentence, shifting toward a compromise direction. While this explains why multi-sentence texts exhibit weaker semantic fidelity, it raises a natural question: *how does semantic diversity itself arise as we move through a text?*

In natural language, semantics evolve gradually. Adjacent sentences typically share strong local coherence, while sentences farther apart may describe different entities, events, or topics. This structured progression is neither random noise nor model-induced drift; rather, it reflects the intrinsic, content-driven evolution of meaning. We refer to this phenomenon as **semantic shift**. In this section, we formalize semantic shift and argue why it offers a more fundamental perspective on embedding pathologies than length or concentration alone.

3.1 Local Semantic Evolution

A natural attempt to capture semantic evolution is to sum the distances between consecutive sentences.

Definition 1 (Local Semantic Evolution)**.** *For sentence embeddings $e_1, \ldots, e_k$, the local semantic evolution up to length k is*

$$\text{Local}(k) = \sum_{i=1}^{k-1} \big(1 - \cos(e_i, e_{i+1})\big). \tag{19}$$

Although intuitive, this Local Semantic Evolution conflates two qualitatively distinct scenarios:

- **Monotonic semantic shift**, where sentences gradually move away from earlier ones in a coherent direction.
- **Semantic clustering**, where sentences fluctuate locally but remain within a compact region around a shared theme.

Both cases may yield similar local cumulative differences, yet their global semantic structures, and thus their impact on pooling and retrieval, differ drastically. Purely local measures cannot distinguish coherent progression from topic mixing, nor can they capture how far the sentence set spreads in the embedding space. Since semantic dilution (Theorem 1) depends on the global diversity of sentence embeddings, a more complete definition must incorporate both local and global information.

3.2 Global Semantic Dispersion

We quantify how semantically dispersed a set of k sentences is using their mean pairwise distance.

Definition 2 (Semantic Dispersion)**.** *Given sentence embeddings $e_1, \ldots, e_k$, the* semantic dispersion *is*

$$\text{Disp}(k) = \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} \big(1 - \cos(e_i, e_j)\big), \tag{20}$$

with the convention $\text{Disp}(1) = 0$.

A larger $\text{Disp}(k)$ indicates that the sentences occupy a wider region in the embedding space.

3.3 Semantic Shift: Integrating Local and Global Structure

Semantic dilution occurs when the local semantic evolution interacts with the global semantic dispersion. When both are small, the text maintains a stable topic; when both are large, semantics evolve into distinct conceptual regions, creating strong dilution under pooling.

We therefore define semantic shift as the interaction between these two factors.

Definition 3 (Semantic Shift)**.** *For a sequence of k sentence embeddings $e_1, \ldots, e_k$, the* semantic shift *is defined as*

$$\text{Shift}(k) = \text{Local}(k) \cdot \text{Disp}(k). \tag{21}$$

4 A New Lens on Length Collapse and Anisotropy

Theoretical results in Sections 2 and 3 suggest that embedding pathologies commonly attributed to text length may, in fact, be driven by semantic shift. In this section, we design controlled experiments to disentangle these two factors and show that semantic shift, rather than length, is the primary determinant of concentration, anisotropy, and retrieval degradation.

4.1 How Semantic Shift Drives Length Collapse and Anisotropy

Previous work has observed that embeddings of longer texts tend to be more concentrated, and recent work attributes length-induced embedding collapse in PLM-based models to attention mechanisms that increase text length accelerates low-pass filtering in the attention matrix, making text embeddings for longer texts more similar(Zhou et al., 2025). Then, this concentration is argued to harm retrieval, clustering, and related tasks.

However, we find that the dominant factor behind this effect is not length, but the *strength of semantic shift* inside the text. In the following, we present controlled experiments to disentangle these factors.

Experimental setup. In the experiments, we used a diverse set of embedding models of different types and scales, including bge-large (Xiao et al., 2024), e5-large (Wang et al., 2022), all-mpnet (Song et al., 2020), gte-large (Li et al., 2023), and text-embedding models (OpenAI, 2024), covering open source and closed source systems. On the corpus side, we also evaluated a broad range of text sources, including academic documents, long-form novels, knowledge-focused articles, and encyclopedic materials. Due to space limitations, we present only a subset of the results in the main paper—showcasing selected models and selected corpora. Additional results, along with full details on the models and datasets used, are provided in the Appendix B.

For each corpus, we segment the text into sentences to obtain an ordered sequence

$$S = (s_1, s_2, \ldots, s_n).$$

We then construct longer "sentences" by concatenating sentences in S according to three patterns, and embed all resulting sequences with a fixed PLM encoder (e.g., bge-large). For each resulting sequence, we measure the embedding concentration using MPD.

Concatenation patterns.

- **Repeat concatenation.** Each sentence is repeated multiple times:

$$\begin{aligned} S2^{\text{rep}} &= (s_1 s_1,\ s_2 s_2,\ \ldots,\ s_n s_n), \\ S5^{\text{rep}} &= (s_1^5,\ s_2^5,\ \ldots,\ s_n^5), \\ S10^{\text{rep}} &= (s_1^{10},\ s_2^{10},\ \ldots,\ s_n^{10}), \end{aligned}$$

  where s_i^m denotes s_i repeated m times. Here, length increases but the underlying semantics of each unit do not change.

- **Sequential concatenation.** Each sentence is concatenated with its immediate successors:

$$\begin{aligned} S2^{\text{seq}} &= (s_1 s_2,\ s_2 s_3,\ \ldots,\ s_{n-1} s_n,\ s_n), \\ S5^{\text{seq}} &= (s_1 \ldots s_5,\ s_2 \ldots s_6,\ \ldots,\ s_n), \\ S10^{\text{seq}} &= (s_1 \ldots s_{10},\ s_2 \ldots s_{11},\ \ldots,\ s_n). \end{aligned}$$

  Here, length increases and semantics evolve smoothly within a local window along the original text.

- **Random concatenation.** Each sentence is concatenated with randomly sampled sentences from the entire corpus:

$$\begin{aligned} S2^{\text{rand}} &= (s_1 s_{i_1},\ s_2 s_{i_2},\ \ldots,\ s_{n-1} s_{i_{n-1}},\ s_n), \\ S5^{\text{rand}} &= (s_1 s_{i_1} s_{j_1} s_{k_1} s_{\ell_1},\ \ldots,\ s_n), \\ S10^{\text{rand}} &= (s_1 \ldots,\ s_2 \ldots,\ \ldots,\ s_n), \end{aligned}$$

  where $s_{i_1}, s_{j_1}, s_{k_1}, s_{\ell_1}, \ldots$ are sentences sampled independently from S. Here, both length and semantic heterogeneity increase.

In all three patterns, we embed the resulting sequences and compute the MPD of the embeddings for S, $S2$, $S5$, and $S10$ (where we omit the superscripts in the figure labels for brevity). A lower MPD indicates a stronger embedding concentration, corresponding to the phenomena described by length collapse and anisotropy.

Results and analysis. Figure 3 summarizes the MPD changes across two corpora (ArXiv (Common Pile and arXiv.org, 2023) and Alice's Adventures in Wonderland (Project Gutenberg; Carroll)) and different concatenation patterns; the embedding model is bge-large.

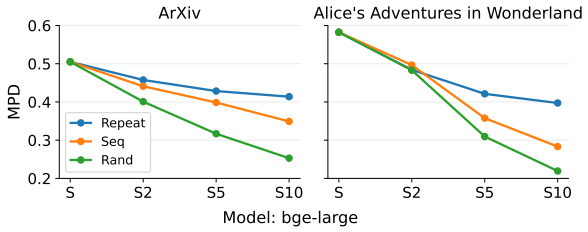

Figure 3: Variation of MPD under different sentence concatenation patterns across two corpora.

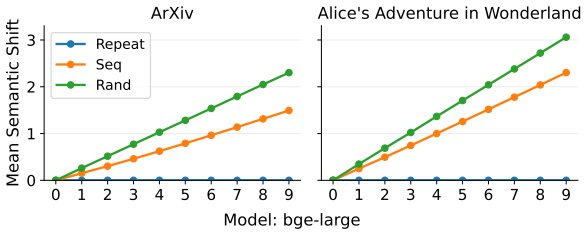

Figure 4: Mean semantic shift increases with hop distance across two corpora under different concatenation modes. The x-axis is the hop distance; the y-axis is mean semantic shift.

For the ArXiv corpus, under repeat and concatenation, MPD decreases slowly as we move from S to $S10$, while under sequential concatenation, the MPD decreases more rapidly. Under random concatenation, MPD decreases much more sharply. From S to $S10$, the drop in MPD is roughly three times larger than in the repeat pattern.

This indicates that pure lengthening (repeat) induces some concentration but not dramatically. In contrast, sequential and random concatenation injects a strong semantic shift within each concatenated unit, causing semantics to be smoothed and diluted, resulting in a much more severe embedding concentration.

For the Alice's Adventures in Wonderland corpus, we observe a similar overall trend, except that the range of variation in MPD is wider, which is likely due to the different types of corpora.

Across two corpora, the MPD drop from S to $S10$ is larger under sequential and random concatenation, again supporting the view that strong internal semantic variation, rather than length alone, is the main driver of severe embedding concentration.

To directly quantify how semantic shift contributes to embedding concentration, we next measure the semantic shift defined in Definition 3 under the three concatenation patterns. Specifically, for each corpus and each pattern, we take the $S10$ variant and compute semantic shift at different hop distances: 1-hop, 2-hop, . . . , 9-hop. This evaluates how semantics evolve as we move further along the concatenated units.

Figure 4 reports the mean semantic shift for the two corpora. In the ArXiv corpus, the random concatenation pattern produces a semantic shift substantially higher than the sequential pattern at all hop distances. This confirms that random mixing injects strong semantic variation even within the same concatenated unit.

For the Alice's Adventures in Wonderland corpus, the semantic shift exhibited by the random and sequential patterns becomes more similar, reflecting the fact that long narrative texts naturally contain topic transitions and plot developments.

Crucially, when we compare Figure 4 with the MPD results in Figure 3, the relationship becomes clear: the degree of embedding concentration aligns almost perfectly with the measured semantic shift. Sequential and random concatenation, which produce a larger semantic shift, also induce significantly stronger MPD reduction.

These results provide quantitative evidence for our central claim. **The semantic shift, rather than the text length, is the dominant factor driving embedding concentration.**

These results also motivate our next question. **When and why does such concentration actually hurt downstream tasks?**

4.2 Impact on Downstream Retrieval and Revisiting Anisotropy

Anisotropy in embedding spaces—where vectors collapse into a narrow cone—is often reported to be harmful to downstream tasks such as retrieval. However, empirical findings are mixed: some work finds clear negative impacts(Gao et al., 2019; Huang et al., 2021), while others observe little to no degradation(Ait-Saada and Nadif, 2023). Our earlier results (Figure 1) also show that some models (e.g. e5-large) exhibit stronger anisotropy than others (e.g. all-mpnet) but do not perform worse on retrieval benchmarks. This suggests that anisotropy per se is not always harmful; the missing piece is *when* and *why* it becomes problematic.

Building on our semantic shift perspective, we hypothesize that anisotropy is harmful primarily when it is induced by strong semantic shift, not when it is mainly caused by length-induced collapse. To test this, we conduct retrieval experiments on the same corpora and concatenation patterns.

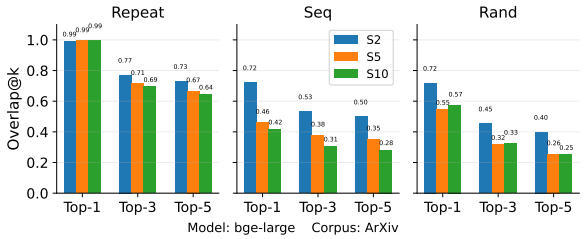

Figure 5: Average self-overlap@k between retrieval results on the original corpus S and its concatenated variants ($S2$, $S5$, $S10$) under repeat, sequential, and random patterns. Higher bars indicate stronger semantic preservation and less retrieval damage.

Self-overlap as a robustness measure. For each corpus $S = (s_1, \ldots, s_n)$, we randomly sample 1000 sentences as query set Q. For each query $q \in Q$, we perform nearest-neighbor search in embedding space under the following settings:

- **Baseline**: retrieve top-k nearest neighbors from the original corpus S.
- **Concatenated variants**: retrieve top-k neighbors from each of $S2, S5, S10$ under the *repeat*, *sequential*, and *random* patterns.

We treat the top-k neighbors from S as a proxy for ground truth, since the set necessarily includes the query itself and its most similar sentences in the original, unmodified corpus. For each variant S' ($S2, S5, S10$) and each query q, we compute the *self-overlap@k*:

$$\text{Overlap@k}(q, S') = \frac{\left|\text{Top@k}(q, S) \cap \text{Top@k}(q, S')\right|}{k}, \quad (22)$$

and then average over all queries. Higher self-overlap@k means that retrieval on the transformed corpus preserves the same neighbors as the original corpus, indicating weaker damage to retrieval.

Results. Figure 5 shows the average overlap@k for $k \in \{1, 3, 5\}$ across concatenation patterns.

For the **repeat** pattern:

- Overlap@1 is almost equal to 1.0 for $S2, S5, S10$, which means that the nearest neighbor is always preserved.
- Overlap@3 and Overlap@5 remain high and stable as length increases.

This confirms that length-induced concentration *without* semantic shift has little impact on retrieval: anisotropy increases (MPD decreases), but relative distances among relevant sentences remain largely intact, so the ranking of neighbors is preserved.

In contrast, for the **sequential** and the **random** patterns:

- Overlap@1 drops to about 0.7 and decreases further as we move from $S2$ to $S10$.
- Overlap@3 and Overlap@5 further deteriorate, with random concatenation consistently yielding the lowest overlap.

Across more corpora and different embedding models (see Appendix E, F for full results), the same pattern holds:

- **Anisotropy driven by length** (repeat) leads to mild embedding concentration and has small harm to retrieval.
- **Anisotropy driven by semantic shift** (sequential and random) simultaneously causes strong concentration and substantial retrieval damage.

5 Conclusion

This paper identifies semantic shift as a fundamental driver of embedding concentration and downstream failures. We provide a principled account of semantic smoothing: pooling-based aggregation in Transformer encoders inevitably yields a compromised representation that shifts away from its constituent sentence embeddings as semantic diversity increases. Building on this insight, we formalize the semantic shift by coupling local semantic evolution with global semantic dispersion and validate it through controlled concatenation studies. Across corpora and embedding models, semantic shift consistently tracks concentration and clarifies when anisotropy becomes harmful to retrieval.

Beyond diagnosis, our findings suggest that the semantic shift can serve as a controllable signal for downstream text processing, offering a path from analysis to practical algorithm design. As a concrete example, we instantiate semantic shift into a shift-aware text segmenter (Semantic Shift Splitter) that adaptively places boundaries while maintaining stable chunk granularity, and observe strong empirical improvements over fixed and semantic splitters. Due to space constraints, we present the splitter and its extensive evaluation in Appendix G.

Limitations

Our analysis interprets Transformer text embeddings through a pooling lens, which cleanly exposes the geometry behind semantic smoothing. Although this abstraction matches common practice (mean/CLS-style pooling) and is empirically supported in our study, it does not attempt to model all fine-grained token interactions across layers.

We define the semantic shift via cosine-distance-based local evolution and global dispersion. Other reasonable choices (e.g., alternative similarity metrics, different window sizes, or discourse-aware weighting) could be plugged into the same framework and may further refine sensitivity in certain domains. Our goal is to establish a simple and computable measure that is stable across models and corpora, not to claim a unique definition.

For readability, the main text presents representative results on a subset of models/corpora, with additional experiments provided in the appendix. Although the observed trends are consistent across all tested settings (models and corpora), extending coverage to more languages and additional specialized domains would further broaden the empirical picture.

References

Mira Ait-Saada and Mohamed Nadif. 2023. Is anisotropy truly harmful? a case study on text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Jane Austen. Pride and prejudice (project gutenberg). https://www.gutenberg.org/ebooks/1342. Accessed 2025-12-24.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

Lewis Carroll. Alice's adventures in wonderland (project gutenberg). https://www.gutenberg.org/ebooks/11. Accessed 2025-12-24.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Team Common Pile and arXiv.org. 2023. arxiv abstracts dataset. https://huggingface.co/datasets/common-pile/arxiv_abstracts. Accessed: 2025-12-24.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Alejandro Fuster-Baggetto and Víctor Fresno. 2022. Is anisotropy really the cause of bert embeddings not being semantic? In *Findings of the association for computational linguistics: EMNLP 2022*, pages 4271–4281.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Marti A Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

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*, pages 238–244.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

LangChain. 2022. Langchain. https://github.com/langchain-ai/langchain.

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)*, pages 9119–9130.

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

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459.

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

LlamaIndex. 2022. Llamaindex. https://github.com/run-llama/llama_index.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

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

Belinda Mo, Kyssen Yu, Joshua Kazdan, Proud Mpala, Lisa Yu, Charilaos I. Kanatsoulis, and Sanmi Koyejo. 2025. KGGen: Extracting knowledge graphs from plain text with language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Niklas Muennighoff and 1 others. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

OpenAI. 2024. New embedding models and api updates. https://openai.com/blog/new-embedding-models-and-api-updates. Accessed: 2025-12-24.

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

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Project Gutenberg. Project gutenberg. https://www.gutenberg.org/. Accessed 2025-12-24.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

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)*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 world wide web conference*, pages 1003–1011.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 16857–16867.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 499–506.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 5065–5075.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Yuqi Zhou, Sunhao Dai, Zhanshuo Cao, Xiao Zhang, and Jun Xu. 2025. Length-induced embedding collapse in plm-based models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28767–28791.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Longembed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816.

A Extended Related Work

Text embeddings and dense representation learning. Learning vector representations for text has long been central to information retrieval and semantic matching. Early approaches focused on static distributional embeddings, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017). More recently, pretrained language models (PLMs), including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019), have enabled context-sensitive representations that significantly improve downstream performance. For retrieval and similarity-based tasks, dense bi-encoder architectures (Karpukhin et al., 2020; Izacard et al., 2022) have become a standard alternative to sparse lexical methods such as BM25 (Robertson and Zaragoza, 2009), while large-scale benchmarks like BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) reveal persistent challenges in robustness and generalization across domains.

Sentence embedding learning and representation geometry. A substantial body of work studies how to learn sentence-level embeddings that faithfully capture semantic similarity. Sentence-BERT (Reimers and Gurevych, 2019) introduced siamese architectures for efficient cosine-based retrieval, which were later enhanced by contrastive learning objectives. Representative methods include SimCSE (Gao et al., 2021), ConSERT (Yan et al., 2021), Mirror-BERT (Liu et al., 2021), PromptBERT (Jiang et al., 2022), DiffCSE (Chuang et al., 2022), and TSDAE (Wang et al., 2021). More recent embedding families, such as E5 (Wang et al., 2022), BGE (Xiao et al., 2024), and INSTRUCTOR (Su et al., 2023), aim to produce general-purpose representations aligned with diverse instructions and tasks.

Beyond accuracy, increasing attention has been paid to the *geometry* of embedding spaces. Prior work shows that contextual embeddings are often highly *anisotropic*, concentrating in a narrow cone rather than being uniformly distributed (Ethayarajh, 2019). Several mitigation strategies have been proposed, including removing dominant directions (Mu and Viswanath, 2018; Arora et al., 2017; Raunak et al., 2019), whitening-based normalization (Su et al., 2021; Huang et al., 2021), and flow-based transformations (Li et al., 2020). Recent analyses caution that global concentration metrics may not reliably predict semantic quality (Timkey and van Schijndel, 2021; Fuster-Baggetto and Fresno, 2022).

Long-context representations and length-induced collapse. Long texts pose a persistent challenge for embedding-based retrieval. Recent work formalizes this phenomenon as length-induced embedding collapse, attributing it to the low-pass filtering behavior of self-attention, where repeated attention operations progressively suppress high-frequency semantic variations and amplify dominant low-frequency components, causing representations of longer texts to become increasingly similar and less discriminative(Zhou et al., 2025). Complementary efforts benchmark long-context embeddings and retrieval (Zhu et al., 2024). Parallel advances in long-context modeling and efficient attention enable substantially longer sequences(Beltagy et al., 2020; Zaheer et al., 2020; Press et al., 2022; Dao et al., 2022). However, length alone does not fully explain retrieval difficulty: texts of identical length can exhibit vastly different embedding behaviors depending on how their semantics evolve internally.

Discourse structure, topic evolution, and semantic shift. The evolution of meaning in a document has been extensively studied in linguistics and discourse theory. Classical frameworks char-

acterize discourse coherence and structure (Grosz and Sidner, 1986; Mann and Thompson, 1988), while computational models operationalize coherence through entity transitions and local continuity (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013). Text segmentation and topic boundary detection formalize discourse evolution, with early unsupervised methods such as TextTiling (Hearst, 1997) and subsequent statistical approaches (Choi, 2000; Utiyama and Isahara, 2001; Galley et al., 2003; Malioutov and Barzilay, 2006; Eisenstein and Barzilay, 2008). Neural models also address segmentation and coherence with supervised or hierarchical architectures (Koshorek et al., 2018).

A related literature studies semantic change over time using temporal embeddings and alignment techniques (Hamilton et al., 2016; Kutuzov et al., 2018; Rudolph and Blei, 2018; Yao et al., 2018). Although studies focus on evolution across corpora or time periods, they share a key insight: semantic variation is best understood as a process with measurable rates, rather than as a single static distance.

B Embedding Models, Corpora, and Preprocessing

This appendix describes the embedding models and corpora used throughout our experiments, together with the unified preprocessing pipeline used to convert each corpus into an ordered sentence sequence $S = (s_1, \ldots, s_n)$.

B.1 Embedding models

We consider a diverse set of embedding models that span open-source and proprietary systems, covering different training paradigms and embedding-space geometries. Table 1 summarizes their key characteristics, including the dimension of the output, the architectural foundation, and the information of the release.

B.2 Corpora

Table 2 summarizes all corpora used in our experiments. These corpora cover heterogeneous discourse regimes (technical abstracts, encyclopedic entries, knowledge essays, and long-form narratives), enabling us to analyze semantic shifts under substantially different topic-evolution patterns. Unless otherwise noted, we split each document into sentences, preserve the original order, and concatenate all sentences into a single ordered sequence $S = (s_1, \ldots, s_n)$. For very large corpora (ArXiv and Wikipedia), we restrict to the first 5,000 documents in the dataset in order to control runtime and improve reproducibility.

B.3 Preprocessing and sentence sequence construction

In all corpora, we convert the raw text into an ordered sentence sequence S using the same pipeline.

Text cleaning. Given a raw text string, we apply a lightweight cleaning function that: (i) strips noisy characters at both ends while deliberately preserving sentence-final punctuation to avoid breaking sentence boundary detection; and (ii) merges repeated whitespace into a single space.

Sentence segmentation. We split each document into sentences via nltk.tokenize.sent_tokenize, which is based on the Punkt sentence tokenizer (Kiss and Strunk, 2006). After splitting, empty sentences are removed, and each sentence is re-cleaned. This yields a list of sentences for each document.

Preserving order and forming S. For each corpus, we preserve the original document order, and, within each document, preserve the original sentence order. We then concatenate all sentence lists into one global ordered sequence $S = (s_1, \ldots, s_n)$. For corpora that naturally consist of many documents (ArXiv, Wikipedia, MINE), this produces a long sequence whose local neighborhoods reflect within-document coherence, while global transitions reflect the dataset's document ordering.

C Comparing Mean Pairwise Distance (MPD) Across Corpora and Embedding Models

This section complements the simplified illustration in Figure 1 (main paper) by reporting the corpus-level MPD statistics for all models and corpora used throughout the paper, and by providing a more careful interpretation of what MPD does—and does not—reveal about embedding geometry and downstream retrieval.

C.1 Metric and computation protocol

Given a corpus-specific ordered sentence sequence $S = (s_1, \ldots, s_n)$ constructed by the preprocessing pipeline in Section B.3, we embed each sentence s_i using an embedding model (Section B.1) to obtain unit-normalized sentence embeddings $\{e_i\}_{i=1}^{n}$. We

Model (full name)	Abbreviation	Architecture / key traits	Dim.	Provider	License	References
bge-large-en-v1.5	bge-large	Transformer bi-encoder for dense retrieval; contrastive training with strong general-purpose embedding behavior.	1024	BAAI	Open source	(Xiao et al., 2024; Chen et al., 2024)
e5-large-v2	e5-large	Transformer bi-encoder trained via weakly-supervised contrastive pretraining; query/passage style prompting is commonly used in the E5 family.	1024	Microsoft	Open source	(Wang et al., 2022)
all-mpnet-base-v2	all-mpnet	Sentence-Transformers bi-encoder built on MPNet-base; mean pooling for sentence embeddings; widely used strong baseline.	768	Sentence-Transformers	Open source	(Reimers and Gurevych, 2019; Song et al., 2020)
gte-large	gte-large	General Text Embeddings (GTE); Transformer encoder optimized for retrieval-style embedding.	1024	Alibaba	Open source	(Li et al., 2023)
text-embedding-3-large	text-embedding	Proprietary API embedding model; high-dimensional embeddings designed for general semantic matching and retrieval.	3072	OpenAI	Closed source	(OpenAI, 2024)

Table 1: Embedding models used in our experiments. "Dim." denotes the output embedding dimensionality.

Corpus (full name)	Abbreviation	Characteristics	References
ArXiv abstracts (common-pile/arxiv_abstracts)	ArXiv	Large-scale scientific paper abstracts. Highly technical, information-dense, and relatively short documents with strong domain-specific terminology. We use the first 5000 abstracts in dataset order.	(Common Pile and arXiv.org, 2023).
Alice's Adventures in Wonderland (Project Gutenberg)	Alice	Single long-form narrative novel (fiction). Natural discourse progression with plot-driven topic transitions and stylistic variation. We treat the entire book as one document and keep the original reading order.	(Project Gutenberg; Carroll).
Pride and Prejudice (Project Gutenberg)	Pride	Single long-form narrative novel (fiction). Long-range thematic development and chapter-level transitions. We treat the entire book as one document and keep the original reading order.	(Project Gutenberg; Austen).
MINE essays (kyssen/kg-gen-evaluation-essays)	MINE	A collection of knowledge-focused essays (multi-document). Each essay is relatively short and expository, often exhibiting clearer local coherence than narratives. We preserve dataset order and concatenate essays to form S.	(Mo et al., 2025).
Wikipedia (English) (google/wiki40b)	Wikipedia	Encyclopedic articles spanning broad topics; expository style with frequent entity/topic changes across documents. We use the first 5,000 documents in dataset order.	(Guo et al., 2020).

Table 2: Corpora used in our experiments. We cover technical, encyclopedic, essay-style, and narrative texts. For large multi-document corpora (ArXiv and Wikipedia), we use the first 5000 documents for efficiency and reproducibility.

then compute the mean pairwise cosine distance (MPD):

$$\mathrm{MPD}(S) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \big(1 - \cos(e_i, e_j)\big), \tag{23}$$

where a smaller MPD indicates a more concentrated (more anisotropic) embedding distribution, while a larger MPD indicates more dispersed sentence embeddings.

The main paper (Figure 1) plots an incremental version of this statistic by computing MPD over the first $1, 2, \ldots, n$ sentences. The plateau observed there motivates a practical summary statistic: the converged MPD value when n is sufficiently large. Table 3 reports this corpus-level MPD computed

Corpus	bge-large	e5-large	all-mpnet	gte-large	text-embedding	Avg.
ArXiv	0.505	0.231↓	0.865↑	0.249	0.789	0.528
Alice's Adventures	0.582	0.232	0.747↑	0.202↓	0.697	0.492
Pride and Prejudice	0.577	0.240	0.774↑	0.209↓	0.718	0.504
MINE	0.575	0.253↓	0.903↑	0.256	0.882	0.574
Wikipedia	0.641	0.293	0.897	0.261↓	0.900↑	0.598
Avg. (across corpora)	0.576	0.250	0.837	0.235	0.797	0.539

Table 3: Mean pairwise cosine distance (MPD) of sentence embeddings across corpora and embedding models. Larger MPD indicates more dispersed sentence embeddings; smaller MPD indicates stronger concentration. The last column reports the average MPD across models for each corpus, and the last row reports averages across corpora for each model.

on all sentences in each corpus after preprocessing. For large multi-document corpora (ArXiv and Wikipedia), we follow Section B.2 and restrict to the first 5000 documents for efficiency and reproducibility.

C.2 Results: model dependence vs. corpus dependence

Table 3 shows that MPD varies substantially across both models and corpora, but qualitatively different ways.

Model dependence dominates the absolute MPD scale. Keeping the corpus fixed, different embedding models can yield dramatically different MPD values. For example, on ArXiv, MPD ranges from 0.231 (e5-large) to 0.865 (all-mpnet), a gap of ≈ 0.634. Similar gaps appear on Wikipedia (from 0.261 to 0.900, gap ≈ 0.639) and MINE (from 0.253 to 0.903, gap ≈ 0.650). This agrees with Figure 1: even when the MPD curves stabilize as n increases, their converged levels are strongly model-dependent.

A consistent ranking also emerges in the averaged row of Table 3: gte-large and e5-large tend to produce the most concentrated sentence embeddings (lowest MPD), bge-large is intermediate, while text-embedding and all-mpnet are substantially more dispersed (higher MPD).

Corpus dependence reflects discourse regime, but with smaller range. Keeping the model fixed, MPD still varies across corpora, indicating that sentence-level semantic diversity differs by domain. However, the range within the model is typically much smaller than the range across the model. For example, under bge-large, MPD ranges from 0.505 (ArXiv) to 0.641 (Wikipedia), a spread of 0.136; under e5-large, the spread is 0.062 (from 0.231 to 0.293); under gte-large, the spread is 0.059 (from 0.202 to 0.261). This pattern suggests that while corpus semantics shape dispersion, the global geometry induced by the embedding model largely determines the overall MPD scale.

At the corpus level, the last column of Table 3 indicates that Wikipedia and MINE have a higher average MPD than the two novels, which is consistent with their broader topical coverage and frequent changes between documents. In contrast, long-form narratives (Alice, Pride) tend to maintain stronger global continuity and recurring entities/themes, which typically reduces global dispersion.

C.3 Discussion: what MPD can (and cannot) explain

MPD is a geometry descriptor, not a performance predictor. MPD (and related anisotropy/concentration measures) summarizes the global spread of embeddings, but it does not directly determine retrieval quality. This is consistent with the paradox highlighted in the main paper: models with very different MPD (e.g., e5-large vs. all-mpnet) can still achieve broadly comparable performance on practical downstream tasks. In other words, the absolute concentration level alone is insufficient to explain when embedding-based retrieval becomes difficult.

Why can different models yield drastically different MPD? The strong model dependence of MPD suggests that it is not merely a property of the corpus. Different training objectives, data mixtures, embedding dimensions, pooling implementations,

and normalization conventions can induce different global angular distributions (i.e. different degrees of anisotropy) even on identical inputs. Therefore, comparing MPD values across models mainly reveals differences in embedding-space geometry, not necessarily differences in semantic fidelity.

Implication for our paper. Taken together, Table 3 and Figure 1 motivate the central question of this paper: if global concentration statistics can vary widely across models and yet do not consistently predict downstream behavior, what content-driven factor explains when embeddings become less discriminative? This motivates our semantic shift perspective in the main paper: instead of treating concentration as the root cause, **we examine how structured semantic evolution within text (semantic shift) interacts with pooling/smoothing mechanisms to produce collapse and retrieval degradation.**

Table 3 should be read as a diagnostic snapshot of the embedding geometry induced by each model in each discourse regime. Its main message is not that "low MPD is bad" or "high MPD is good", but that MPD is strongly model-dependent and therefore cannot by itself serve as a universal explanation of retrieval difficulty. This observation sets the stage for the controlled semantic-shift experiments analyzed in subsequent sections.

D Further Analysis of Transformer-Based Embedding Models and Extended Experiments on Theorem 1

This appendix extends the empirical validation of Theorem 1 to five embedding models and five corpora (summarized in Section B). Our goal is to test the central claim under a realistic encoding pipeline: **even when a multi-sentence text is encoded directly by a Transformer encoder (instead of being explicitly averaged over sentence embeddings), sentence-level semantic diversity still monotonically increases the discrepancy between the text embedding and its constituent sentence embeddings.**

D.1 Protocol: controlling sentence-level semantic diversity

We follow the unified preprocessing pipeline in Section B to convert each corpus into an ordered sentence sequence. We then fix the group size to k=10 and construct sentence groups under three controlled diversity regimes: *local* (consecutive sentences within a document), *medium* (non-adjacent sentences within a document) and *high* (sentences sampled uniformly from the corpus). This design varies sentence-level semantic diversity while holding k fixed, allowing a direct test of the monotonicity predicted by Theorem 1 beyond idealized pooling.

D.2 Metrics and evaluation

For each sampled group $(s_1, \ldots, s_k)$, we compute sentence embeddings $(e_1, \ldots, e_k)$ by encoding each sentence separately, and compute a text embedding z by concatenating the k sentences (with standard separators) and encoding the resulting multi-sentence text once. We then measure:

$$C_{\text{pair}} = \frac{2}{k(k-1)} \sum_{i<j} \left(1 - \cos(e_i, e_j)\right),$$

$$C_{\text{mean}} = \frac{1}{k} \sum_{i=1}^{k} \left(1 - \cos(e_i, z)\right).$$

C_{pair} quantifies sentence-level semantic diversity, while C_{mean} quantifies how much the encoded text representation deviates from its constituent sentences (semantic dilution).

To quantify monotonic dependence without assuming linearity, we report Spearman's rank correlation ρ and Kendall's τ between C_{pair} and C_{mean} for each corpus–model pair.

D.3 Results: Strong Cross-Model and Cross-Corpus Monotonicity

Figure 6 reports scatter plots of C_{mean} versus C_{pair} for each corpus, with points stratified by the three diversity regimes. Across all corpora and embedding models, we observe a clear monotonic trend: higher sentence-level semantic diversity (C_{pair}) consistently yields larger text–sentence discrepancy (C_{mean}), matching the qualitative behavior predicted by Theorem 1.

To quantify monotonic dependence without assuming linearity, we compute Spearman's ρ and Kendall's τ for each corpus-model pair. The results are summarized in Table 4. Correlations are uniformly high: *Spearman* ρ *ranges from* 0.82 *to* 0.99 *and Kendall* τ *ranges from* 0.63 *to* 0.91 *across all settings.* Notably, the knowledge-oriented MINE corpus exhibits near-saturated correlations across all models (e.g., $\rho \geq 0.97$), while long-form narratives (Alice / Pride) remain strongly monotonic but slightly noisier—consistent with the fact that narrative texts contain richer discourse phenomena (e.g.,

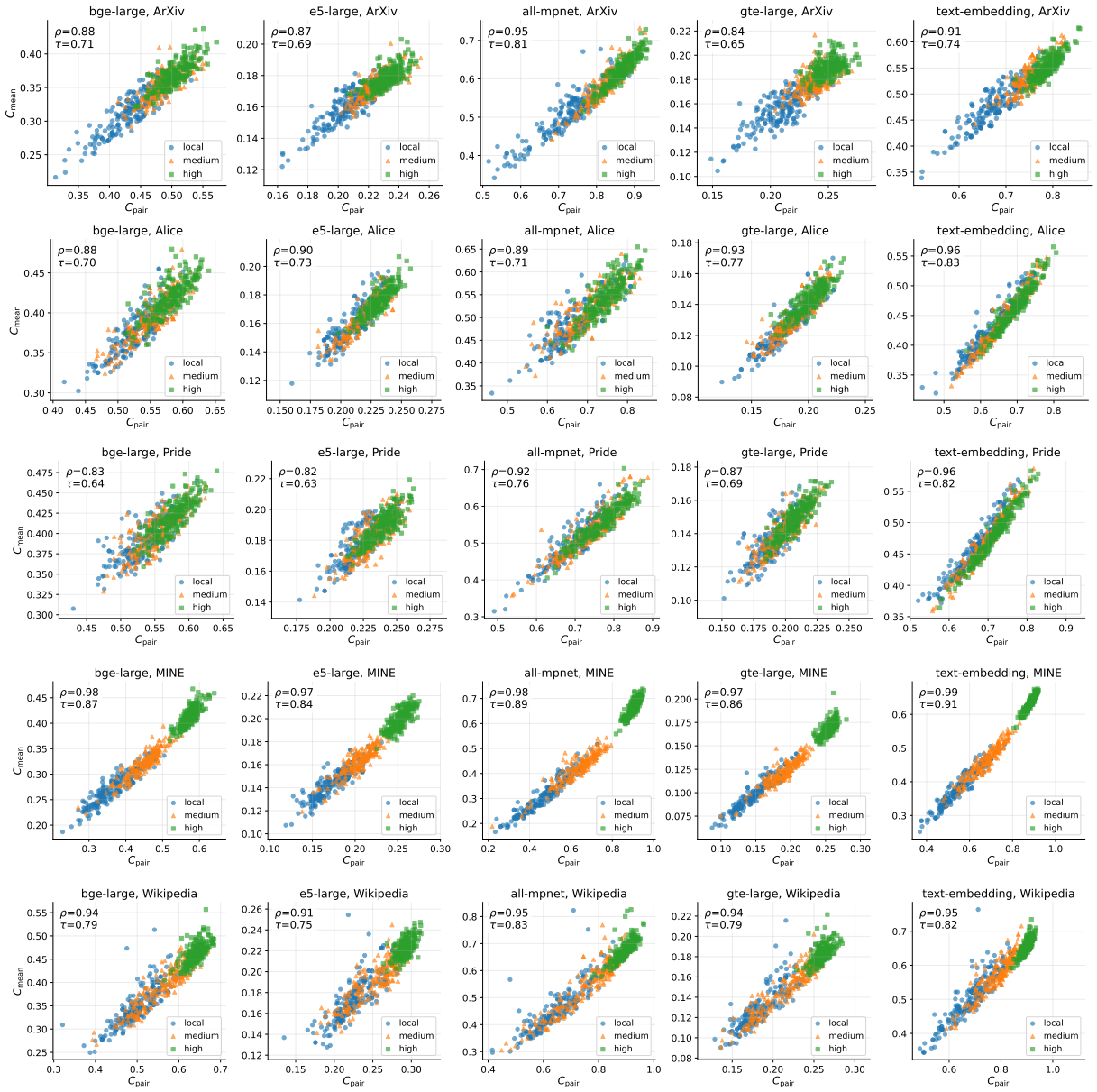

Figure 6: Empirical verification of Theorem 1 across five corpora and five embedding models. Each subplot shows C_{mean} (text–sentence discrepancy) versus C_{pair} (sentence-level semantic diversity) under three controlled diversity regimes: *local*, *medium*, and *high*. Rank correlations (Spearman's ρ, Kendall's τ) are reported in each subplot.

gradual topic drift, character/event re-entrance) that can introduce additional variability in embedding behavior.

Why scale differences do not affect the conclusion. We emphasize that the absolute magnitudes of C_{pair} and C_{mean} can vary substantially across models due to differences in embedding-space geometry (e.g., global angular concentration, normalization conventions, and training objectives). This is precisely why we report rank-based statistics (Spearman/Kendall): they are invariant to monotone rescaling and directly test the theoretical prediction of monotonicity. Therefore, model-dependent scale differences do not change the conclusion that semantic diversity reliably drives semantic dilution under practical Transformer encoders.

In general, the extended results in Figure 6 and Table 4 confirm that Theorem 1 captures a robust property of real embedding models and diverse corpora, rather than a peculiarity of a specific architecture, dataset, or idealized pooling assumption. Sentence-level semantic diversity consistently induces larger text–sentence discrepancy even under direct encoding of concatenated text, providing a solid empirical foundation for the semantic shift

Corpus	bge-large	e5-large	all-mpnet	gte-large	text-embedding
ArXiv	ρ=0.88, τ=0.71	ρ=0.87, τ=0.69	ρ=0.95, τ=0.81	ρ=0.84, τ=0.65	ρ=0.91, τ=0.74
Alice	ρ=0.88, τ=0.70	ρ=0.90, τ=0.73	ρ=0.89, τ=0.71	ρ=0.93, τ=0.77	ρ=0.96, τ=0.83
Pride	ρ=0.83, τ=0.64	ρ=0.82, τ=0.63	ρ=0.92, τ=0.76	ρ=0.87, τ=0.69	ρ=0.96, τ=0.82
MINE	ρ=0.98, τ=0.87	ρ=0.97, τ=0.84	ρ=0.98, τ=0.89	ρ=0.97, τ=0.86	ρ=0.99, τ=0.91
Wikipedia	ρ=0.94, τ=0.79	ρ=0.91, τ=0.75	ρ=0.95, τ=0.83	ρ=0.94, τ=0.79	ρ=0.95, τ=0.82

Table 4: Extended empirical validation of Theorem 1 across five corpora and five embedding models. We report Spearman's ρ and Kendall's τ between C_{pair} and C_{mean}. High values across all settings indicate a robust monotonic relationship.

perspective developed in the main paper.

E Additional Results: Semantic Shift vs. Length Collapse Across Embedding Models

This appendix complements Sec. 4.1 by reporting results on two additional embedding models (e5-large and all-mpnet) beyond the main-embedding model (bge-large). Across all three models and both corpora (ArXiv and Alice's Adventures in Wonderland), we observe highly consistent qualitative patterns: (i) embedding concentration (measured by MPD) strengthens as the constructed text units become longer, but the magnitude of this effect depends primarily on how strongly semantics are mixed within each unit; and (ii) our semantic-shift metric measured on the same constructed units tracks the severity of MPD reduction almost monotonically. These results support the claim that semantic shift is a model-robust explanatory variable for when length collapse and anisotropy become severe.

Figures. For each model, we report (1) MPD under the three concatenation patterns (repeat, sequential, random) for S, $S2$, $S5$, and $S10$, and (2) mean semantic shift measured on the $S10$ variant across hop distances $1 \ldots 9$. Figures 7 and 8 summarize the complete set of results.

(1) MPD results are consistent across models. Across all models, MPD decreases from $S \rightarrow S10$ in all concatenation patterns (Fig. 7), confirming that lengthening tends to increase concentration. However, the rate and extent of MPD reduction depend strongly on how semantics are composed within each constructed unit: *repeat* produces the mildest MPD drop, *sequential* produces a noticeably larger drop, and *random* produces the sharpest decline, indicating the strongest concentration. This ordering (Repeat < Seq < Rand in collapse severity) holds on both corpora for all models (Figs. 7).

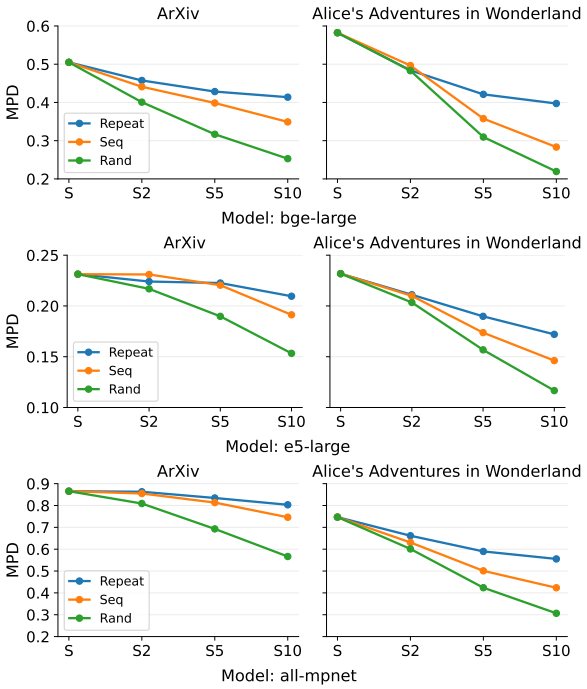

Figure 7: Embedding concentration measured by MPD under different concatenation patterns on ArXiv and *Alice's Adventures in Wonderland*. Lower MPD indicates stronger concentration (i.e., more severe length collapse / anisotropy).

In addition, the absolute MPD level is clearly model-dependent. e5-large operates in a substantially more concentrated regime overall (lower MPD throughout), while all-mpnet is the least concentrated (higher MPD), and bge-large lies in between. This reproduces a familiar empirical fact: different embedding families can exhibit markedly different global geometry (e.g., anisotropy), even when their downstream performance is broadly comparable. However, critically, these baseline differences do not change the central pattern: *semantic diversity consistently amplifies concentration far more than pure lengthening via repetition*.

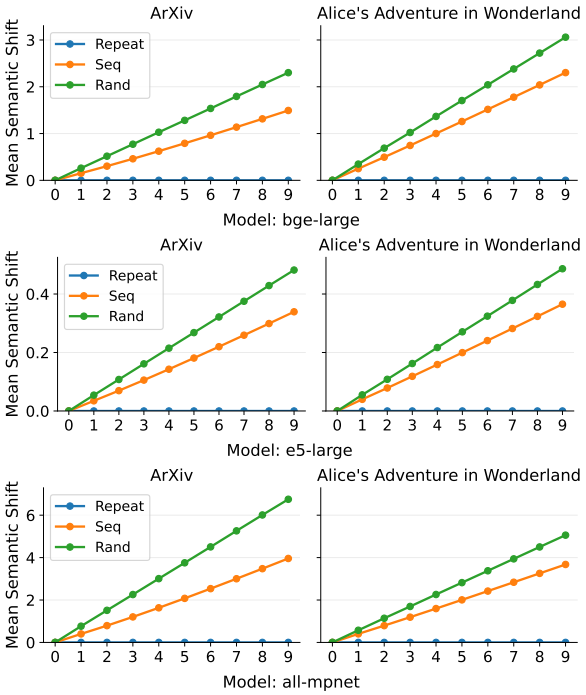

Figure 8: Mean semantic shift on the $S10$ variant as a function of hop distance under repeat/sequential/random concatenation. Across models, repeat yields zero shift, sequential yields moderate shift, and random yields the largest shift.

(2) Semantic shift curves show the same ranking across models. Figure 8 reports the mean semantic shift in the variant $S10$ across hop distances. Three robust patterns emerge across all models and both corpora: (i) Repeat yields zero shift across hop distances, as expected because the constructed units preserve the same sentence semantics and only increase length; (ii) Sequential and random shifts increase with hop distance, indicating that semantic divergence accumulates progressively as we move farther along the sequence; and (iii) Random tends to exceed sequential most clearly on ArXiv, reflecting stronger semantic heterogeneity induced by global mixing (Figs. 8).

For Alice's Adventures in Wonderland, the gap between random and sequential becomes smaller than that on ArXiv across all models (Fig. 8). This is consistent with the narrative nature of the corpus: even local sequential windows naturally include topic transitions and plot development, so sequential concatenation already induces non-trivial within-unit semantic evolution, partially closing the gap to random mixing.

(3) Semantic shift explains MPD reduction better than length alone. Comparing Figs. 7 and 8 reveals a consistent alignment: patterns with larger measured semantic shift (Seq/Rand) also produce stronger MPD reductions, while repeat concatenation produces zero semantic shift and only mild concentration. Importantly, this alignment persists across: **(a)** embedding models with very different baseline geometry (overall MPD levels), and **(b)** corpus types with different discourse properties (technical articles vs. long-form narrative). Therefore, the expanded results strengthen the main-text conclusion: *the severity of length collapse / anisotropy is primarily controlled by the strength of within-unit semantic shift, rather than by length itself.*

Cross-model invariance. A particularly salient observation is that the e5-large model is globally more anisotropic (lower MPD across all settings), which could prima facie suggest that "anisotropy alone" should predict retrieval difficulty. However, across bge-large, e5-large, and all-mpnet, we consistently observe the same monotonic relationship: **larger semantic shift ⇒ larger MPD reduction (stronger collapse)**. In other words, even when a model starts from a more concentrated geometry, *semantic shift still governs how rapidly embeddings further collapse as we inject semantic heterogeneity*. This invariance indicates that semantic shift is not a model-specific artifact; rather, it functions as a stable explanatory variable that generalizes across embedding families with different training objectives and baseline anisotropy.

Observed scaling differences. While the ranking of semantic shift is consistent, the absolute scale of the shift values can differ across models (Fig. 8). This is expected, because our semantic shift metric is computed from cosine-based distances between embeddings, and different encoders induce different global angular distributions due to architectural and training choices (e.g., normalization conventions, contrastive temperature/regularization, and how aggressively the representation space is "compressed" around dominant directions). As a result, the same underlying semantic transition in text may correspond to a larger or smaller cosine-distance change depending on the model's intrinsic geometry. Crucially, our claims in this section rely on *within-model, across-pattern comparisons* (Repeat vs. Seq vs. Rand on the same corpus), where the metric is applied under a fixed encoder. Under this controlled setting, the relative ordering and monotonic alignment between shift

and MPD reduction remains stable, making the conclusion robust to cross-model scale differences.

Implications. Taken together, these additional results clarify two points that are easy to conflate: (1) baseline anisotropy (global MPD level) is model-dependent and does not by itself determine when embeddings become unreliable; and (2) the incremental collapse induced by lengthening is strongly modulated by the degree of semantic mixing inside each unit, which is captured by semantic shift. Therefore, semantic shift offers a more predictive lens for diagnosing when long-text embeddings will collapse (and when anisotropy is likely to translate into retrieval failures), beyond explanations that attribute collapse primarily to length alone.

Design implications and real-world correspondence. The three controlled concatenation patterns in Section 4.1 are not merely synthetic stress tests; they closely mirror how long "documents" arise in practice. **Repeat concatenation** approximates long inputs with high redundancy (e.g., templated pages, boilerplate-heavy documents, repetitive logs, or duplicated passages), where length increases without introducing new semantic components. **Sequential concatenation** resembles organically long documents (e.g. academic papers, books, or well-edited articles) in which content evolves through locally coherent discourse, introducing semantic change gradually. In contrast, **random concatenation** serves as a proxy for heterogeneous aggregation commonly produced by real pipelines: concatenating multiple sources into a single context window (multi-page PDFs, scraped web pages with sidebars and unrelated blocks, forum threads, stitched meeting notes, or retrieval-augmented prompts that combine snippets from different topics). These settings differ less in length than in within-unit semantic heterogeneity, precisely the factor captured by the semantic shift.

This mapping helps to clarify why "length alone" is an incomplete predictor of collapse and downstream degradation. If length-induced embedding collapse were primarily a function of sequence length, then all long inputs of comparable length should degrade similarly. Our results instead show that long inputs can be relatively benign when semantic shift is weak (Repeat; and sometimes Sequential on structured corpora), but can collapse severely when semantic shift is strong (Random; and Sequential on narrative texts with frequent topic transitions). Therefore, the operational driver behind collapse in realistic workloads is often not just the token budget, but how many distinct semantic components are mixed into the same embedding unit and how fast these components evolve across the text.

From a system-design perspective, this suggests that mitigation strategies should be shift-aware rather than purely length-aware. For example, chunking policies in retrieval or RAG pipelines are often tuned by length heuristics (fixed token windows or simple overlap). Our findings imply that such heuristics can be suboptimal: they may unnecessarily split redundant but coherent spans (low shift) while failing to separate heterogeneous spans (high shift) that are most likely to collapse. Instead, semantic-shift signals can be used to identify semantic boundary points where topic transitions accelerate, which are precisely the locations where aggregation is most harmful. More broadly, semantic shift provides a principled diagnostic for long-text embedding reliability: documents with a high shift within the unit should be decomposed, indexed, or retrieved at finer granularity, whereas low-shift documents can tolerate larger units without substantial collapse. This perspective also reconciles why embeddings of identical length can exhibit widely different concentration behaviors (Fig. 7) and why anisotropy is not uniformly harmful across settings: it becomes most damaging when it is induced by strong semantic shift rather than by lengthening.

F Additional Retrieval Results Across Models and Corpora

This appendix extends Sec. 4.2 by reporting self-overlap@k results for three embedding models (bge-large, e5-large, all-mpnet) on two corpora (ArXiv and Alice's Adventures in Wonderland). The main text shows only bge-large on ArXiv for brevity. Here, we demonstrate that the key conclusion is invariant across models and corpora: *anisotropy becomes harmful primarily when induced by strong semantic shift (sequential/random mixing), whereas anisotropy caused by lengthening (repeat) has limited impact on retrieval robustness.*

Figures. Figures 9, 10 and 11 summarize the complete set of results. Each subfigure reports average self-overlap@k ($k \in \{1, 3, 5\}$) between the retrieval of the original corpus S and its concatenated variants ($S2$, $S5$, $S10$) in repeat/sequential/random patterns.

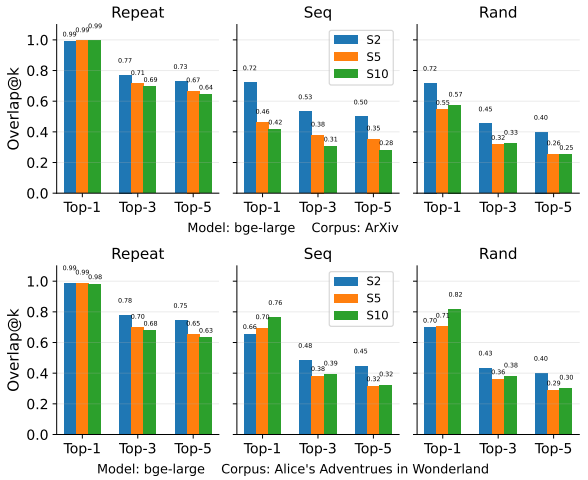

Figure 9: (Model: bge-large) Average self-overlap@k between retrieval results on the original corpus S and its concatenated variants ($S2$, $S5$, $S10$) under repeat, sequential, and random patterns. Higher overlap indicates stronger semantic preservation and less retrieval damage.

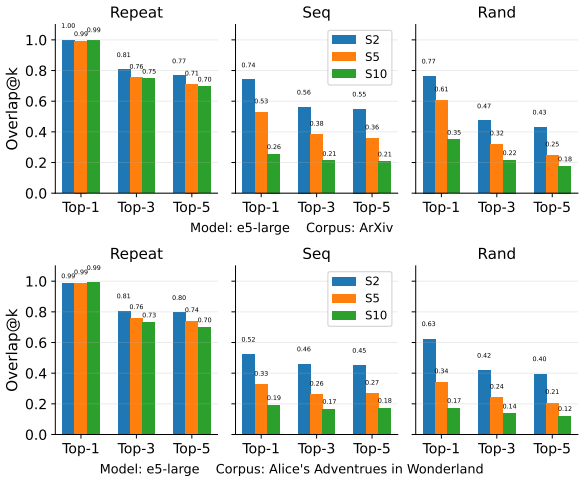

Figure 10: (Model: e5-large) Average self-overlap@k between retrieval results on the original corpus S and its concatenated variants ($S2$, $S5$, $S10$) under repeat, sequential, and random patterns. Higher overlap indicates stronger semantic preservation and less retrieval damage.

(1) Repeat concatenation: benign anisotropy with minimal retrieval damage. Across all three models and both corpora, the repeat pattern consistently yields the highest overlap and remains stable as we move from $S2$ to $S10$. In particular, Overlap@1 is essentially preserved (typically ≈ 0.98–1.00) in all settings, and Overlap@3/5 also stays relatively high (often ≈ 0.7–0.8). This supports the main-text claim: *anisotropy/concentration induced mainly by lengthening (without semantic diversification) tends to preserve relative neighborhoods and thus has limited impact on retrieval robustness.*

(2) Sequential concatenation: retrieval degrades as the semantic window expands. Under sequential concatenation, overlap decreases substantially and typically worsens with longer windows ($S2 \rightarrow S10$), especially for larger k. On ArXiv, the trend is clear across models: for bge-large, Overlap@1 drops to roughly $0.72 \rightarrow 0.42$ from $S2$ to $S10$, and Overlap@5 drops to roughly $0.50 \rightarrow 0.28$ (Fig. 9); for e5-large, the degradation is even sharper on larger-k neighborhoods (e.g., Overlap@5 around $0.55 \rightarrow 0.21$; Fig. 10); and all-mpnet exhibits a similar monotone deterioration (Fig. 11). On Alice, sequential concatenation remains harmful across all models as well (Figs. 9, 10, 11), consistent with the fact that narrative discourse can accumulate semantic change even within locally adjacent spans, so enlarging the sequential window injects increasing within-unit semantic variation.

(3) Random concatenation: strongest retrieval damage. Random concatenation consistently yields the lowest overlap@k and the fastest degradation with window size. On ArXiv, bge-large shows Overlap@1 decreasing from about 0.72 (S2) to ≈ 0.57 (S10), and Overlap@3/5 similarly collapsing (Fig. 9); e5-large shows substantial drops especially for larger k (e.g., Overlap@5 reaching ≈ 0.18 on S10; Fig. 10); and all-mpnet follows the same pattern (Fig. 11). On Alice, random remains highly damaging across models,with `e5-large` showing particularly low overlap for larger-k neighborhoods (Fig. 10). Overall, random mixing, which maximizes within-unit semantic heterogeneity, produces the most severe loss of neighborhood preservation, aligning with our thesis that retrieval failure is driven by *semantic shift* rather than concentration alone.

Cross-model invariance. Although models differ in baseline anisotropy (e.g., e5-large is often more concentrated globally than all-mpnet), Figures 9, 10, and 11 show stronger and more general regularity. **across all three models, the ordering of retrieval robustness is consistent:**

$$\text{Repeat} > \text{Sequential} > \text{Random}.$$

That is, even when a model starts from a more anisotropic embedding space, what determines retrieval degradation under lengthening is how se-

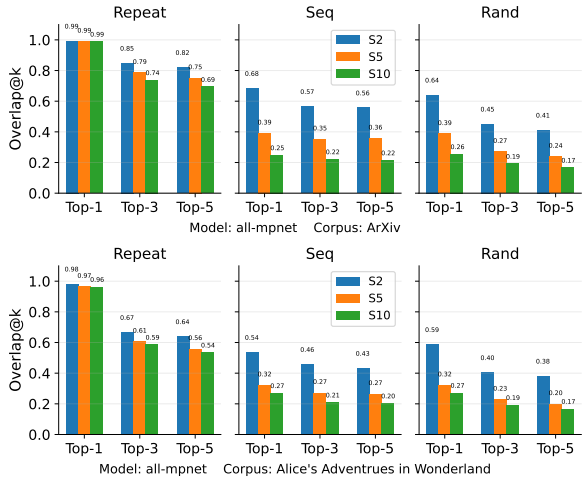

Figure 11: (Model: all-mpnet) Average self-overlap@k between retrieval results on the original corpus S and its concatenated variants ($S2$, $S5$, $S10$) under repeat, sequential, and random patterns. Higher overlap indicates stronger semantic preservation and less retrieval damage.

mantic content is mixed within each embedded unit. This invariance mirrors our concentration results (MPD/shift) and supports semantic shift as a model-agnostic driver of when anisotropy becomes harmful.

Across all embedding models and all corpora, the retrieval experiments consistently support: (i) length-only induced concentration (repeat) is largely benign for retrieval robustness; (ii) shift-inducing transformations (sequential/random) substantially disrupt nearest-neighbor rankings; and (iii) the strength of retrieval degradation correlates with how much semantic shift is injected within each unit, providing a principled explanation for why anisotropy is not uniformly harmful and when it becomes problematic.

G Semantic Shift Splitter: From Analysis to a Practical Segmenter

G.1 Motivation: Turning Semantic Shift into a Segmentation Signal

The main paper characterizes the semantic shift as a systematic shift of embedding representations as text grows, driven jointly by (i) local semantic transitions between adjacent units and (ii) the global dispersion among all units within the same context. Beyond serving as a diagnostic lens for embedding-based retrieval, this shift signal can be directly operationalized for sentence-level segmentation.

Chunking is a core primitive in retrieval-augmented generation (RAG): an effective chunker should (a) place boundaries near meaningful topic/section transitions and (b) produce chunks with controllable granularity and stable size, since excessive size variance can lead to a mixture of tiny fragments (weak evidence) and oversized passages (token-inefficient and harder to rank). These requirements motivate a Semantic Shift Splitter, which forms a chunk online and cuts precisely when the accumulated shift of the current segment indicates that continuing would create a semantically unstable (or internally dispersed) chunk.

G.2 Principle: Semantic Shift within a Candidate Segment

We segment a document at the sentence level. Let a document be a sequence of sentences $\{s_1, \ldots, s_n\}$ and let e_i denote the embedding of s_i (we use bge-large in all experiments). For a candidate segment containing k ordered sentences, we reuse the semantic-shift definition from Definition 3:

$$\mathrm{Shift}(k) \;=\; \mathrm{Local}(k) \cdot \mathrm{Disp}(k).$$

Here, $\mathrm{Local}(k)$ measures ordered stepwise shift across adjacent sentences, while $\mathrm{Disp}(k)$ measures global semantic spread among all sentences within the segment. Their product amplifies segments that are simultaneously (i) shifting along the reading order and (ii) internally dispersed, matching the instability patterns highlighted in the main paper. This makes $\mathrm{Shift}(\cdot)$ a natural boundary signal: when adding the next sentence sharply increases shift, the current segment is likely crossing a semantic transition.

G.3 Algorithm: Shift-Aware Online Chunking with Adaptive Threshold

The splitter constructs chunks left-to-right. Starting from an empty chunk, it appends sentences one by one. Before appending sentence s_i, it evaluates the hypothetical shift $\mathrm{Shift}(|C|+1)$; if this value exceeds a threshold τ, it cuts before appending and starts a new chunk at s_i. We also enforce a hard token cap to avoid overly long chunks, which is essential in RAG. See Algorithm 1 for algorithm details.

Adaptive threshold estimation. We estimate τ per document rather than fixing it globally. For each position t, we compute the shift of a local

Algorithm 1 Semantic Shift Splitter

Require: sentences $\{s_i\}_{i=1}^n$, embeddings $\{e_i\}_{i=1}^n$, percentile p, token cap T, min sentences per chunk m

Estimate τ by window shifts: $\tau \leftarrow \text{Percentile}(\{\hat{s}_t\}_{t=1}^n, p)$
Initialize empty current chunk $C \leftarrow [\,]$ and state for Local, Disp
for $i = 1$ to n **do**
  if $|C| \geq 1$ **and** tokens(C)+tokens(s_i) $> T$ **then**
    output C; reset $C \leftarrow [\,]$
  end if
  if $|C| \geq m$ **then**
    compute hypothetical Shift($|C|+1$) if appending s_i
    if Shift($|C|+1$) $> \tau$ **then**
      output C; reset $C \leftarrow [\,]$
    end if
  end if
  append s_i into C and update state
end for
if $C \neq [\,]$ **then**
  output C
end if

window of embeddings with radius b, producing window-shift values $\{\hat{s}_t\}_{t=1}^n$, and set

$$\tau \;=\; \text{Percentile}\big(\{\hat{s}_t\}_{t=1}^n,\, p\big),$$

where p (shift_percentile) controls how aggressively we cut: smaller p yields a smaller τ and thus more boundaries. This document-adaptive threshold makes the splitter robust to differences in writing style, length, and topical density.

Efficiency. A naive Disp(k) computation is $O(k^2)$, but the online construction admits incremental updates: when adding a new sentence, we only compute similarities between the new embedding and the embeddings already in the current chunk, yielding $O(k)$ per step. In practice, k is bounded by both the shift threshold and the token cap, which makes the method efficient for typical chunk sizes.

G.4 Experimental Setup and Fair Comparison Protocol

Datasets. We evaluate on two paragraph-annotated sources that reflect different discourse structures and segmentation cues.

(1) ArXiv Abstracts. We use scientific abstracts from ArXiv (Common Pile and arXiv.org, 2023), where ground-truth boundaries are defined by natural paragraph breaks. This setting emphasizes fine-grained rhetorical and topical transitions in compact, information-dense text (Table 5).

(2) MINE (KG-Gen Evaluation Essays). We use the essay dataset MINE(Mo et al., 2025). Each instance is a short essay consisting of multiple paragraphs; we treat the paragraph boundaries as the ground-truth segmentation. Compared with ArXiv, MINE contains more narrative/expository transitions, providing a complementary testbed for semantic chunking beyond scientific writing.

For both datasets, we segment at the sentence level and define a gold boundary whenever a paragraph break occurs between two consecutive sentences.

Baselines. We compare our proposed Semantic Shift Splitter against two widely-used document tiling strategies:

(1) Fixed-Length Splitting: A standard heuristic-based baseline that partitions text into chunks of a fixed number of sentences or tokens. This method ignores the underlying semantic structure but serves as a fundamental benchmark for retrieval efficiency.

(2) Standard Semantic Splitter: A dynamic splitting strategy popularized by frameworks like LlamaIndex (LlamaIndex, 2022) and LangChain (LangChain, 2022). This approach determines boundaries by calculating the cosine dissimilarity between adjacent sentence embeddings and setting breakpoints at a specific percentile of local dissimilarity scores.

(3) Semantic Shift Splitter (Ours): Our proposed method, which leverages semantic shift to achieve more contextually coherent partitions.

Metrics. We report boundary Precision/Recall/F1, P_k, and WindowDiff (WD), and additionally track chunk-size statistics **avg_sents/chunk** and **var_sents/chunk**. The variance term is practically important for RAG, since large variance introduces unstable evidence granularity and can bias retrieval and reranking.

Matching chunk granularity. To avoid confounding segmentation quality with chunk size, we compare methods under **approximately matched avg_sents/chunk**. Fixed controls granularity via k (sentences per chunk), while Semantic/Shift mainly

Granularity	Splitter	P	R	F1	Pk↓	WD↓	avg_sents	var_sents
	Fixed	0.1998	0.3322	0.2495	0.5353	0.5410	2.998	0.004
≈**3**	Semantic	0.1684	0.2688	0.2071	0.4088	0.4533	3.123	5.395
	Shift(Ours)	**0.3809**	**0.6244**	**0.4731**	**0.3733**	**0.3894**	3.041	0.977
	Fixed	0.2010	0.2003	0.2007	0.4854	0.4884	4.998	0.002
≈**5**	Semantic	0.1555	0.1553	0.1554	0.3914	0.4074	4.990	16.371
	Shift(Ours)	**0.3452**	**0.3406**	**0.3429**	**0.3763**	**0.3827**	5.049	1.489
	Fixed	0.2254	0.1603	0.1873	0.4352	0.4382	7.000	0.000
≈**7**	Semantic	0.1384	0.0968	0.1139	**0.3944**	**0.4007**	7.117	41.046
	Shift(Ours)	**0.3014**	**0.2104**	**0.2478**	0.4051	0.4104	7.134	1.600

Table 5: ArXiv paragraph-based segmentation: comparison of Fixed Splitter, Semantic Splitter, and Semantic Shift Splitters under matched granularity (≈3/5/7 sentences per chunk). Higher is better for P/R/F1; lower is better for Pk and WindowDiff. Chunk statistics (avg_sents/chunk, var_sents/chunk) are reported in the last two columns.

Granularity	Splitter	P	R	F1	Pk↓	WD↓	avg_sents	var_sents
	Fixed	0.2845	0.3077	0.2956	0.4772	0.4779	3.000	0.000
≈**3**	Semantic	0.3304	0.3543	0.3420	0.4679	0.5501	3.026	7.287
	Shift(Ours)	**0.4203**	**0.4487**	**0.4340**	**0.3907**	**0.4014**	3.039	1.078
	Fixed	0.2962	0.1923	0.2332	0.5246	0.5246	4.995	0.016
≈**5**	Semantic	0.3070	0.1993	0.2417	0.5160	0.5494	4.995	22.941
	Shift(Ours)	**0.3485**	**0.2238**	**0.2725**	**0.5124**	**0.5135**	5.049	2.304
	Fixed	0.3417	0.1585	0.2166	0.5343	0.5346	6.985	0.090
≈**7**	Semantic	0.3026	0.1375	0.1891	0.5408	0.5573	7.128	43.186
	Shift(Ours)	**0.3753**	**0.1737**	**0.2375**	**0.5318**	**0.5329**	7.003	3.269

Table 6: MINE paragraph-based segmentation: comparison of Fixed Splitter, Semantic Splitter, and Semantic Shift Splitters under matched granularity (≈3/5/7 sentences per chunk). Higher is better for P/R/F1; lower is better for Pk and WindowDiff. Chunk statistics (avg_sents/chunk, var_sents/chunk) are reported in the last two columns.

use semantic_percentile and shift_percentile, respectively (smaller percentile ⇒ more cuts). In practice, we (i) set Fixed to a target k, (ii) sweep a small set of percentile values for Semantic and Shift, and (iii) select configurations whose avg_sents/chunk best matches the target. This protocol produces a fair head-to-head comparison where improvements reflect better boundary placement and segmentation consistency rather than simply generating finer chunks.

G.5 Results and Observations

Tables 5 and 6 summarize results on ArXiv and MINE under three matched granularities (≈3/5/7 sentences per chunk).

Boundary quality: Semantic Shift Splitter yields consistently higher F1 at matched granularity. Across both datasets, the Semantic Shift Splitter achieves the strongest boundary F1 in *all* three regimes. On ArXiv (Table 5), Shift improves F1 substantially over Fixed and the standard Semantic Splitter at ≈3/5/7 (0.4731/0.3429/0.2478 vs. 0.2495–0.2007–0.1873 for Fixed and 0.2071–0.1554–0.1139 for Semantic). The same pattern holds on MINE (Table 6), where Shift achieves the best F1 at ≈3/5/7 (0.4340/0.2725/0.2375), outperforming both baselines in each regime. Notably, these gains are typically driven by improved recall while maintaining strong precision, consistent with the intuition that shift detects when a segment becomes semantically unstable and should be cut.

Window metrics: Semantic Shift Splitter improves Pk/WD on ArXiv and remains competitive on MINE. On ArXiv, Semantic Shift Splitter achieves the best (lowest) P_k and WD across the first two granularities (Table 5), indicating not only better point-wise boundary alignment (higher F1) but also stronger global consistency under window-based evaluation. On MINE, Shift attains the lowest P_k and WD at ≈3, and remains close to Fixed at ≈5 and ≈7 (Table 6). Overall, Shift provides a clearer and more reliable trade-off: it improves boundary F1 consistently, while maintaining competitive window metrics across two distinct do-

mains.

Chunk-size stability: Semantic Shift Splitter sharply reduces variance relative to Semantic splitting. A persistent observation across both datasets is that the standard Semantic Splitter produces highly uneven chunk sizes even when avg_sents/chunk is matched. Its variance grows rapidly with granularity (ArXiv: 5.395/16.371/41.046; MINE: 7.287/22.941/43.186), indicating a mixture of very short and very long chunks (Tables 5 and 6). By contrast, Semantic Shift Splitter maintains much lower variance (ArXiv: 0.977/1.489/1.600; MINE: 1.078/2.304/3.269), approaching the regularity of Fixed splitting while remaining content-adaptive. This stability is practically important for RAG, as it reduces both fragmented evidence (overly short chunks) and token-inefficient passages (overly long chunks), making retrieval and reranking behavior more predictable.

Across ArXiv and MINE datasets, and under matched granularity, the Semantic Shift Splitter consistently improves boundary F1 and generally yields better or competitive P_k/WD, while dramatically reducing chunk-size variance compared to the standard Semantic Splitter (Tables 5 and 6). These results support the claim that explicitly combining local semantic transitions with global dispersion leads to a more accurate and controllable segmentation strategy.

G.6 Summary and Commentary

This appendix turns the semantic shift from an analytic quantity into a practical segmentation mechanism. The Semantic Shift Splitter cuts when a segment's joint local shift and global dispersion indicate semantic instability, using a document-adaptive threshold and an online construction compatible with RAG constraints. Empirically, under matched granularity, it consistently improves boundary F1 over Fixed and Semantic splitting, while dramatically reducing chunk-size variance compared to the Semantic Splitter. These results support the broader takeaway of the main paper: semantic shift is not only a fundamental challenge for embeddings and retrieval but also a useful, controllable signal for building more reliable text processing components.