Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Multi-Level Paragraph Representation through Semantic States, Continuity, and Transitions

**Ricky Y. T. Hou**

rickyhou@bnbu.edu.cn

Beijing Normal-Hong Kong Baptist University

**Wentao Fan**

Beijing Normal-Hong Kong Baptist University

**Patrick Chau**

Beijing Normal-Hong Kong Baptist University

**Gang Liu**

Beijing Normal-Hong Kong Baptist University

**Wenting Deng**

Beijing Normal-Hong Kong Baptist University

**Yandan Lei**

Beijing Normal-Hong Kong Baptist University

**Xueling Luo**

Beijing Normal-Hong Kong Baptist University

**Research Article**

**Keywords:** Paragraph similarity, Sentence embeddings, Multi-level representation, Semantic states, Semantic continuity, Semantic transitions, Optimal transport, Sinkhorn algorithm, Semantic organization

**Posted Date:** November 4th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11398234/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Multi-Level Paragraph Representation through Semantic States, Continuity, and Transitions

**Abstract**

Paragraph similarity is commonly assessed by representing each paragraph as a single vector, which can capture overall semantic content but may obscure sentence-level organization and semantic changes across a paragraph. This paper proposes a multi-level paragraph representation that combines three complementary components: semantic states (L1), local semantic continuity (L2), and semantic transitions in embedding space (L3). The components are compared using optimal transport and combined through weighted fusion to evaluate both semantic similarity and structural sensitivity. Experiments across multiple datasets and evaluation settings show that the components capture different aspects of paragraph similarity. In a semantic similarity evaluation using SBERT embeddings, the fused representation achieves a Spearman correlation of 0.8881, comparable to 0.8876 for the SBERT mean-pooling baseline, although its Pearson correlation is lower. In a controlled sentence-reordering experiment on ROCStories, L2 and L3 achieve ROC-AUC scores of 1.000 for distinguishing original from reordered stories under the tested conditions. These findings suggest that multi-level representation can provide explicit sensitivity to sentence-level organization while maintaining competitive rank-based semantic similarity performance, rather than uniformly outperforming single-vector representations.

**Keywords:** Paragraph similarity; Sentence embeddings; Multi-level representation; Semantic states; Semantic continuity; Semantic transitions; Optimal transport; Sinkhorn algorithm; Semantic organization.

## 1. Introduction

### 1.1. Paragraph Similarity and Representation

Paragraph similarity commonly begins by converting variable-length text into a fixed-dimensional numerical representation. Traditional approaches such as Doc2Vec (Le and Mikolov 2014) directly learn a dense vector for a sentence, paragraph, or document, providing a compact representation that can be used for downstream similarity and classification tasks.

More recent approaches commonly construct paragraph representations from sentence embeddings. For example, Sentence-BERT (SBERT) (Reimers and Gurevych 2019) produces semantically meaningful embeddings for individual sentences that can be compared using cosine similarity. A paragraph can consequently be represented by encoding its sentences, $P = (s_1, s_2, \cdots, s_n) \rightarrow E = (e_1, e_2, \cdots, e_n)$, and then aggregating the sentence embeddings into a single paragraph vector. A simple and widely applicable strategy is mean pooling, $v_P = \frac{1}{n}\sum_{i=1}^{n} e_i$.

Simple averaging-based representations have also been shown to provide surprisingly strong baselines for sentence-level semantic representation (Arora et al 2017). Other sentence-embedding approaches, such as Sent2Vec (Pagliardini et al 2018), similarly provide distributed representations that can serve as building blocks for representing longer text. These approaches reflect a dominant paragraph-to-single-vector paradigm: $P \rightarrow v_P$.

Once a paragraph has been converted to $v_P$, its similarity to another paragraph can be computed using a vector similarity or distance measure. This paradigm is attractive because it produces a compact, fixed-dimensional representation that is convenient for efficient comparison and downstream applications.

However, the compression of a paragraph into a single vector primarily emphasizes its aggregate semantic content. The paragraph itself is an ordered sequence of sentences, and information concerning how those sentences relate to one another, how semantic content develops across the sequence, and how the paragraph moves between semantic states is not explicitly represented in $v_P$.

Two paragraphs may discuss similar concepts and therefore have similar aggregate semantic representations, while differing substantially in how those concepts are organized. In particular, differences

in sentence ordering, local continuity, semantic progression, and transitions can produce different paragraph structures even when the overall semantic content remains similar.

Representing a paragraph as a single vector is effective for capturing overall semantic content, but the compression may not explicitly preserve the relationships between consecutive sentences or the trajectory of semantic change across the paragraph. This distinction is related to research on textual coherence (Barzilay and Lapata 2008), which emphasizes that the organization and relationships among sentences contribute to how a text is structured and interpreted. Thus, aggregate semantic similarity and sequential structural similarity are related but distinct properties (Lapata and Barzilay 2005).

### 1.2. A Multi-Level View of Paragraph Semantics

A paragraph contains not only a collection of semantic meanings but also an organization of those meanings across sentences. A single aggregate representation may not explicitly distinguish how individual semantic states are connected or how meaning changes as the paragraph progresses. A multi-level view therefore provides a more informative representation by separating these complementary aspects of paragraph semantics. Specifically, it captures the semantic content of the paragraph, the local relationships between adjacent semantic states, and the semantic movement between consecutive states. A more detailed design of these three levels is presented in Section 3.

### 1.3. Motivation for Optimal Transport

The multi-level representation produces sequences of semantic states, continuity values, and transition vectors rather than a single fixed-length paragraph vector. Comparing such representations therefore requires a similarity measure that can accommodate differences in their internal structure. In particular, two paragraphs may contain different numbers of sentences or exhibit different local alignments between their semantic components.

Optimal Transport (OT) (Kusner et al 2015) (Peyré and Cuturi 2019) provides a natural framework for this purpose. Rather than requiring a strict one-to-one correspondence between elements, OT compares two collections through a transport plan that minimizes the overall cost of matching their components. This idea has been successfully applied to text comparison; for example, Word Mover's Distance (WMD) (Kusner et al 2015) compares documents by transporting semantic mass between embedded words according to their distances in the embedding space.

This flexibility is useful for the proposed paragraph representation because corresponding semantic elements need not occur at identical positions or have an exact one-to-one alignment. The comparison can instead be based on the semantic cost of matching components across the two representations. Sinkhorn's entropic regularization further provides an efficient approximation to the optimal transport problem, making OT more practical for repeated comparisons (Cuturi 2013).

Accordingly, OT is used in this work as a common comparison mechanism for the three representation levels. It allows three levels to be compared through their respective cost structures while retaining the flexibility of distribution-level matching.

### 1.4. Research Gap

Existing paragraph representations, including Doc2Vec, paragraph-level SBERT, and sentence-embedding aggregation, primarily characterize the aggregate semantic content of a paragraph (Le and Mikolov 2014). In contrast, research on discourse coherence shows that relationships between sentences also contribute to the organization of meaning.

The gap addressed in this work is the lack of an explicit, unified representation that considers both sentence-level semantic states and their sequential organization for paragraph similarity. We therefore propose a three-level representation with Sinkhorn OT used to compare the resulting representations (Peyré and Cuturi 2019).

Our central hypothesis is that sequential semantic information provides complementary information to aggregate semantic content. In particular, we hypothesize that explicitly modeling continuity and transitions can reveal structural differences between paragraphs that are not directly represented by a single aggregate paragraph vector.

Based on the motivation and research gap described above, this study investigates the following four

research questions:

(a) Does the proposed multi-level representation provide an effective representation of paragraph similarity compared with conventional paragraph-level representations?
(b) Do semantic continuity and semantic transition provide information that is complementary to sentence-level semantic states when measuring paragraph similarity?
(c) Can the proposed representation detect changes in paragraph structure, including sentence reordering and variations in the density and extent of structural perturbations, while the overall semantic content remains similar?
(d) How robust is the proposed framework to different sentence encoders, paragraph lengths, fusion weights, and sampling conditions?

Together, these questions examine not only whether the proposed representation can measure paragraph similarity, but also whether its multiple semantic levels capture complementary sequential information and remain meaningful under different structural and experimental conditions.

### 1.5. Contributions and Paper Organization

This work makes four main contributions:

(a) We propose a representation that models paragraph semantics at three levels: semantic state, semantic continuity, and semantic transition, extending the conventional single-vector representation with explicit sequential information.
(b) We develop a unified OT-based framework for comparing the three levels, allowing their respective semantic structures to be evaluated within a common similarity framework.
(c) We systematically examine the sensitivity of the representation to sentence reordering, perturbation density, and paragraph length, providing direct tests of its ability to capture semantic organization through local continuity and semantic transitions.
(d) We evaluate component complementarity, encoder robustness, fusion-weight sensitivity, statistical reproducibility, and performance against established representations including SBERT, Doc2Vec, and Sent2Vec.

Together, these contributions provide an empirical assessment of whether explicitly modeling sequential semantic structure adds useful information beyond aggregate paragraph representations.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 presents the proposed method, and Section 4 describes the experimental methodology. Section 5 reports the experimental results, while Section 6 provides analysis and discussion. Finally, Section 7 concludes the paper and outlines directions for future work.

## 2. Related Work

This section reviews the evolution of text representation from paragraph-level embeddings to sentence-level and sequential representations, with particular attention to their ability to capture semantic similarity and structural information. It concludes by summarizing the relevant literature and identifying the context in which the proposed multi-level representation framework is developed.

### 2.1. Evolution of Text Representation

Text representation has evolved from word-level vector models toward increasingly expressive representations of sentences and longer text. Early neural embedding methods such as Word2Vec represented individual words as dense vectors learned from their contextual usage, providing a continuous representation of semantic and syntactic relationships (Apidianaki 2023). GloVe subsequently incorporated global word–context statistics into word representations, while fastText extended this approach by representing words through character n-grams, improving the treatment of morphology and rare or unseen words (Pennington et al 2014) (Bojanowski 2017).

The focus then shifted from individual words to longer textual units. Sent2Vec extended the distributional embedding paradigm to sentences by learning sentence representations from compositional word and n-gram information (Pagliardini et al 2018). Doc2Vec, or Paragraph Vector, similarly introduced fixed-length distributed representations for variable-length text such as sentences, paragraphs, and documents (Le and Mikolov 2014). These approaches enabled an entire textual unit to be represented by a single dense vector, providing a compact representation suitable for downstream comparison and classification.

More recently, transformer-based models have produced contextual sentence representations. Sentence-BERT (SBERT), for example, adapts pretrained BERT using siamese and triplet network structures to generate semantically meaningful sentence embeddings that can be efficiently compared using cosine similarity (Reimers and Gurevych 2019). Unlike paragraph-level approaches that directly map an entire paragraph to one vector, sentence encoders can represent a paragraph as a sequence of sentence embeddings. This provides a useful basis for examining not only the semantic content of individual sentences but also their continuity and transitions, which motivates the representation framework developed in this study.

### 2.2. Paragraph-Level Representation

Paragraph-level representation methods aim to map a variable-length paragraph or document to a single fixed-dimensional vector. Doc2Vec is an early example, learning a dense representation for each document or text segment from its constituent words. Its PV-DM and PV-DBOW architectures provide fixed-length document representations that can be directly used for comparison or downstream learning tasks (Le and Mikolov 2014).

More recent approaches can also produce a single vector by applying a sentence encoder to the entire paragraph. Sentence-embedding models such as SBERT have also been applied to paragraph similarity tasks, providing a compact representation of paragraph-level semantics (Maslak and Mitkov 2021) (Jin 2025). Sentence-transformer models were specifically developed to produce semantically meaningful fixed-dimensional embeddings that can be efficiently compared using similarity measures such as cosine similarity (Reimers and Gurevych 2019).

Other document-embedding approaches similarly seek to compress longer text into a fixed-size representation, including neural sentence/document encoders (Cer et al 2018) and pooling-based representations. These methods differ in how the vector is learned or constructed, but share a common representational granularity: the paragraph is ultimately represented by one vector. In this sense, paragraph-level representation provides a compact representation of overall semantic content, while the internal sentence-level structure is not represented as a separate sequence of vectors.

### 2.3. Sentence-Level and Sequential Representation

An alternative to single-vector paragraph representations is to retain the paragraph as a sequence of sentence-level representations. Sentence embeddings provide a vector for each sentence, allowing a paragraph to be represented as $P \rightarrow E = (e_1, e_2, \cdots, e_n)$ rather than as a single vector (Huang 2023). Hierarchical models further exploit this structure by first representing individual sentences and then modeling their relationships to obtain a document-level representation (Wu et al 2021). For example, Hierarchical Attention Networks explicitly construct representations at both word and sentence levels and use sentence-level attention to form a document representation (Yang et al 2016).

Sequential and discourse-oriented approaches go beyond retaining sentence representations by explicitly modeling sentence relationships and ordering (Liu 2020). Sentence-ordering and coherence research treats the organization of sentences as an important aspect of coherent text, with models capturing local and global dependencies between sentences (Oh et al 2019) (Prabhumoye et al 2020). These approaches demonstrate that sentence identity alone does not fully describe a paragraph: relationships between neighboring or non-neighboring sentences can also provide information about its organization and coherence (Iter et al 2020).

### 2.4. Semantic Similarity and Structural Similarity

Semantic similarity generally measures the extent to which two texts convey similar meanings or semantic content. This objective underlies semantic textual similarity (STS) evaluation, where sentences or text segments are assigned similarity scores according to their semantic equivalence or relatedness (Agirre et al 2012). Modern sentence-embedding models such as SBERT provide dense representations that can be compared efficiently using cosine similarity, making them widely applicable to semantic similarity tasks (Reimers and Gurevych 2019).

Structural or sequential similarity addresses a related but different property: whether the internal organization and relationships among textual units are similar. Research on discourse coherence, for example, models local coherence through sentence-to-sentence transitions and patterns of entity continuity (Barzilay and Lapata 2008) (Farag et al 2020). Sentence-ordering research likewise demonstrates that the arrangement

of otherwise similar sentences can provide information about textual coherence and organization (Oh et al 2019) (Xu et al 2019).

This distinction is relevant to paragraph-level similarity because two paragraphs may contain similar semantic content while differing in sentence order or internal progression. In the experiments of this study, the proposed framework treats paragraph similarity as potentially involving both semantic content and semantic organization, rather than assuming that aggregate semantic similarity fully characterizes paragraph similarity.

### 2.5. Summary of Related Work

The literature reviewed above shows a progression from word-level representations to sentence- and paragraph-level representations, with increasing attention to contextual meaning and textual structure. Paragraph-level methods generally compress a paragraph into a single vector, whereas sentence-level and sequential approaches retain some information about the constituent sentences and their relationships. These approaches therefore differ not only in modeling architecture but also in the granularity of the representation and the aspects of similarity that can be examined, as shown in Table 1.

Comparison between our proposed framework and existing approaches.

Table 1 Comparison between our proposed framework and existing approaches.

| Representation approach | Basic unit | Sentence order retained? | Internal relationships represented? |
|---|---|---|---|
| Doc2Vec | Paragraph/document | No | No |
| Whole-paragraph SBERT | Paragraph | No | No |
| Mean-pooled sentence embeddings | Sentences | No | No |
| Sentence embeddings | Individual sentences | Yes | Not explicitly |
| **Proposed L1/L2/L3 framework** | Sentence sequence + derived structures | **Yes** | **Yes** |

Overall, existing approaches provide complementary ways of representing textual meaning and organization. This review motivates treating a paragraph not only as an aggregate semantic unit but also as an ordered sequence of semantic states and relationships, which provides the conceptual basis for the representation framework described in the next section.

## 3. Proposed Method

This section presents the proposed framework for representing and comparing paragraphs through sentence-level semantic information and multi-level structural representations. It first describes paragraph representation and sentence encoding, followed by the construction of three levels of semantic representation and their comparison using Optimal Transport (Peyré and Cuturi 2019) with multi-level fusion. Finally, the computational procedure and implementation details are provided to clarify how the proposed framework is applied in practice.

### 3.1. Paragraph Representation and Sentence Encoding

We represent a paragraph as an ordered sequence of sentences, $P = (s_1, s_2, \cdots, s_n)$, where $s_i$ denotes the $i$-th sentence and $n$ is the number of sentences in the paragraph. Rather than directly mapping the entire paragraph to a single vector, each sentence is independently encoded into a $d$-dimensional semantic embedding: $e_i = f(s_i)$, where $f(\cdot)$ denotes a sentence encoder, such as Sent2Vec or SBERT. The paragraph is therefore represented as an ordered sequence of sentence embeddings, $P \to E = (e_1, e_2, \cdots, e_n)$.

This sequence serves as the fundamental input to the proposed multi-level representation. From $E$, three complementary levels are constructed: L1 Semantic State, which represents the semantic content of individual sentences; L2 Semantic Continuity, which characterizes the local semantic relationship between adjacent sentences; and L3 Semantic Transition, which represents the change in semantic state from one

sentence to the next. The subsequent stages compare these three representations using Sinkhorn OT (Cuturi 2013) and combine their resulting similarity scores. Figure 1 shows the five main steps of the proposed framework.

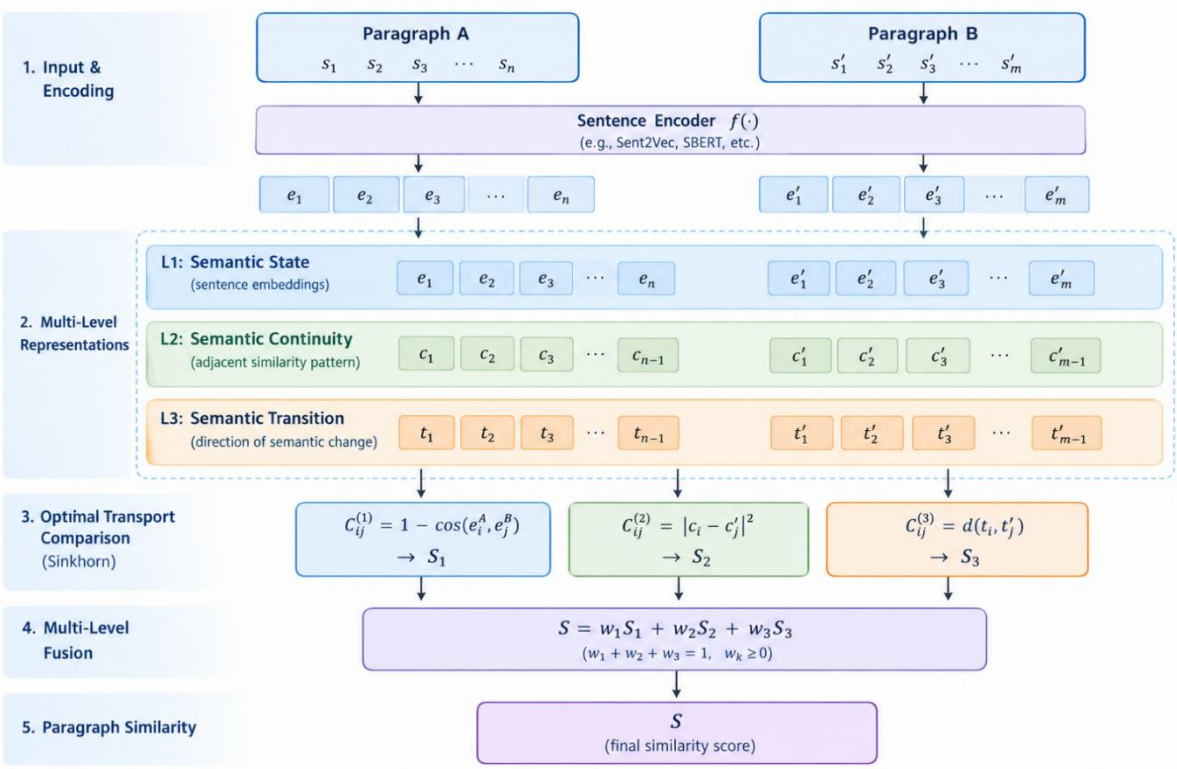



**Fig. 1.** Overview of the proposed multi-level paragraph representation and comparison framework. Five steps of paragraph comparison: input and sentence encoding, three-level representation, optimal transport comparison, weighted fusion, and final output of similarity.

### 3.2. Multi-Level Semantic Representation

The sentence-embedding sequence $E$ is further represented at three complementary levels: semantic state, semantic continuity, and semantic transition. Each level captures a different aspect of paragraph structure.

(1) Semantic State (L1)

The first level directly uses the sequence of sentence embeddings,

$$E = (e_1, e_2, \cdots, e_n), \tag{1}$$

to represent the semantic states of the paragraph. Each $e_i$ describes the semantic content of an individual sentence. L1 therefore captures what semantic information is present in the paragraph, without explicitly modeling the relationships between successive sentences.

(2) Semantic Continuity (L2)

To characterize the local relationship between adjacent sentences, we calculate their cosine similarity:

$$c_i = cos(e_i, e_{i+1}), \ i = 1, \cdots, n-1. \tag{2}$$

The paragraph is consequently represented by a continuity sequence,

$$C = (c_1, c_2, \cdots, c_{n-1}). \tag{3}$$

L2 captures the local continuity pattern of a paragraph—how closely the semantic content of neighboring sentences is connected. Unlike L1, it therefore provides information about relationships between successive semantic states rather than the states themselves.

(3) Semantic Transition (L3)

To represent how semantic content changes from one sentence to the next, we define the transition vector as

$$t_i = e_{i+1} - e_i,\ i = 1, \cdots, n-1. \tag{4}$$

The resulting transition sequence is

$$T = (t_1, t_2, \cdots, t_{n-1}). \tag{5}$$

L3 captures the direction of semantic change between consecutive sentences, providing information about how the paragraph moves from one semantic state to another.

The proposed framework encodes sequential information through adjacent-sentence continuity and embedding-space transition features before comparing the resulting representations using optimal transport. The three levels provide complementary views of a paragraph. Together, these representations model what a paragraph contains, how its semantic states are connected, and how they change.

### 3.3. Optimal Transport Comparison and Multi-Level Fusion

The three representations provide different views of paragraph semantics and are compared independently using Sinkhorn optimal transport (OT) (Cuturi 2013). For level $k \in \{1, 2, 3\}$, let $X_k^A$ and $X_k^B$ denote the corresponding representations of paragraphs $A$ and $B$. The distance at each level is defined as

$$D_k = OT(X_k^A, X_k^B),\ k \in \{1, 2, 3\}. \tag{6}$$

For L1, the cost between two sentence embeddings is based on cosine distance:

$$C_{ij}^{(1)} = 1 - cos(e_i^A, e_j^B). \tag{7}$$

For L2, the cost measures the difference between continuity values:

$$C_{ij}^{(2)} = \left| c_i^A - c_j^B \right|^2. \tag{8}$$

For L3, the cost measures the distance between semantic transition vectors:

$$C_{ij}^{(3)} = 1 - \frac{t_i^A \cdot t_j^B}{\|t_i^A\| \cdot \|t_j^B\|}, \tag{9}$$

where $t_i^A$ and $t_j^B$ denote transition vectors from paragraphs $A$ and $B$, respectively. Sinkhorn OT (Cuturi 2013) then provides a soft alignment between the two representations without requiring a fixed one-to-one correspondence between their elements (see Fig. 2). This is particularly useful when paragraphs differ in sentence count or when corresponding semantic states occur at different positions.

The level-specific Sinkhorn distances $D_1$, $D_2$, and $D_3$ are transformed into bounded similarity scores using $S_k = 1/(1 + D_k)$. This monotonic transformation maps all scores to the interval $(0,1]$, enabling weighted fusion across the three semantic levels. The three level-specific similarity scores are subsequently combined into a single paragraph similarity score:

$$S = w_1 S_1 + w_2 S_2 + w_3 S_3, \tag{10}$$

subject to

$$w_1 + w_2 + w_3 = 1,\ w_k \geq 0. \tag{11}$$

The weights determine the relative contribution of semantic state, continuity, and transition information. In the initial heuristic configuration, the weights were set to 0.40/0.10/0.50, although subsequent experiments investigated alternative weight combinations rather than assuming that this configuration is universally optimal.

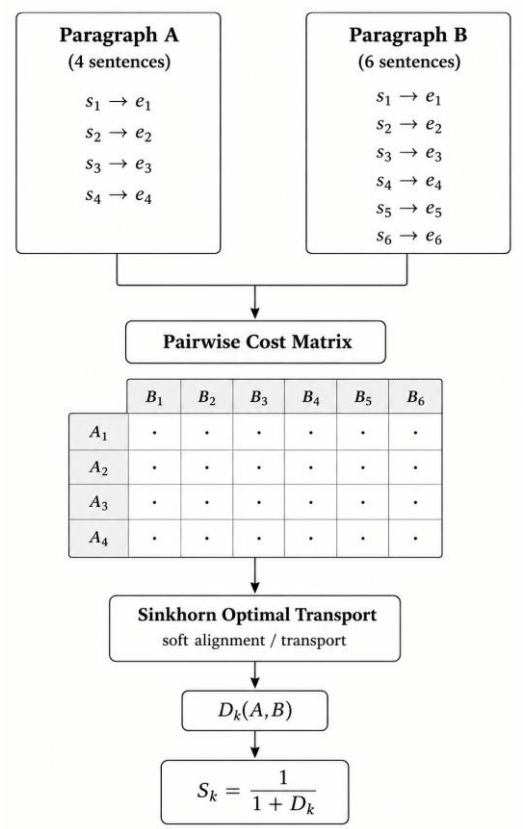



**Fig. 2.** Illustration of the OT comparison between variable-length paragraph representations.

The three similarity scores are intended to provide complementary information. $S_1$ primarily measures similarity in semantic content, $S_2$ measures similarity in local semantic continuity, and $S_3$ measures similarity in semantic transitions. Specifically, $S_2$ and $S_3$ assess differences in the patterns of local semantic relationships and semantic transitions between two paragraphs. The proposed fusion therefore extends paragraph comparison beyond differences in individual semantic states to also consider how those states are connected and how they evolve throughout the paragraph.

### 3.4. Computational Procedure and Implementation

Given two paragraphs $A$ and $B$, the proposed method follows a sequential processing pipeline. First, each paragraph is segmented into sentences and each sentence is encoded using the selected sentence encoder. The resulting sentence embeddings are then used to construct the three representations. Sinkhorn OT (Cuturi 2013) is subsequently applied independently to the three levels to obtain $S_1$, $S_2$, and $S_3$. Finally, these similarities are combined using the specified fusion weights to produce the overall similarity score $S$.

The computational procedure (Algorithm 1) can be summarized as follows:

---

**Algorithm 1: Multi-Level Paragraph Similarity**

---

1. Segment paragraphs $A$ and $B$ into sentences.
2. Encode each sentence to obtain the embedding sequences.
3. Construct the three representations.
4. Construct the corresponding level-specific cost matrices.
5. Apply Sinkhorn OT to obtain $S_1$, $S_2$, and $S_3$ (Section 3.3).
6. Fuse the three similarity scores.
7. Return $S$ as the final similarity.

---

The framework separates sentence encoding from multi-level comparison. Thus, the same L1/L2/L3 and

Sinkhorn procedure can be applied with different sentence encoders. In the experiments, both Sent2Vec and SBERT were used, with SBERT based on all-mpnet-base-v2. This separation allows the effect of the sentence encoder to be evaluated independently from the proposed multi-level architecture.

From a computational perspective, sentence encoding constitutes a major part of the processing cost, particularly for long documents, whereas the construction of L1, L2, and L3 is relatively lightweight (see Experiment 5 in Section 5.2). This separation is useful for implementation because sentence embeddings can be computed once and cached, after which the different representations and comparisons can be evaluated without repeatedly encoding the same sentences.

## 4. Experimental Methodology

### 4.1. Datasets

The datasets used in this study are grouped into two categories: (1) standard similarity datasets and (2) structural evaluation datasets.

(1) Standard Similarity Datasets

The pszemraj/synthetic-text-similarity dataset (Pszemraj 2026) (paragraphs containing at least two sentences) provides document pairs with automatically generated similarity scores. These scores are derived from cosine similarities between document embeddings, and pairs are sampled based on the resulting similarity-score distribution. The scores are rescaled to the range [0, 1]. We use this dataset to examine how closely the proposed multi-level representation agrees with these embedding-based similarity scores. It consists of document pairs with associated similarity scores that reflect semantic relatedness. This dataset was used in Experiments 1, 2, 7 and 8.

(2) Structural Evaluation Datasets

These datasets were used to evaluate sentence reordering, sentence replacement, coherence disruption, and semantic-transition changes.

Experiment 3 used DiscoEval DCwiki (Glick 2023), which contains coherent six-sentence Wikipedia sequences for discourse-coherence evaluation. The test split was used, and 1,000 coherent paragraphs were selected as the original texts. Rather than using the dataset's existing incoherent samples, controlled perturbations were generated from coherent paragraphs to provide better experimental control.

Experiment 4 used ROCStories (Jupally 2024), a corpus of five-sentence narratives, to examine sensitivity to sentence ordering. This experiment directly evaluated whether the proposed representation preserves sequential and discourse-level information that may be lost by mean-based paragraph representations.

Experiments 5 and 6 used EleutherAI/ wikitext_document_level (EleutherAI 2023) (configuration: wikitext-103-raw-v1, train split). A total of 500 WikiText articles were selected to extend the analysis to substantially longer natural documents. Unlike ROCStories, WikiText provides naturally occurring long-form texts, enabling evaluation across a wide range of document lengths.

### 4.2. Sentence Encoders and Traditional Baselines

Two sentence encoders were evaluated: Sent2Vec and SBERT. Sent2Vec was used as the initial experimental encoder, while SBERT all-mpnet-base-v2 was subsequently introduced as a stronger transformer-based encoder. The L1/L2/L3 representation and Sinkhorn OT procedure remained unchanged across encoders.

The Doc2Vec baselines are PV-DBOW and PV-DM. Both represent each paragraph as an embedding, and paragraph embeddings are compared using cosine similarity. In addition, we have the following two more baselines: Whole-paragraph SBERT followed by cosine similarity and sentence-average SBERT, $v_P = \frac{1}{n}\sum_i e_i$ followed by cosine similarity. The last two can show the actual impacts of SBERT in each area.

### 4.3. Experimental Design

The experiments are designed as a coherent progression, moving from representation validity and component complementarity to structural sensitivity, robustness and reproducibility, and finally benchmarking against established methods.

(1) Representation validity and component complementarity

Experiment 1 evaluated the validity of the proposed three-level representation. Experiment 2 examined

encoder robustness, the effect of a stronger modern sentence encoder, and the contributions of the sentence encoder and multi-level architecture. Experiment 3 examined whether the proposed representation captures paragraph structure, what information is captured at each semantic level, and how sensitivity changes with different levels of structural perturbation.

(2) Structural sensitivity

Experiments 4 to 6 shift from paragraph-level semantic correlation to controlled structural evaluation. Experiment 4 examined sensitivity to sentence reordering, Experiment 5 examined whether this sensitivity persists as document length increases, and Experiment 6 investigated whether the observed length effect is primarily due to perturbation dilution rather than an intrinsic limitation on longer documents.

(3) Weighting sensitivity, reproducibility, and benchmarking

Experiment 7 examined the robustness of the multi-level fusion across different weighting configurations and quantified uncertainty through bootstrap analysis (for estimating how reliable the results are). Experiment 8 benchmarked the proposed representation against established paragraph/document representations and statistically quantified the differences in performance.

## 5. Experiment Results

This section presents the experimental results in three stages. It first evaluates representation validity and component complementarity, then examines structural sensitivity under controlled perturbations, and finally assesses weighting sensitivity, reproducibility, and performance relative to established methods. Together, these results evaluate both the semantic and structural capabilities of the proposed multi-level representation.

### 5.1. A. Representation Validity and Component Complementarity

(1) Experiment 1

Experiment 1 used 1,000 valid paragraph pairs from the pszemraj/synthetic-text-similarity dataset (Pszemraj 2026), with both paragraphs required to contain at least two sentences. Sentence representations were generated using Sent2Vec (wiki_unigrams.bin). Similarity comparison was modeled using adjacent sentence cosine similarities and compared via Sinkhorn OT. Pearson and Spearman correlations served as the primary evaluation metrics. Experiments examined the effectiveness of individual semantic levels, their combinations, and two fusion schemes, namely equal weighting and the default weighting of 0.40/0.10/0.50.

Table 2 reports the paragraph-similarity performance of each level and their combinations.

Table 2 Performance of each individual component and their combinations.

| Representation | Pearson | Spearman |
|---|---|---|
| L1 | 0.643 | 0.680 |
| L2 | 0.425 | 0.457 |
| L3 | 0.679 | 0.749 |
| L1 + L2 | 0.678 | 0.690 |
| L1 + L3 | 0.690 | 0.751 |
| L2 + L3 | 0.693 | 0.706 |
| Equal 3-level | 0.714 | 0.745 |
| 0.40/0.10/0.50 | 0.702 | 0.758 |

As shown in the table, the individual components L1, L2, and L3 achieve Spearman correlations of 0.680, 0.457, and 0.749, respectively. Among the three levels, L3 provides the strongest individual performance, while L2 exhibits the weakest standalone correlation. Nevertheless, when the three levels are combined using the default fusion weights (0.40/0.10/0.50), the overall performance improves to 0.702 Pearson and 0.758 Spearman correlation.

A comparison between the two-level and three-level models further demonstrates the contribution of Semantic Continuity. For example, the L1+L3 model achieves a Pearson correlation of 0.690, whereas the default three-level fusion achieves 0.702, representing an improvement of 0.012. Similarly, the Spearman

correlation increases from 0.751 to 0.758, yielding a gain of 0.007. Although L2 has relatively weak standalone performance, these improvements indicate that Semantic Continuity contributes complementary information that is not fully captured by either Semantic State or Semantic Transition.

To further investigate the relationships among the three semantic levels, we computed their pairwise correlations as shown in Table 3:

Table 3 Pairwise correlations of L1, L2, and L3.

| | L1 | L2 | L3 |
|---|---|---|---|
| L1 | 1.000 | 0.332 | 0.831 |
| L2 | 0.332 | 1.000 | 0.370 |
| L3 | 0.831 | 0.370 | 1.000 |

The correlation between L1 and L2 is 0.332, while the correlation between L2 and L3 is 0.370. In contrast, L1 and L3 exhibit a substantially higher correlation of 0.831. These results suggest that Semantic Continuity captures information that is largely distinct from both Semantic State and Semantic Transition. Therefore, despite its weaker individual predictive power, L2 provides complementary semantic information that contributes to improved overall performance when integrated with the other levels. The findings support the inclusion of Semantic Continuity as a separate semantic dimension within the proposed three-level framework when using the Sent2Vec encoder.

(2) Experiment 2

Experiment 2 had four objectives: (1) to evaluate the robustness of the proposed representation to different sentence encoders; (2) to determine whether a stronger encoder improves performance; (3) to assess the contribution of the sentence encoder; and (4) to separate encoder effects from those of the proposed multi-level architecture.

Paragraphs were first segmented into sentences and encoded using SBERT all-mpnet-base-v2. The resulting sentence embeddings were transformed into the L1, L2, and L3 representations, which were compared using Sinkhorn OT and subsequently fused into a final similarity score.

The main experimental settings were as follows. Sentences were encoded using SBERT all-mpnet-base-v2, producing 768-dimensional embeddings. The experiment used 1,000 valid paragraph pairs from the pszemraj/synthetic-text-similarity dataset. Similarity was computed using Sinkhorn OT with a regularization parameter of 0.05, and the three semantic levels were fused using weights of 0.40/0.10/0.50. Doc2Vec PV-DBOW and Doc2Vec PV-DM (50 inference epochs) served as baselines. Performance was evaluated using Pearson and Spearman correlations, as shown in Table 4.

Table 4 Experiment 2 results.

| **Method** | **Pearson** | **Spearman** |
|---|---|---|
| Doc2Vec PV-DBOW | 0.7477 | 0.7478 |
| Doc2Vec PV-DM | 0.7356 | 0.7344 |
| SBERT L1 | 0.7603 | 0.8685 |
| SBERT L2 | 0.4074 | 0.4302 |
| SBERT L3 | 0.6335 | 0.8642 |
| Proposed L1+L2+L3 | 0.7461 | 0.8755 |

The most notable effect of Experiment 2 was the improvement in L1. Compared with Experiment 1 (Pearson = 0.643, Spearman = 0.680), SBERT-based L1 achieved Pearson = 0.7603 and Spearman = 0.8685, corresponding to improvements of approximately 0.117 and 0.189, respectively. This result highlights the importance of the underlying sentence encoder and indicates that the proposed framework benefits substantially from stronger sentence representations.

The fused L1+L2+L3 representation achieved Pearson = 0.7461 and Spearman = 0.8755, obtaining the highest Spearman correlation among all methods evaluated. Overall, Experiment 2 demonstrates that upgrading the sentence encoder substantially improves the semantic-state representation, while the proposed multi-level framework contributes complementary structural information that yields the strongest rank-based

performance.

(3) Experiment 3

Experiment 3 investigated whether the proposed representation is sensitive to paragraph structure, what information is captured by each semantic level, and how the representation responds to different types of perturbations.

The experiment used 1,000 coherent six-sentence Wikipedia paragraphs from DiscoEval DCwiki (Glick 2023). Four perturbation types were generated from the original paragraph $P = [s1, s2, s3, s4, s5, s6]$: (a) an adjacent sentence swap ($\rightarrow [s1, s3, s2, s4, s5, s6]$), (b) a non-adjacent sentence swap ($\rightarrow [s1, s4, s3, s2, s5, s6]$), (c) a single intruder replacement, in which one sentence was replaced with a sentence from another coherent paragraph, and (d) two intruder replacements, in which two sentences were replaced, producing a stronger structural perturbation. For each original-perturbed pair, similarities were computed for L1 ($S_1$), L2 ($S_2$), L3 ($S_3$), and the fused three-level representation ($S_{123}$).

Sentences were encoded using SBERT all-mpnet-base-v2. Similarities were computed using Sinkhorn OT with fusion weights of 0.40/0.10/0.50. Structural sensitivity was evaluated using $\Delta S_1$, $\Delta S_2$, $\Delta S_3$, and $\Delta S_{123}$, as shown in Table 5.

Table 5 Experiment 3 results.

| **Condition** | $\Delta S_1$ | $\Delta S_2$ | $\Delta S_3$ | $\Delta S_{123}$ |
|---|---|---|---|---|
| Original | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Adjacent swap | 0.0000 | 0.0154 | 0.2244 | 0.1138 |
| Non-adjacent swap | 0.0000 | 0.0168 | 0.2671 | 0.1353 |
| Single intruder | 0.1371 | 0.0519 | 0.1959 | 0.1580 |
| Two intruders | 0.2381 | 0.1035 | 0.3281 | 0.2696 |

The results reveal distinct roles for the three semantic levels. Under sentence reordering, L1 remained unchanged, indicating that it primarily captures semantic content rather than sentence order. In contrast, L3 exhibited the largest response, showing strong sensitivity to semantic transitions and sentence-to-sentence progression, while L2 detected structural changes more weakly.

The intruder conditions affected both semantic content and structure. Consequently, both L1 and L3 increased substantially as the severity of perturbation increased. The fused representation exhibited a clear severity ordering, with larger similarity reductions observed for increasingly disruptive perturbations. Overall, Experiment 3 demonstrates that L2 and especially L3 capture structural information beyond sentence-level semantic content.

### 5.2. Structural Sensitivity

In Experiments 4, 5, and 6, L1, L2, and L3 are compared using position-aligned distance measures, rather than implying that all three levels are compared using OT.

(1) Experiment 4

Experiment 4 had three objectives: to test whether the proposed representation detects sentence-order changes, to identify the information lost through mean aggregation, and to evaluate the specific roles of L2 and L3. The experiment used ROCStories, a corpus of five-sentence narratives containing natural temporal progression, causal relationships, and narrative structure, making it well suited for sentence-order evaluation.

Each original story $(s1, s2, s3, s4, s5)$ was compared with four reordered versions: an adjacent swap, a non-adjacent swap, a reversal $(s5, s4, s3, s2, s1)$, and a middle swap $(s1, s2, s4, s3, s5)$. Sentences were encoded using SBERT all-mpnet-base-v2, and similarity was measured using the fused L1/L2/L3 representation with Sinkhorn OT. SBERT-Mean, Sent2Vec-Mean, Doc2Vec PV-DBOW, and Doc2Vec PV-DM served as baselines. Performance was evaluated using ROC-AUC, which is used to evaluate how well a model distinguishes between two classes. A value of 0.50 indicates random guessing, whereas 1.00 indicates perfect discrimination. Table 6 reports the results.

We evaluated seven methods on 1,000 five-sentence ROCStories under four sentence-order perturbations. Original stories were labeled negative (0), and perturbed stories positive (1). The distance between each

original–perturbed pair served as the score, with larger distances indicating greater structural change. ROC-AUC was calculated separately for each method and perturbation type. Uncertainty was estimated using 5,000 bootstrap resamples of story IDs, with a random seed of 42. We report 95% percentile confidence intervals based on the 2.5th and 97.5th percentiles.

Table 6 ROC-AUC under sentence reordering.

| Representation | AUC under sentence reordering |
|---|---|
| L1 | ≈ 0.50 |
| SBERT-Mean | ≈ 0.50 |
| Sent2Vec-Mean | ≈ 0.50 |
| Doc2Vec PV-DBOW | ≈ 0.50 |
| Doc2Vec PV-DM | 0.775–0.854 |
| L2 | 1.000 |
| L3 | 1.000 |

The results show that L1 and conventional mean-based representations are essentially order-invariant, achieving ROC-AUC values close to random guessing (≈ 0.50). In contrast, L2 and L3 perfectly distinguished the original stories from their reordered versions, demonstrating strong sensitivity to sentence order. Although Doc2Vec PV-DM exhibited some order awareness, its performance remained substantially below that of L2 and L3. These findings indicate that L2 and L3 capture sequential information that is largely absent from conventional paragraph representations.

(2) Experiment 5

Experiment 5 investigated whether the structural sensitivity of the proposed representation remains effective as document length increases. In particular, the experiment examined the effects of local and global sentence reordering, the dependence of L2 and L3 on document length, and the computational scalability of the framework.

The experiment used 500 articles from WikiText-103 (wikitext-103-raw-v1, train split). Documents of length 5, 10, 20, 40, 80, and 160 sentences were constructed and subjected to the same four sentence-reordering perturbations used in Experiment 4. Sentences were encoded using SBERT all-mpnet-base-v2, with SBERT-Mean serving as a baseline.

The L1 results were essentially zero across all document lengths and perturbation types, indicating that L1 captures aggregate semantic content but is largely insensitive to sentence order (see Table 7).

Table 7 L1 response under sentence reordering.

| Length | Adjacent | Non-adjacent | Reverse | Middle |
|---|---|---|---|---|
| 5 | $2.94\times10^{-8}$ | $3.84\times10^{-8}$ | $4.57\times10^{-8}$ | $2.94\times10^{-8}$ |
| 10 | $2.01\times10^{-8}$ | $2.66\times10^{-8}$ | $5.32\times10^{-8}$ | $2.60\times10^{-8}$ |
| 20 | $1.33\times10^{-8}$ | $1.78\times10^{-8}$ | $6.36\times10^{-8}$ | $1.96\times10^{-8}$ |
| 40 | $8.91\times10^{-9}$ | $1.19\times10^{-8}$ | $7.95\times10^{-8}$ | $1.39\times10^{-8}$ |
| 80 | $5.80\times10^{-9}$ | $7.95\times10^{-9}$ | $1.03\times10^{-7}$ | $9.91\times10^{-9}$ |
| 160 | $3.82\times10^{-9}$ | $5.40\times10^{-9}$ | $1.37\times10^{-7}$ | $6.92\times10^{-9}$ |

The L2 results reveal two distinct behaviors (see Table 8). For local perturbations (adjacent, non-adjacent, and middle swaps), structural sensitivity decreased approximately in proportion to $1/n$. In contrast, the reversal condition remained relatively stable across document lengths, indicating persistent sensitivity to global structural changes.

Table 8 L2 response under sentence reordering.

| Length | Adjacent | Non-adjacent | Reverse | Middle |
|---|---|---|---|---|
| 5 | 0.0789 | 0.1548 | 0.1614 | 0.0789 |

| | | | | |
|---|---|---|---|---|
| 10 | 0.0351 | 0.0688 | 0.1481 | 0.0325 |
| 20 | 0.0166 | 0.0326 | 0.1639 | 0.0156 |
| 40 | 0.0081 | 0.0159 | 0.1730 | 0.0077 |
| 80 | 0.0040 | 0.0078 | 0.1741 | 0.0037 |
| 160 | 0.0020 | 0.0039 | 0.1786 | 0.0018 |

A similar pattern was observed for L3, as shown in Table 9. Local perturbations weakened as document length increased, whereas reversal remained nearly constant across all lengths.

Table 9 L3 response under sentence reordering.

| **Length** | **Adjacent** | **Non-adjacent** | **Reverse** | **Middle** |
|---|---|---|---|---|
| 5 | 1.040 | 1.080 | 1.264 | 1.040 |
| 10 | 0.462 | 0.480 | 1.544 | 0.473 |
| 20 | 0.219 | 0.227 | 1.533 | 0.224 |
| 40 | 0.107 | 0.111 | 1.528 | 0.110 |
| 80 | 0.0527 | 0.0547 | 1.531 | 0.0549 |
| 160 | 0.0262 | 0.0272 | 1.533 | 0.0272 |

Overall, Experiment 5 shows that L2 and L3 encode local sequential relationships that are sensitive to sentence reordering in long documents. However, fixed local perturbations become diluted approximately as $1/n$, whereas global structural disruption remains clearly detectable.

Experiment 5 also demonstrated the computational practicality of the framework. The experiment processed approximately 79,592 unique sentences using an NVIDIA RTX A400 GPU. SBERT encoding required approximately 19.25 minutes (72.4 sentences/s), while the subsequent L1/L2/L3 computations required only a few seconds, as shown in Table 10.

Table 10 Computation time for L1/L2/L3.

| **Length** | **L1/L2/L3 calculation** |
|---|---|
| 5 | 0.408 s |
| 10 | 0.420 s |
| 20 | 0.575 s |
| 40 | 0.793 s |
| 80 | 2.386 s |
| 160 | 2.523 s |

These results indicate that the computational cost is dominated by sentence encoding, while the proposed multi-level analysis introduces relatively little additional overhead.

(3) Experiment 6

Experiment 6 investigated whether the length effect observed in Experiment 5 was primarily a dilution effect, separated the influence of document length from perturbation density, and further validated the roles of L1, L2, and L3.

The experiment used the same 500 validated WikiText-103 articles as Experiment 5. Document lengths of 10, 20, 40, 80, and 160 sentences were evaluated under seven conditions: the original document (P0), a single adjacent swap (P1), 5% perturbation (P2), 10% perturbation (P3), 25% perturbation (P4), 50% perturbation (P5), and complete reversal (P6). Sentences were encoded using SBERT all-mpnet-base-v2, with SBERT-Mean serving as a control baseline.

L1 and SBERT-Mean remained approximately zero across all conditions, indicating that they are largely insensitive to sentence-order perturbations. For a fixed one-swap perturbation (P1), both L2 and L3 decreased as document length increased, as shown in Table 11.

Table 11 Effect of document length under P1 (fixed one-swap).

| Length (n) | L2 | L3 |
|---|---|---|
| 10 | 0.0351 | 0.463 |
| 20 | 0.0166 | 0.219 |
| 40 | 0.0081 | 0.107 |
| 80 | 0.0040 | 0.0527 |
| 160 | 0.0020 | 0.0262 |

As document length approximately doubled, the structural signal approximately halved, suggesting that both L2 and L3 decrease roughly in proportion to $1/n$ for fixed local perturbations.

Increasing perturbation density restored the structural signal. At $n = 160$, both L2 and L3 increased steadily as the proportion of perturbed sentences increased, as shown in Table 12. In our implementation, each sentence transition is represented by the difference between consecutive sentence embeddings. The L3 response is the mean Euclidean distance between corresponding transition vectors in the original and perturbed paragraphs. Because these vectors are not normalized, the resulting L3 distance can exceed 1.

Table 12. Effect of perturbation density ($n$ = 160).

| Perturbation | L2 | L3 |
|---|---|---|
| Fixed 1 swap | 0.0020 | 0.0262 |
| 5% | 0.0055 | 0.0934 |
| 10% | 0.0131 | 0.2023 |
| 25% | 0.0354 | 0.5286 |
| 50% | 0.0721 | 1.0705 |
| 100% reversal | 0.1801 | 1.5351 |

The results show a clear dose-response pattern: larger perturbation densities produce stronger L2 and L3 responses. Fixed local perturbations become diluted as documents grow longer, whereas global structural disruption remains readily detectable.

Overall, Experiment 6 demonstrates that the decline in L2/L3 sensitivity with increasing document length is primarily a dilution effect rather than an inherent limitation of the proposed representation. When perturbation density is maintained, both measures, particularly L3, continue to provide strong and scalable structural signals.

### 5.3. Weighting Sensitivity, Reproducibility, and Benchmarking

(1) Experiment 7

Experiment 7 was designed to evaluate component effectiveness, statistical reproducibility, and fusion-weight sensitivity.

Experiment 7A used 1,000 new paragraph pairs encoded with SBERT all-mpnet-base-v2. Individual components and multiple fusion strategies were evaluated to assess the contributions of L1, L2, and L3.

Table 13 Experiment 7A results.

| Method | Pearson | Spearman |
|---|---|---|
| L1 | **0.7848** | **0.8781** |
| L2 | 0.4514 | 0.5198 |
| L3 | 0.5949 | **0.8651** |
| L1 + L2 | **0.8115** | 0.8822 |
| L1 + L3 | 0.7340 | 0.8872 |
| L2 + L3 | 0.6607 | 0.8360 |
| Equal 3-level | 0.7610 | 0.8857 |
| 40/10/50 | 0.7334 | **0.8880** |

L1 achieved strong standalone performance, while L3 exhibited substantially higher Spearman than Pearson correlation, suggesting stronger rank-order preservation. The best Pearson correlation was obtained by L1+L2 (0.8115), indicating that L2 provides complementary information beyond L1. The default 40/10/50 fusion achieved the highest Spearman correlation (0.8880) (see Table 13).

Experiment 7B evaluated statistical reproducibility using 5,000 bootstrap resamples. The bootstrap estimates closely matched the original results, supporting the stability and reliability of the reported correlations, as shown in Table 14.

Table 14 Experiment 7B bootstrap validation for Pearson and Spearman.

| Method | Exp7A P | Bootstrap P | Exp7A S | Bootstrap S |
|---|---|---|---|---|
| L1 | 0.7848 | 0.7855 | 0.8781 | 0.8775 |
| L2 | 0.4514 | 0.4501 | 0.5198 | 0.5195 |
| L3 | 0.5949 | 0.5991 | 0.8651 | 0.8646 |
| L1+L2 | 0.8115 | 0.8119 | 0.8822 | 0.8816 |
| L1+L3 | 0.7340 | 0.7357 | 0.8872 | 0.8867 |
| Equal 3 | 0.7610 | 0.7623 | 0.8857 | 0.8852 |
| 40/10/50 | 0.7334 | 0.7352 | 0.8880 | 0.8874 |

Across all methods, bootstrap estimates differed from the original results only marginally, typically in the third decimal place. For example, the 95% bootstrap confidence intervals for L1 were [0.7607, 0.8090] for Pearson and [0.8627, 0.8912] for Spearman, indicating good reproducibility.

Experiment 7C investigated fusion-weight sensitivity using 25 weight combinations. L1 weights ranged from 30% to 50%, while L2 weights ranged from 0% to 20%, with the remaining weight assigned to L3. The corresponding Pearson and Spearman results are illustrated in Figures 3 and 4.

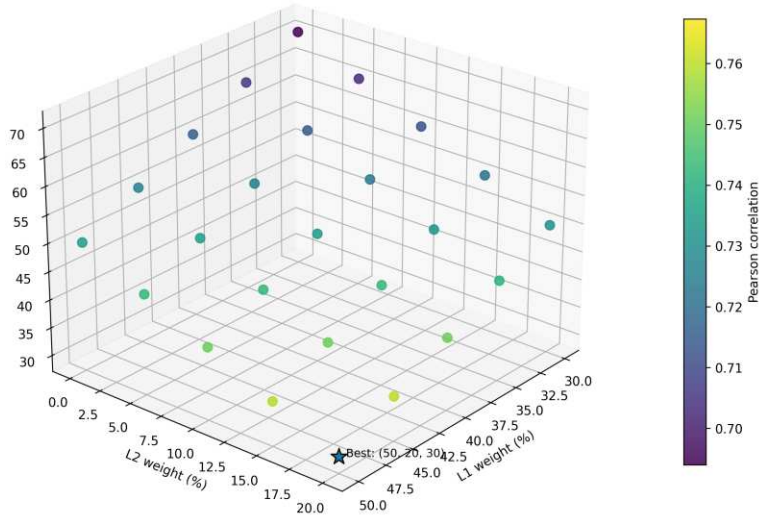



**Fig. 3.** Analysis of Pearson weight sensitivity, the best combination is highlighted.

The best Spearman performance was obtained with a 45/10/45 weight configuration, achieving Pearson = 0.7425 and Spearman = 0.8881. The best Pearson performance was achieved by 50/20/30, yielding Pearson = 0.7673 and Spearman = 0.8878. Overall, the relatively small variation across high-performing configurations suggests that the proposed framework is reasonably robust to moderate changes in fusion weights.

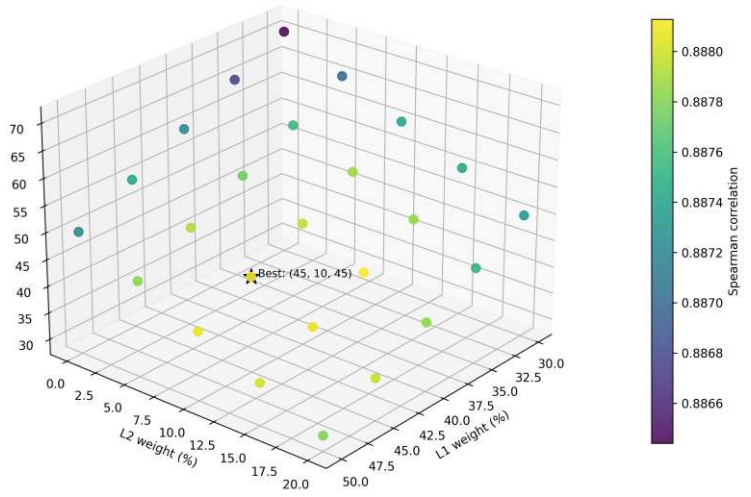



**Fig. 4.** Analysis of Spearman weight sensitivity, the best combination is highlighted.

(2) Experiment 8

Experiment 8 had two objectives: benchmarking against established paragraph representations and statistically evaluating the observed performance differences.

Experiment 8A used the same 1,000 paragraph pairs as Experiment 7 and compared three fusion configurations (45/10/45, 40/10/50, and equal three-level fusion) against four benchmark methods: SBERT Mean, Doc2Vec PV-DBOW, Doc2Vec PV-DM, and Sent2Vec Mean. The results were presented in Table 15.

Table 15 Experiment 8A benchmark comparison results.

| Method | Pearson | Spearman |
|---|---|---|
| **SBERT Mean** | **0.8736** | 0.8876 |
| L1 | 0.7848 | 0.8781 |
| L2 | 0.4514 | 0.5198 |
| L3 | 0.5949 | 0.8651 |
| **45/10/45 Fusion** | 0.7425 | **0.8881** |
| 40/10/50 Fusion | 0.7334 | 0.8880 |
| Equal 3-Level Fusion | 0.7610 | 0.8857 |
| Doc2Vec PV-DBOW | 0.7688 | 0.7686 |
| Doc2Vec PV-DM | 0.7598 | 0.7658 |
| Sent2Vec Mean | 0.7140 | 0.7270 |

SBERT Mean achieved the highest Pearson correlation (0.8736), while the 45/10/45 fusion achieved the highest Spearman correlation (0.8881). These results suggest that SBERT Mean provides the strongest linear agreement with the reference scores, whereas the proposed framework offers slightly better rank-order agreement. Both the proposed fusion methods and SBERT-based representations substantially outperformed the Doc2Vec and Sent2Vec baselines in Spearman.

Experiment 8B performed paired bootstrap comparisons between the best-performing fusion model (45/10/45) and the benchmark methods, as shown in Table 16.

Table 16 Experiment 8B paired bootstrap comparison.

| Comparison | Pearson Diff. | 95% CI | Spearman Diff. | 95% CI |
|---|---|---|---|---|
| Fusion − SBERT Mean | −0.1293 | [−0.1678, −0.0899] | +0.00044 | [−0.0107, +0.0109] |
| Fusion − Doc2Vec PV-DBOW | −0.0241 | [−0.0690, +0.0231] | +0.1196 | [+0.0874, +0.1537] |

| | | | | |
|---|---|---|---|---|
| Fusion − Doc2Vec PV-DM | −0.0153 | [−0.0562, +0.0280] | +0.1221 | [+0.0932, +0.1527] |
| Fusion − Sent2Vec Mean | +0.0306 | [−0.0207, +0.0829] | +0.1611 | [+0.1300, +0.1943] |

The bootstrap analysis showed that the 45/10/45 fusion achieved significantly lower Pearson correlation than SBERT Mean, but no statistically significant difference in Spearman correlation. In contrast, the fusion model achieved substantially higher Spearman correlations than all Doc2Vec and Sent2Vec baselines, with confidence intervals entirely above zero.

Overall, Experiment 8 indicates that the proposed state-continuity-transition representation achieves rank-order performance comparable to strong SBERT mean pooling while substantially outperforming the evaluated Doc2Vec and Sent2Vec baselines.

## 6. Analysis and Discussion

This section discusses the main findings from the experiments and their implications for multi-level paragraph representation. It considers structural sensitivity and document length, followed by the effects of encoder choice, benchmark comparisons, and computational considerations.

### 6.1. Multi-Level Representation Is Meaningful

The experimental results support the view that paragraph similarity contains both semantic-content and semantic-organization information. Across the similarity experiments, L1 generally provides the strongest representation of semantic content, particularly with the SBERT encoder. In Experiment 7A, for example, L1 achieved Pearson and Spearman correlations of 0.7848 and 0.8781, respectively. This confirms that sentence-level semantic states provide a strong basis for paragraph-level similarity.

By contrast, L2 is weaker as an independent similarity measure, but its relatively low correlation with L1 and L3 indicates that it captures information that is not fully redundant with the other levels. In Experiment 1, the correlations between L2 and L1/L3 were only 0.332 and 0.370, compared with 0.831 between L1 and L3. This suggests that semantic continuity provides a distinct structural signal despite its weaker standalone performance.

L3 shows particular sensitivity to semantic movement and sentence reordering. Experiments 3–6 demonstrate that transition-based representation responds strongly to changes in sentence order, while L1 remains essentially unchanged under pure reordering. Thus, L3 contributes information about how semantic content progresses through a paragraph rather than only what content is present.

Taken together, these findings support multi-level representation as a complementary rather than redundant strategy. The fusion results show that combining semantic content with continuity and transition information can retain aspects of paragraph structure that single-level representations do not capture. The benefit is therefore not simply a stronger semantic representation, but a representation that reflects both what a paragraph contains and how its semantic content is organized and progresses.

### 6.2. Structural Sensitivity and Document Length

Experiments 3–6 show that the proposed representation is sensitive not only to semantic content but also to changes in sentence organization. In Experiment 3, sentence reordering produced little change in L1, whereas L2 and particularly L3 responded to the resulting structural disruption. Intruder replacement further increased the changes across the levels, indicating that semantic and structural perturbations affect the representation differently.

The longer-document experiments in Experiments 5 and 6 reveal an important length effect. For a fixed local perturbation, both L2 and L3 responses decreased approximately in proportion to $1/n$, where $n$ is the number of sentences. This reflects perturbation dilution: as the document becomes longer, a fixed number of altered sentence relations contributes a smaller fraction of the overall representation. In contrast, increasing the perturbation density produced progressively larger structural signals, while global reversal remained clearly detectable even for long documents.

These results indicate that structural sensitivity depends jointly on the type and density of the perturbation and the length of the document. Local changes become less influential as paragraph length increases, whereas widespread or global reordering remains detectable. This provides an important qualification for interpreting

structural similarity in longer texts.

### 6.3. Encoder, Benchmark, and Computational Considerations

The experiments also highlight the importance of the underlying sentence encoder. Experiment 2 showed that replacing Sent2Vec with SBERT substantially improved the semantic-state representation, with L1 increasing from 0.643/0.680 in Experiment 1 to 0.7603/0.8685 in Pearson/Spearman correlation. This indicates that the proposed framework is influenced by encoder quality and should therefore be viewed as a representation framework built on top of sentence embeddings rather than as an encoder-independent method.

Experiment 8 further shows that the proposed multi-level representation provides a different performance profile from conventional paragraph representations. The 45/10/45 fusion achieved a Spearman correlation of 0.8881, compared with 0.8876 for SBERT-Mean, while its Pearson correlation was lower (0.7425 versus 0.8736). Paired bootstrap analysis in Experiment 8B indicates that the Spearman difference was not statistically distinguishable from zero, whereas the Pearson difference was significant. Thus, the results support complementary structural information rather than a general performance advantage over paragraph-level SBERT. Experiments 7A and 7C and the bootstrap analysis in Experiment 7B also show that the main findings are reasonably stable across resampling and moderate changes in fusion weights.

Finally, Experiment 5 indicates that the additional multi-level calculations introduce relatively limited computational overhead compared with sentence encoding. For approximately 79,592 unique sentences, SBERT encoding required about 19.25 minutes, whereas the subsequent L1/L2/L3 calculations for the tested document sets required only seconds. With sentence embeddings cached and reused, the proposed structural analysis is therefore computationally practical for the tested document lengths, although the cost of sentence encoding and OT computation remains relevant for larger-scale applications.

## 7. Conclusion and Future Work

### 7.1. Main Findings and Contributions

This study investigated paragraph representation beyond the conventional single-vector paradigm by modeling a paragraph as a sequence of semantic states. The experimental results show that L1 provides a strong representation of semantic content, while L2 and L3 contribute complementary information about semantic continuity and transitions. Although L2 is weaker as a standalone similarity measure, its relatively low correlation with the other levels indicates that it captures distinct structural information. L3, in particular, provides sensitivity to semantic movement and sentence ordering.

The results further show that multi-level fusion can preserve structural information that order-invariant paragraph representations overlook. Experiments 3–6 demonstrate clear sensitivity to sentence reordering and other structural perturbations, while the similarity experiments show that the fused representation remains competitive in rank-based similarity evaluation. Although the results do not establish universal superiority over SBERT-Mean, they support the value of explicitly modeling sequential semantic structure as a complement to conventional paragraph-level representations.

### 7.2. Future Work

Several directions follow from the current study. First, the representation and fusion strategy could be further developed by learning the fusion weights rather than specifying them manually, and by exploring richer formulations of semantic continuity and transition. Second, the dependence on the underlying sentence encoder and embedding geometry warrants investigation using additional encoders and alternative embedding spaces. Finally, broader evaluation across datasets and downstream tasks would help assess generality, while different sentence-segmentation strategies and more efficient implementations could improve the applicability of the framework to longer and larger-scale documents.

## References

(Agirre et al 2012) E. Agirre, D. Cer, M. Diab and A. Gonzalez-Agirre, (2012) “SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity,” SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Montreal, Canada, pp. 385-393.

(Apidianaki 2023) M. Apidianaki, (2023) “From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation,” Computational Linguistics, vol. 49, no. 2, pp. 465-523.

(Arora et al 2017) S. Arora, Y. Liang, and T. Ma, (2017) “A simple but tough-to-beat baseline for sentence embeddings,” 5th International Conference on Learning Representations, Toulon, France.

(Barzilay and Lapata 2008) R. Barzilay and M. Lapata, (2008) “Modeling Local Coherence: An Entity-Based Approach,” Computational Linguistics, vol. 34, no. 1, pp. 1-34.

(Bojanowski 2017) P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, (2017) “Enriching Word Vectors with Subword Information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146.

(Cer et al 2018) D. Cer et al., (2018) “Universal Sentence Encoder for English,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 169-174.

(Cuturi 2013) M. Cuturi, (2013) “Sinkhorn Distances: Lightspeed Computation of Optimal Transport,” in Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, pp. 2292-2300.

(EleutherAI 2023) EleutherAI, (2023) “WikiText Document Level,” Hugging Face, Mar. 2023. [Online]. Available: https://huggingface.co/datasets/EleutherAI /wikitext_document_level. [Accessed: Oct. 2, 2026].

(Farag et al 2020) Y. Farag, J. Valvoda, H. Yannakoudakis, and T. Briscoe, (2020) “Analyzing Neural Discourse Coherence Models,” in Proceedings of the First Workshop on Computational Approaches to Discourse, Online, pp. 102-112.

(Glick 2023) O. Glick, (2023) “DiscoEval Benchmark Datasets,” Hugging Face, Nov. 2023. [Online]. Available: https://huggingface.co/datasets/OfekGlick/DiscoEval. [Accessed: Oct. 2, 2026].

(Huang 2023) J. Y. Huang et al., (2023) “Bridging Continuous and Discrete Spaces: Interpretable Sentence Representation Learning via Compositional Operations,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 14584-14595.

(Iter et al 2020) D. Iter, K. Guu, L. Lansing, and D. Jurafsky, (2020) “Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 4859-4870.

(Jin 2025) X. Jin, (2025) “Improving Paragraph Similarity by Sentence Interaction With BERT,” Expert Systems, vol. 42, no. 3.

(Jupally 2024) M. Jupally, (2024) “ROCStories,” Hugging Face, May 2024. [Online]. Available: https://huggingface.co/datasets/mintujupally/ROCStories. [Accessed: Oct. 2, 2026].

(Kusner et al 2015) M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, (2015) “From Word Embeddings to Document Distances,” in Proceedings of the 32nd International Conference on Machine Learning, vol. 37, Lille, France, pp. 957-966.

(Lapata and Barzilay 2005) M. Lapata and R. Barzilay, (2005) “Automatic Evaluation of Text Coherence: Models and Representations,” in Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1085-1090.

(Le and Mikolov 2014) Q. Le and T. Mikolov, (2014) “Distributed Representations of Sentences and Documents,” in Proceedings of the 31st International Conference on Machine Learning, vol. 32, no. 2, Beijing, China, pp. 1188-1196.

(Liu 2020) S. Liu, S. Zeng, and S. Li, (2020) “Evaluating Text Coherence at Sentence and Paragraph Levels,” in Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, pp. 1695-1703.

(Maslak and Mitkov 2021) H. Maslak and R. Mitkov, (2021) “Paragraph Similarity Matches for Generating Multiple-choice Test Items,” in Proceedings of the Student Research Workshop Associated with RANLP 2021, Online, pp. 99-108.

(Oh et al 2019) B. Oh et al., (2019) “Topic-Guided Coherence Modeling for Sentence Ordering by Preserving Global and Local Information,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, pp. 2273-2283.

(Pagliardini et al 2018) M. Pagliardini, P. Gupta, and M. Jaggi, (2018) “Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, New Orleans, Louisiana, USA, pp. 528-540.

(Pennington et al 2014) J. Pennington, R. Socher, and C. Manning, (2014) “GloVe: Global Vectors for Word Representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, pp. 1532-1543.

(Peyré and Cuturi 2019) G. Peyré and M. Cuturi, (2019) “Computational Optimal Transport,” Foundations and Trends in Machine Learning, vol. 11, no. 5-6, pp. 355-607.

(Prabhumoye et al 2020) S. Prabhumoye, R. Salakhutdinov, and A. W. Black, (2020) “Topological Sort for Sentence Ordering,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 2783-2792.

(Pszemraj 2026) Pszemraj, (2026) “Synthetic Text Similarity,” Hugging Face. [Online]. Available: https://huggingface.co/datasets/pszemraj/synthetic-text-similarity. [Accessed: Oct. 2, 2026].

(Reimers and Gurevych 2019) N. Reimers and I. 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, Hong Kong, China, pp. 3982-3992.

(Wu et al 2021) C. Wu, F. Wu, T. Qi, and Y. Huang, (2021) “Hi-Transformer: Hierarchical Interactive Transformer for Efficient and Effective Long Document Modeling,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, pp. 848-853.

(Xu et al 2019) P. Xu et al., (2019) “A Cross-Domain Transferable Neural Coherence Model,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 678-687.

(Yang et al 2016) Z. Yang et al., (2016) “Hierarchical Attention Networks for Document Classification,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480-1489, San Diego, California.