stats

MDPI

Communication

Ordering and Quantifying Textual Cohesion via Semantic, Geometric and Statistical Structure

Stelios Arvanitis

Department of Economics, Athens University of Economics and Business, 10434 Athens, Greece; stelios@aueb.gr; Tel.: +30-2108203437

Abstract

We propose a semantic, geometric, and statistical framework for quantifying and ordering textual cohesion in long-form discourse. Sentences are embedded into a semantic similarity graph and Ollivier–Ricci curvature is used to extract sentence- and document-level structural profiles, represented as step functions on a normalized rhetorical-time axis. On this functional space we define the Weighted Utopia Index (wUI), a corpus-relative measure of weighted shortfall from an upper-envelope profile under a dominance-type ordering. The rhetorical-time weighting function is learned self-supervised: we generate controlled sentence-order perturbations with known ordinal coherence degradation and estimate the weight parameters via an ordered probit model on a training split. We evaluate ordering recovery on held-out State of the Union speeches using rank correlations, pairwise and adjacent ordering accuracy, and violation-localization diagnostics with bootstrap uncertainty. Across these criteria, wUI systematically outperforms embedding-only adjacent-similarity baselines, while a Nash-type aggregation provides an interpretable semantic–structural trade-off score. An application to later-period speeches illustrates how the method yields interpretable cohesion rankings and curvature-profile diagnostics without requiring external annotations.

Keywords: textual cohesion; discourse structure; semantic similarity graphs; Ollivier–Ricci curvature; Weighted Utopia Index; self-supervised evaluation; ordered probit; ordinal recovery

1. Introduction

Quantifying textual coherence—the degree to which a text exhibits consistent semantic and structural progression—is central to applications ranging from discourse analysis and summarization to automated text evaluation and generation. While recent neural approaches have achieved strong performance by learning sentence-level plausibility from large corpora, they often lack explicit representations of discourse structure and provide limited interpretability regarding how coherence is formed across a document.

We propose a framework for measuring textual coherence that integrates semantic representations with geometric structure extracted from sentence similarity graphs. Sentences are embedded using pretrained transformer models and represented as weighted graphs, from which we compute Ollivier–Ricci curvature to capture structural smoothness, bottlenecks, and transitions in discourse flow. This yields a curvature profile along the rhetorical progression of a text. We introduce a statistically trained structural alignment score that compares a text's curvature profile to corpus-level reference envelopes under a dominance-type ordering. Specifically, we define the Weighted Utopia Index (wUI) as a normalized, weighted shortfall functional over rhetorical time. Crucially, the weighting

Check for updates

Academic Editor: Wei Zhu

Received: 12 January 2026
Revised: 12 February 2026
Accepted: 24 February 2026
Published: 3 March 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

function is not specified ad hoc. Instead, it is learned in a self-supervised manner from known ordinal coherence structure induced by controlled sentence-order perturbations. We generate document variants of increasing permutation severity, which yields an exact ground-truth ordering of coherence by construction, and estimate the rhetorical-time weighting parameters via an ordered probit model on a training split. This links a latent coherence utility to observed ordinal categories determined by perturbation severity, yielding a probabilistically grounded and internally identified weighting profile. The learned weighting function is itself interpretable and reveals which portions of a document's rhetorical progression are most informative for coherence discrimination.

Evaluation is conducted entirely within the same corpus family, avoiding cross-genre transfer and external annotation noise. Using held-out documents, we assess how well different cohesion measures recover the known ordinal ordering induced by perturbations. We report rank correlations, pairwise and adjacent ordering accuracy, violation-localization diagnostics, and bootstrap confidence intervals. Across these criteria, the proposed wUI consistently outperforms embedding-only adjacency baselines, while a Nash-type aggregation provides an interpretable semantic–structural trade-off score.

An application to later-period political speeches illustrates how the framework can be used as an interpretive tool rather than a validation exercise. By decomposing cohesion into semantic continuity and global structural alignment, the method reveals rhetorical patterns that are not visible to local similarity measures alone.

The proposed framework is interpretable, modular, and statistically grounded, offering a geometry-aware alternative to purely neural coherence models and a principled evaluation paradigm based on exact ordinal supervision.

2. Materials and Methods

For a finite corpus of texts $\mathcal{D} = \{d_1, \ldots, d_N\}$, where each text d_k consists of a non-empty ordered sequence of sentences $\mathcal{S}_k = \{s_1^{(k)}, \ldots, s_{n_k}^{(k)}\}$, the goal is to construct numerical scores that approximate the internal cohesion of each d_k by analyzing semantic and geometric information over $\mathcal{S}_k$.

The current section develops the methodological pipeline. We begin by constructing semantic similarity graphs using SBERT-based sentence embeddings (see [1]), linking sentences via cosine similarity to form weighted, undirected graphs in the spirit of TextRank [2]. The graph-theoretic Ollivier–Ricci curvature (see [3,4]) is then computed on these graphs using optimal transport distances between neighborhood distributions, yielding sentence-level curvature profiles, represented by text-by-text comparable step function embeddings of the curvature vectors.

To assess structural alignment using information from the curvature profiles, we introduce a "reverse First order Dominance-type" order relation and produce a Weighted Utopia Index (wUI). The latter is an adaptation of the Utopia Index associated with Second Order Stochastic Dominance and introduced in [5]. Given the set of curvature profiles for the corpus, the wUI quantifies the severity of the obstacles to order maximality of a speech's curvature profile against corpus-wide bounds.

The wUI is computed using a learned weight function

$$w_\theta(x) = \exp\big(\theta_1 x + \theta_2 x^2\big),$$

where $\theta \in \mathbb{R}^2$ is estimated via an ordered probit model under self-supervised ordinal supervision. Ordinal labels are induced by controlled sentence-order perturbations of increasing severity, which generate an exact ground-truth ordering of coherence by construction. The

probit likelihood links a latent coherence utility indexed by the wUI to these severity-based ordinal categories, yielding a statistically identified weighting function over rhetorical time.

This learning procedure endows the wUI with sensitivity to structural coherence patterns that are empirically informative for distinguishing progressively degraded discourse within a homogeneous corpus. The resulting weight function can be interpreted as a data-driven estimate of which rhetorical regions contribute most to coherence discrimination under the induced ordinal ordering. To complement structural metrics, we compute the SBERT-based Coherence Score (SBERT-CS), which measures semantic continuity as the average cosine similarity between adjacent SBERT-sentence embeddings. This average partially captures local flow and topical progression across the discourse, leveraging contextual sentence representations produced by a pre-trained transformer.

Finally, the Nash cohesion score is defined as the geometric mean of the SBERT-based coherence score and the Weighted Utopia Index. It summarizes a trade-off between semantic plausibility and structural alignment. This can be thought of as a Pareto-scalarization of the two-dimensional score obtained from vectorising SBERT-CS with wUI.

2.1. TextRank and SBERT Semantic Graph

Let $\mathcal{S}_k = \{s_1^{(k)}, \ldots, s_{n_k}^{(k)}\}$ be a finite set of sentences extracted from an arbitrary document in the corpus. The goal of the TextRank algorithm [2] is to compute a score $r_i^{(k)}$ for each sentence $s_i^{(k)}$, reflecting its importance in terms of semantic similarity to other sentences.

The relationships between sentences are represented using an undirected weighted graph $G^{(k)} = (V^{(k)}, E^{(k)}, W^{(k)})$, where:

- $V^{(k)} = \mathcal{S}_k$ is the set of nodes (sentences);
- $(s_i^{(k)}, s_j^{(k)}) \in E^{(k)}$ if and only if $i \neq j$ and $\text{sim}(s_i^{(k)}, s_j^{(k)}) > \tau$ where sim represents some similarity index between sentences (to be defined in the following paragraph), and $\tau \in [0, 1]$ is a similarity threshold;
- $w^{(k)} : E^{(k)} \to \mathbb{R}_+$ assigns edge weights $w_{ij}^{(k)} = \text{sim}(s_i^{(k)}, s_j^{(k)})$.

Similarity via Sentence Embeddings with SBERT

To capture sentence-level semantic meaning, and thus define the aforementioned sim function, we utilize Sentence-BERT (SBERT), a modification of the original Google's BERT (see [6]) architecture specifically designed to produce high-quality embeddings for sentence similarity tasks [1]. While BERT (Bidirectional Encoder Representations from Transformers) achieves state-of-the-art performance in many NLP benchmarks, it is aknowledged that it is not directly optimized for measuring sentence-to-sentence similarity. Its limitations are associated to the fact that BERT produces contextualized token embeddings, not sentence embeddings. Aggregating these into a fixed-size representation (e.g., using mean pooling) yields inconsistent semantic distances. Furthermore, in similarity tasks, vanilla BERT requires joint input of any pair of two sentences which can be computationally inefficient for large-scale or many-to-many comparisons.

The SBERT modification addressed these issues by introducing a Siamese network architecture (see [7]) that encodes each sentence independently using a shared transformer model. Sentence embeddings are then obtained by applying a pooling operation (typically mean-pooling) over the output token embeddings. Each sentence $s^{(k)} \in \mathcal{S}_k$ is mapped to a vector representation $\mathbf{v}_s^{(k)} \in \mathbb{R}^d$. The resulting fixed-size vectors lie in a semantically meaningful space, allowing cosine similarity to approximate semantic sentence similarity.

SBERT is fine-tuned on large-scale natural language inference (NLI) and semantic textual similarity (STS) datasets. This training encourages the model to place semantically similar sentences near each other in the embedding vector space. The resulting

architecture supports fast, meaningful, and interpretable vector-based comparisons, making it highly suitable for applications in coherence evaluation, clustering, and semantic graph construction.

Given then an arbitrary pair of SBERT sentence embedding vectors $\mathbf{v}_{s_i^{(k)}}, \mathbf{v}_{s_j^{(k)}}$, the edge weight between them is defined as:

$$\mathbf{w}_{ij}^{(k)} = \mathrm{sim}(s_i^{(k)}, s_j^{(k)}) := \frac{\langle \mathbf{v}_{s_i^{(k)}}, \mathbf{v}_{s_j^{(k)}} \rangle}{\|\mathbf{v}_{s_i^{(k)}}\|_2 \cdot \|\mathbf{v}_{s_j^{(k)}}\|_2}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2$ respectively denote the Euclidean inner product and norm. The embeddings are then used to define the Edge set $E^{(k)}$ and the final weight function $W^{(k)}$ that complete the construction of the TextRank-type graph; namely, the edge $e_{ij}^{(k)}$ is retained in $E^{(k)}$ if and only if $\mathbf{w}_{ij}^{(k)} > \tau$ for a fixed threshold where $\tau \in (0, 1)$ is a similarity threshold controlling graph sparsity; optionally the top-k most similar neighbors could be retained for each node in order to enforce k-sparsity. The reduced weight function is then $w_{ij}^{(k)} := \begin{cases} \mathbf{w}_{ij}^{(k)}, \text{ if } \mathbf{w}_{ij}^{(k)} > \tau \\ 0, \text{ if } \mathbf{w}_{ij}^{(k)} \leq \tau \end{cases}$.

For each text d_k text in the corpus at hand, the resulting SBERT-based sentence-based TextRank graph can be interpreted as a sentence-level semantic network. Its topological and geometrical properties could reflect semantic coherence which could be recovered using graph curvature and/or entropy measures.

2.2. Ollivier–Ricci Curvature and Dominance on Sentence Graphs

The TextRank graph is constructed over sentence embeddings with a geometric notion of curvature to reveal the structural properties of the sentence network for each text in the corpus. In particular, the notion of Ollivier–Ricci curvature [4] is employed, adapted to weighted graphs.

2.2.1. Metric Measure Space Structure

Let $G^{(k)} = (V^{(k)}, E^{(k)}, w^{(k)})$ be the undirected weighted graph defined in the previous paragraph. To each node (sentence) $i \in V^{(k)}$ a discrete probability distribution $\mu_i^{(k)}$ supported its neighbors is defined via:

$$\mu_i^{(k)}(j) := \begin{cases} \frac{w_{ij}^{(k)}}{\sum_{m \in \mathcal{N}(i)} w_{im}^{(k)}}, & \text{for } j \in \mathcal{N}(i) \\ 0, & \text{for } j \notin \mathcal{N}(i) \end{cases},$$

where $\mathcal{N}(i)$ denotes the neighborhood of i inside $G^{(k)}$. This makes $(V^{(k)}, d_{G_{(k)}}, \{\mu_i^{(k)}, i \in V^{(k)}\})$ a discrete metric measure space, where $d_{G_{(k)}}(i, j)$ is the shortest-path distance on $G^{(k)}$, and $\{\mu_i^{(k)}, i \in V^{(k)}\}$ is the process on $V^{(k)}$ comprised of the abovementioned discrete measures.

2.2.2. Between-Nodes Ollivier–Ricci Curvature

Following [4], the Ollivier–Ricci curvature $\kappa(i, j)$ along the edge (i, j) in $G^{(k)}$ is defined via optimal transport between the measures $\mu_i^{(k)}$ and $\mu_j^{(k)}$:

$$\kappa^{(k)}(i, j) := 1 - \frac{\mathcal{W}_1(\mu_i^{(k)}, \mu_j^{(k)})}{d_{G_{(k)}}(i, j)},$$

where $\mathcal{W}_1$ is the 1-Wasserstein distance between $\mu_i^{(k)}$ and $\mu_j^{(k)}$; the latter in the present discrete framework is defined as $\mathcal{W}_1(\mu_i^{(k)}, \mu_j^{(k)}) := \min_{\gamma \in \Gamma(\mu_i^{(k)}, \mu_j^{(k)})} \sum_{(m,n) \in \mathcal{N}_i \times \mathcal{N}_j} \gamma(m,n) d_{G_{(k)}}(m,n)$, where $\Gamma(\mu_i^{(k)}, \mu_j^{(k)})$ denotes the set of Borel probability distributions on $\mathcal{N}_i \times \mathcal{N}_j$ that have respective marginals $\mu_i^{(k)}, \mu_j^{(k)}$ (see [8]). The computation of the Wasserstein distance and thus of the between-nodes curvature has a convenient LP representation—see for example Chapter 8 of [9].

The curvature quantifies how well the neighborhoods of i and j are aligned. Positive curvature could be interpreted as local cohesion, while negative curvature may imply structural bottlenecks or bridges. Edges with negative curvature identify transitions across semantically distant or structurally weakly connected regions. By averaging $\kappa(i,j)$ over incident edges, we define a sentence-level curvature score:

$$\bar{\kappa}^{(k)}(i) := \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \kappa^{(k)}(i,j),$$

which captures the local semantic structural integrity around sentence $s_i^{(k)}$. High positive values may indicate redundant or predictable contexts, while negative scores may indicate pivotal sentences or "non-smooth" transitions.

Intuition for Discourse Analysis

For NLP practitioners, Ollivier–Ricci curvature can be interpreted as a comparison of local semantic neighborhoods. For a sentence i, consider the probability distribution over its neighbors, where mass is proportional to cosine similarity. If two adjacent sentences i and j have neighborhoods that are themselves semantically similar, the optimal transport cost between these two distributions is small, yielding positive or near-zero curvature. This corresponds to locally redundant or smoothly evolving discourse.

By contrast, if the neighborhoods differ substantially—for example, when a sentence acts as a bridge between otherwise weakly connected semantic regions—the transport cost is large and curvature becomes negative. In discourse terms, negative curvature signals a semantic bottleneck, topic shift, or structural transition in the flow of ideas.

Thus, curvature does not measure lexical similarity directly; rather, it captures whether local semantic contexts evolve smoothly or introduce structural breaks in the sentence graph.

2.2.3. Step-Function Embedding of Curvature Profiles

Each sentence-level Ricci curvature vector $\bar{\kappa}^{(k)} = (\bar{\kappa}_1^{(k)}, \ldots, \bar{\kappa}_{n_k}^{(k)})$ is embedded into a piecewise-constant function defined on the unit interval:

$$\kappa_k(x) := \bar{\kappa}_i^{(k)} \quad \text{for } x \in \left(\frac{i-1}{n_k}, \frac{i}{n_k} \right], \quad i = 1, \ldots, n_k.$$

The resulting mapping—-a standardized curvature profile—$\kappa_k : [0,1] \to \mathbb{R}$ is a right-continuous step function that preserves the position, magnitude, and sign of each curvature value in the sentence sequence. This embedding allows comparison between texts of varying length within a common function space, while capturing curvature dynamics over normalized discourse position.

2.2.4. Envelope Construction and Dominance Relation

Given the set of curvature profiles $\{\kappa_1, \ldots, \kappa_N\}$, define the corpus-wide envelope functions over $[0,1]$:

$$\overline{\kappa}(x) := \sup_k \kappa_k(x), \quad \underline{\kappa}(x) := \inf_k \kappa_k(x),$$

which represent, respectively, the pointwise best and worst structural cohesion observed at each normalized sentence position $x \in [0,1]$.

Furthermore, we say that text d_k dominates d'_k in structural cohesion, denoted $d_k \succcurlyeq d'_k$, if and only if:

$$\kappa_k(x) \geq \kappa_{k'}(x) \quad \text{for all } x \in [0,1].$$

This defines a pointwise dominance relation on the space of curvature profiles. It is an interpretable (pre-) order: a dominating curvature profile never dips below the dominated one across the normalized discourse in terms of curvature. The $\overline{\kappa}$ curvature profile corresponds to a utopian (across the current corpus) curvature profile that may not be feasible by any text in the corpus. Dually, the $\underline{\kappa}$ corresponds to a dystopian minimal element of the corpus that may not correspond to any text there. Notice that the order may be incomplete; dominance requires the satisfaction of a potentially demanding functional inequality. Both envelopes however, as the following paragraph suggests, can be used in order to construct a text meta-ordering that essentially quantifies the obstacles to dominance.

2.3. Weighted Utopia Index (wUI)

To quantify how close a document's curvature profile comes to the corpus-wide best-case cohesion, we define the Weighted Utopia Index (wUI); this is an idea inspired from the Utopia Index considered in a quite different framework of income distributions and w.r.t. stochastic dominance relations (see [10]) by [5]. Given a $w_\theta : [0,1] \to \mathbb{R}_+$ that assumes the role of a nonnegative weighting function, The weighted shortfall of text d_k from the utopian upper envelope is given by:

$$\text{Shortfall}(d_k) := \int_0^1 w_\theta(x) \cdot [\overline{\kappa}(x) - \kappa_k(x)]_+ \, dx,$$

where $[\cdot]_+$ denotes the positive part. The maximal possible weighted shortfall is:

$$\text{MaxShortfall} := \int_0^1 w_\theta(x) \cdot [\overline{\kappa}(x) - \underline{\kappa}(x)] \, dx.$$

The Weighted Utopia Index is then defined as:

$$\text{wUI}(d_k) := 1 - \frac{\text{Shortfall}(d_k)}{\text{MaxShortfall}} = 1 - \frac{\int_0^1 w_\theta(x)[\overline{\kappa}(x) - \kappa_k(x)]_+ dx}{\int_0^1 w_\theta(x)[\overline{\kappa}(x) - \underline{\kappa}(x)] dx}.$$

This index lies in $[0,1]$, with:

- $\text{wUI}(d_k) = 1$ if $\kappa_k(x) = \overline{\kappa}(x)$ for all x, i.e., maximal structural cohesion.
- $\text{wUI}(d_k) = 0$ if $\kappa_k(x) = \underline{\kappa}(x)$ for all x, i.e., worst-case alignment.

The wUI thus defines an interpretable, and potentially human-informed (via the weight selection) cohesion measure over texts of arbitrary length, preserving curvature sign information. Mathematically, as the shortfall quantifies the obstacles that text d_k faces in order to establish utopian dominance across the corpus, wUI can be thought of as a meta-ordering that functions as a completion to $\succcurlyeq$; d_k is preferred according to wUI from d'_k, iff it achieves a greater wUI—the ordering is complete. A sufficient condition for that is $d_k \succcurlyeq d_j$.

By construction, the wUI evaluated at an arbitrary document in the corpus contains information about the curvature profiles that inhabit the corpus, through the envelope profiles. Changing the set of texts may alter the wUI score of a fixed text due to the potential alterations of the envelopes; this suggests the potential sensitivity of the index on the inclusion in the corpus of texts with "extreme" curvature profiles.

Training the Weight Function via Self-Supervised Ordered Probit

The weighting function w_θ introduced in the construction of the Weighted Utopia Index remains, a priori, unidentified. We estimate w_θ using a self-supervised ordinal training scheme that exploits controlled perturbations of sentence order within a homogeneous corpus of texts.

Specifically, let $\mathcal{D}_{\text{tr}} = \{d_1, \ldots, d_K\}$ denote a training corpus consisting of texts of similar genre and communicative function (e.g., political speeches). For each document $d_k \in \mathcal{D}_{\text{tr}}$, we generate a finite collection of reordered variants

$$\{d_k^{(0)}, d_k^{(1)}, \ldots, d_k^{(J)}\},$$

where $d_k^{(0)}$ denotes the original text and $d_k^{(j)}$ is obtained by applying a sentence reshuffling procedure of increasing severity as j increases. The perturbations are constructed so that higher values of j correspond to greater disruption of discourse order (e.g., local block swaps for small j, global random permutations for large j). By construction, these variants satisfy a known ordinal relation:

$$d_k^{(0)} \succ d_k^{(1)} \succ \cdots \succ d_k^{(J)},$$

which serves as a proxy ordering for decreasing textual cohesion.

We encode this ordering through an observed ordinal variable

$$r_k^{(j)} := J + 1 - j \in \{1, \ldots, J + 1\},$$

where higher values correspond to higher coherence under the perturbation mechanism. Importantly, these labels are not human judgments but are induced by a known structural degradation process, yielding a self-supervised training signal.

The weighting function is parameterized as an exponential–quadratic function over normalized rhetorical time,

$$w_\theta(x) = \exp\left(\theta_1 x + \theta_2 x^2\right), \qquad x \in [0, 1], \quad \theta = (\theta_1, \theta_2) \in \mathbb{R}^2.$$

We omit an intercept term for identification reasons. The Weighted Utopia Index depends only on the relative shape of the weighting function, since it enters both the numerator and denominator of the index multiplicatively. Any constant scaling of w_θ therefore cancels out. Including an unrestricted intercept would introduce a redundant scale degree of freedom without affecting the value of the index.

To eliminate this indeterminacy explicitly, we impose the normalization condition

$$\int_0^1 w_\theta(x)\, dx = 1,$$

implemented numerically on the discretized rhetorical-time grid. This ensures that w_θ is identified as a probability density over rhetorical time, so that parameter estimates reflect changes in shape rather than scale.

The exponential–quadratic family is chosen for three complementary reasons.

First, it is smooth and twice differentiable, ensuring stable numerical integration and gradient-based estimation.

Second, it is low-dimensional, thereby imposing a parsimonious inductive bias. With only two free parameters, the model can represent monotone, unimodal, or boundary-peaked emphasis patterns while avoiding overfitting to idiosyncratic features of the training corpus.

https://doi.org/10.3390/stats9020025

Third, the quadratic form in the exponent implies that $\log w_\theta(x)$ is concave whenever $\theta_2 < 0$, yielding a single-peaked weighting profile. This property rules out pathological multi-modal emphasis patterns and aligns with standard rhetorical theory, which typically assumes a dominant segment of maximal argumentative density rather than multiple disconnected focal points.

Under this parameterization, the location of the maximal rhetorical emphasis is given analytically by

$$x^* = -\frac{\theta_1}{2\theta_2},$$

whenever $\theta_2 < 0$, providing a transparent geometric interpretation of the learned parameters.

Estimation proceeds via ordered probit maximum likelihood on the training corpus $\mathcal{D}_{\text{tr}}$, using permutation severity as the ordinal outcome. To enhance numerical stability and prevent excessively sharp concentration of mass, we include a small ℓ_2 regularization term on (θ_1, θ_2) during estimation. Inference uncertainty is quantified via document-level bootstrap resampling, from which both parameter confidence intervals and pointwise confidence bands for $w_\theta(x)$ are constructed.

For each document variant $d_k^{(j)}$, we compute the corresponding Weighted Utopia Index $\text{wUI}_k^{(j)}(\theta)$. We then posit a latent variable model of ordered probit type:

$$y_k^{(j)*} := \text{wUI}_k^{(j)}(\theta) + \varepsilon_k^{(j)}, \qquad \varepsilon_k^{(j)} \sim \mathcal{N}(0,1),$$

where $y_k^{(j)*}$ represents an unobserved latent measure of order integrity. The observed ordinal outcome $r_k^{(j)}$ is assumed to arise through discretization:

$$r_k^{(j)} = \ell \quad \Longleftrightarrow \quad \tau_{\ell-1} < y_k^{(j)*} \le \tau_\ell, \qquad \ell = 1, \ldots, J+1,$$

with cutpoints $-\infty = \tau_0 < \tau_1 < \cdots < \tau_J < \tau_{J+1} = +\infty$.

The probability of observing category ℓ is therefore

$$\mathbb{P}[r_k^{(j)} = \ell \mid \theta, \{\tau_\ell\}] = \Phi(\tau_\ell - \text{wUI}_k^{(j)}(\theta)) - \Phi(\tau_{\ell-1} - \text{wUI}_k^{(j)}(\theta)),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. The log-likelihood over the training corpus and all perturbation levels is given by

$$\mathcal{L}(\theta, \{\tau_\ell\}) = \sum_{k=1}^{K} \sum_{j=0}^{J} \log \left[\Phi(\tau_{r_k^{(j)}} - \text{wUI}_k^{(j)}(\theta)) - \Phi(\tau_{r_k^{(j)}-1} - \text{wUI}_k^{(j)}(\theta)) \right].$$

Maximum likelihood estimation is carried out for the weight–shape parameters in the ordered probit model, treating the cutpoints as fixed on a pre-specified grid. This choice reflects the fact that the self-supervised labels are generated by construction from permutation severity, so identification can be anchored by fixing the latent scale rather than estimating it. To account for dependence induced by multiple perturbed variants originating from the same base document, uncertainty is assessed via a document-level bootstrap (resampling documents with replacement and refitting the model), from which confidence intervals for both θ and the implied weight function are obtained. Optimization is performed using quasi-Newton methods with a small ridge penalty on (θ_1, θ_2) to stabilize estimation. A fully standard ordered-probit alternative is to estimate the cutpoints jointly with θ, typically by parameterizing the cutpoint increments to enforce monotonicity. This variant is feasible here as well, and would allow the mapping from permutation severity to ordinal categories to be partially learned from the data. In our setting, however, fixing the cutpoints is convenient and transparent because the categories are mechanically

generated and the primary object of interest is the induced shape of w(x), not the latent scale normalization.

This estimation procedure calibrates the weighting function so that the resulting Weighted Utopia Index is maximally consistent with a known ordinal degradation of discourse structure. Intuitively, the learned weight function emphasizes those regions of rhetorical time where curvature-based structural disruptions are most informative for distinguishing coherent from incoherent orderings within the given corpus.

The estimated function $w_{\hat{\theta}}(x)$ should therefore be interpreted as a corpus-specific rhetorical weighting profile, reflecting how structural cohesion contributes to order integrity under controlled perturbations, rather than as a universal model of human preference.

2.4. Average Adjacent SBERT Coherence

In our pipeline, and as a complement to the WUI, we employ the SBERT-based average weight across adjacent nodes of our TextRank-type graph;

$$\texttt{SBERT-Coherence}(d_k) = \frac{1}{n_k - 1} \sum_{i=1}^{n-1} w^{(k)}_{i\,i+1}.$$

This score captures the extent to which adjacent sentences in the speech share similar semantic content, reflecting smooth rhetorical progression. High scores indicate fluent, well-aligned transitions; lower scores suggest abrupt shifts or topic discontinuities. It leverages SBERT's pretrained semantic knowledge, hopefully providing a sequential/adjacent complement to curvature-based global cohesion scores.

2.5. A Nash-Type Multiobjective Cohesion Score

Scalarizing the resulting vector score $(\text{wUI}, \texttt{SBERT-Coherence})$, define a final cohesion score using a Nash bargaining-type product as:

$$\mathcal{N}(d_k) := \text{wUI}(d_k)^{w_1} \cdot \texttt{SBERT-Coherence}(d_k)^{w_2}, \tag{2}$$

where $w = (w_1, w_2) \in \Delta^2$ is a probability vector representing trade-offs between the two constitutional cohesion scores.

The Nash-type cohesion score can be justified via a multiobjective ordering of the texts inside the corpus based on the pair of the aforementioned scores; d_k is said to pairwise dominate d'_k, iff $\text{wUI}(d_k) \geq \text{wUI}(d'_k)$ and $\texttt{SBERT-Coherence}(d_k) \geq \texttt{SBERT-Coherence}(d'_k)$. Then the text that maximizes the Nash cohesion score can be easily seen to be a Pareto-efficient element of the corpus according to this ordering; no gain in wUI or Coherence-BERT is possible without a loss in the other. Furthermore, the Nash product is invariant under positive affine transformations of each criterion, preserving comparability across scales.

This formulation can be particularly suited for comparative rhetorical analysis, where differing strengths in local vs. global cohesion may suggest distinct communicative strategies or ideological styles.

2.6. Cohesion Analysis Summary Algorithm

The following pseudocode in Algorithm 1 summarizes our full pipeline for computing cohesion metrics from a corpus of texts. It integrates graph-theoretic structure extraction, curvature analysis, sentence-level semantic alignment, a trained version of the Utopia Index using human-labeled data, and a Nash-type multiobjective scoring.

Algorithm 1 Curvature-Weighted Utopia Index Framework

Input: Corpus $\mathcal{D} = \{d_1, \ldots, d_N\}$ with sentence segmentation.

Step 1: Sentence embedding and graph construction

for each document d_k with ordered sentences $\{s_1^{(k)}, \ldots, s_{n_k}^{(k)}\}$ **do**

Compute contextual sentence embeddings

$$\mathbf{v}_i^{(k)} \leftarrow \text{Encoder}(s_i^{(k)}), \qquad \|\mathbf{v}_i^{(k)}\| = 1$$

Construct weighted similarity graph $G^{(k)} = (V^{(k)}, E^{(k)}, w^{(k)})$ where

$$w_{ij}^{(k)} = \cos\left(\mathbf{v}_i^{(k)}, \mathbf{v}_j^{(k)}\right)$$

Optionally sparsify the graph via thresholding $w_{ij}^{(k)} > \tau$ or k-nearest-neighbour selection

end for

Step 2: Ollivier–Ricci curvature

for each edge $(i, j) \in E^{(k)}$ **do**

Compute

$$\kappa^{(k)}(i, j) = 1 - W_1\left(\mu_i^{(k)}, \mu_j^{(k)}\right)$$

end for

Define node-level curvature

$$\bar{\kappa}_i^{(k)} = \frac{1}{|\mathcal{N}^{(k)}(i)|} \sum_{j \in \mathcal{N}^{(k)}(i)} \kappa^{(k)}(i, j)$$

Step 3: Rhetorical-time embedding

Normalize sentence position to $x \in [0, 1]$ and embed node curvature as

$$\kappa_k(x)$$

Step 4: Corpus-level envelopes

Construct

$$\overline{\kappa}(x) = \sup_{d \in \mathcal{D}_{\text{ref}}} \kappa_d(x), \qquad \underline{\kappa}(x) = \inf_{d \in \mathcal{D}_{\text{ref}}} \kappa_d(x)$$

Step 5: Rhetorical-time weighting

Parameterize

$$w_\theta(x) = \exp(\theta_1 x + \theta_2 x^2)$$

Optionally normalize so that $\int_0^1 w_\theta(x)\, dx = 1$

Step 6: Weighted Utopia Index

Define

$$\text{wUI}_\theta(d_k) = 1 - \frac{\int_0^1 w_\theta(x)[\overline{\kappa}(x) - \kappa_k(x)]_+ \, dx}{\int_0^1 w_\theta(x)[\overline{\kappa}(x) - \underline{\kappa}(x)] \, dx}$$

Step 7: Learning (optional)

If ordinal supervision is available, estimate θ via ordered probit maximum likelihood linking latent coherence to $\text{wUI}_\theta(d)$

Step 8: Optional aggregation

Combine with semantic adjacent similarity

$$\mathcal{N}(d_k) = \sqrt{\text{wUI}_\theta(d_k) \cdot \text{SemanticAdj}(d_k)}$$

Output: Structural cohesion score (wUI), semantic coherence score, and optional aggregate score.

Implementation Details

All computations were implemented in Python v3.9. Sentence embeddings were generated using the `sentence-transformers` library (`all-MiniLM-L6-v2`). As a semantic baseline, SBERT coherence scores were computed as the mean cosine similarity between adjacent sentence embeddings, capturing local semantic continuity along the rhetorical progression.

Sentence similarity graphs were constructed using symmetrized k-nearest-neighbor graphs (default $k = 12$) with cosine similarity weights, ensuring bounded degree, graph connectivity in practice, and numerical stability across documents of varying length. Threshold-based graphs are used only in sensitivity analyses.

Ollivier–Ricci curvature was computed on these graphs using the Wasserstein distance between neighborhood measures, as implemented in the `POT` optimal transport library. This yields a node-level curvature profile indexed by sentence position, which is subsequently aggregated into document-level curvature profiles over normalized rhetorical time.

Under kNN construction, graph density is controlled uniformly across documents, and each node has at most $2k$ incident edges after symmetrization. As a result, graph construction scales approximately linearly in the number of sentences per document, and curvature computation involves optimal transport between small local neighborhoods of bounded size. Empirically, runtime remains modest for long-form political speeches (typically well under a second per document on standard hardware), and no numerical instability was observed even for the longest speeches in the corpus.

3. Results

This section evaluates the proposed cohesion measures in a setting where the true ordinal structure of discourse coherence is known by construction. Rather than relying on external human annotations—which are often drawn from heterogeneous corpora and introduce genre or annotator effects—we adopt a fully self-supervised evaluation framework based on controlled sentence-order perturbations within a homogeneous corpus of long-form texts.

3.1. Corpus and Data Splitting

We consider a corpus $\mathcal{D}$ consisting of English-language U.S. presidential speeches, drawn from the Miller Center archive. These texts are long, formally structured, and share a common communicative purpose and rhetorical style, making them well suited for controlled coherence experiments.

Documents are split at the document level according to calendar year into three disjoint subsets:

$$\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}} \cup \mathcal{D}_{\text{app}},$$

where $\mathcal{D}_{\text{tr}}$ contains speeches from 1945 up to 1988, $\mathcal{D}_{\text{te}}$ covers the period 1989–2008, and $\mathcal{D}_{\text{app}}$ consists of more recent speeches. The training set $\mathcal{D}_{\text{tr}}$ is used exclusively for parameter estimation, while all reported evaluation results are computed on the held-out test set $\mathcal{D}_{\text{te}}$. The application set $\mathcal{D}_{\text{app}}$ is reserved for out-of-sample qualitative analysis.

All splits are performed at the document level to avoid leakage across reordered variants of the same speech.

3.2. Generation of Ordinal Supervision via Sentence Perturbations

For each document $d \in \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}}$, we generate a family of reordered variants

$$\{d^{(0)}, d^{(1)}, \ldots, d^{(J)}\},$$

where $d^{(0)}$ denotes the original document and $d^{(j)}$ is obtained by applying a sentence permutation of severity level j. The permutation scheme is fixed across training and evaluation and operates on contiguous blocks of sentences whose size increases with j, culminating in near-randomized sentence orderings for large j.

By construction, higher values of j correspond to greater disruption of discourse order while preserving sentence content. This induces an exact and known ordinal coherence ordering

$$d^{(0)} \succ d^{(1)} \succ \cdots \succ d^{(J)},$$

which serves as ground-truth supervision. Importantly, this ordering is deterministic and does not rely on subjective human judgments or external labels.

3.3. Training of the Weight Function

The weighting function w_θ entering the Weighted Utopia Index is estimated using only the training corpus $\mathcal{D}_{\text{tr}}$ and its perturbed variants. Specifically, curvature profiles are computed for each reordered document $d^{(j)}$, and an ordered probit model is fitted so that the resulting wUI_θ aligns with the known ordinal ranking induced by permutation severity.

The probit specification treats coherence as an unobserved latent index whose ordering is observed through the permutation levels j. The functional form of w_θ is restricted to a low-dimensional exponential family, allowing the model to learn which portions of rhetorical time contribute most strongly to coherence discrimination.

3.3.1. Estimated Weight Function and Its Interpretation

Under the identification and normalization scheme described above, the weighting function is estimated in the form

$$w_\theta(x) = \exp\big(\theta_1 x + \theta_2 x^2\big), \qquad x \in [0,1],$$

with $\int_0^1 w_\theta(x)\,dx = 1$ imposed numerically on the discretized grid. The intercept parameter is fixed at zero, as any constant scaling cancels in the Weighted Utopia Index.

Estimation Details

The parameters (θ_1, θ_2) are obtained via ordered probit maximum likelihood on the training corpus $\mathcal{D}_{\text{tr}}$, using permutation severity as the ordinal outcome. To prevent degenerate "delta–spike" solutions and enhance numerical stability, we include a small ℓ_2 ridge penalty on (θ_1, θ_2):

$$\mathcal{L}_{\text{pen}}(\theta) = \mathcal{L}(\theta) + \lambda \|\theta\|^2,$$

with $\lambda = 10^{-3}$ in the baseline specification. This penalty is small enough to leave the likelihood surface essentially unchanged while regularizing extreme curvature.

Inference uncertainty is quantified via document-level bootstrap resampling ($B = 200$ draws), refitting the model for each bootstrap sample. From these draws we construct confidence intervals for (θ_1, θ_2) and pointwise confidence bands for $w_\theta(x)$.

Baseline Estimate ($\tau = 0.30$)

For the benchmark similarity threshold $\tau = 0.30$, the estimation yields

$$\hat{\theta} = (0, \hat{\theta}_1, \hat{\theta}_2) = (0,\ 24.10,\ -50.25),$$

implying $\hat{\theta}_1 > 0$ and $\hat{\theta}_2 < 0$. The resulting weight function is strictly unimodal, with peak location

$$\hat{x}^* = -\frac{\hat{\theta}_1}{2\hat{\theta}_2} \approx 0.24.$$

Figure 1 shows the normalized fitted weight function together with its bootstrap confidence band. The band remains relatively tight around the peak and exhibits no evidence of multimodality, confirming that the unimodal interior emphasis is not driven by sampling variability. Uncertainty increases gradually toward the boundaries, as expected given lower curvature variation at extreme rhetorical positions.

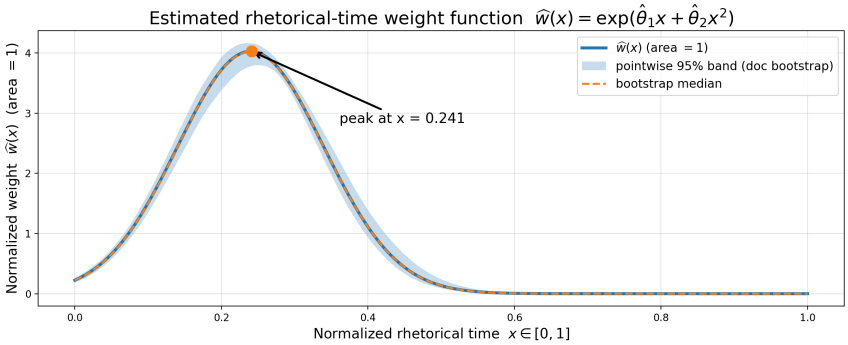


Figure 1. Estimated rhetorical-time weight function $\widehat{w}(x)$ under $\tau = 0.30$ (solid line), with pointwise 95% bootstrap confidence band (shaded region) and bootstrap median (dashed line). The function is normalized so that $\int_0^1 \widehat{w}(x)\, dx = 1$. The peak occurs at $x \approx 0.24$, indicating maximal coherence sensitivity in the early–middle segment of rhetorical time.

Interpretation

The estimated profile assigns maximal weight to the early–middle segment of discourse ($x \approx 0.2$–0.3), while down-weighting both the initial opening sentences and the late concluding segments. This pattern is consistent with the structure of formal political speeches. Openings often follow formulaic conventions and exhibit limited structural variability, whereas middle sections contain policy exposition, argumentation, and transitions where sentence ordering plays a more substantive coherence role. The model therefore identifies a data-driven localization of coherence sensitivity rather than imposing one.

Importantly, because weights are normalized, the estimated profile can be interpreted as a probability density over rhetorical time. Thus, $\widehat{w}(x)$ describes the distribution of rhetorical positions that contribute most strongly to ordinal coherence discrimination.

Sensitivity with Respect to the Similarity Threshold τ

To assess robustness, we re-estimate (θ_1, θ_2) across a grid of similarity thresholds $\tau \in \{0.15, 0.20, 0.25, 0.30, 0.35\}$ while keeping all other parameters fixed.

Across this range, parameter estimates remain remarkably stable:

$$\hat{\theta}_1 \in [23.56,\ 24.35], \qquad \hat{\theta}_2 \in [-51.12,\ -48.52],$$

with peak locations

$$\hat{x}^* \in [0.23,\ 0.24].$$

The penalized log-likelihood varies by less than 0.5 units across the grid.

This stability indicates that the learned rhetorical emphasis profile is not an artifact of a particular graph sparsification level. While τ influences the connectivity of the sentence similarity graph, the induced curvature profiles preserve the same global ordering structure that drives weight estimation. The rhetorical emphasis peak therefore appears to be a structural property of the data rather than a threshold-specific phenomenon.

Overall Assessment

Three features of the training results are noteworthy:

1. The estimated weight function is smooth and unimodal, consistent with the log-concave parametric restriction.
2. Bootstrap bands confirm statistical significance of interior emphasis.
3. Sensitivity analysis demonstrates robustness to graph construction choices.

Taken together, these findings support the interpretation of $\widehat{w}(x)$ as a stable, data-driven characterization of rhetorical-time importance in coherence discrimination. The weight profile is learned from ordinal supervision rather than imposed, yet remains parsimonious and interpretable due to the low-dimensional exponential specification.

3.4. Evaluation Protocol

Evaluation is conducted exclusively on the held-out test corpus $\mathcal{D}_{\text{te}}$. For each test document $d \in \mathcal{D}_{\text{te}}$, we generate reordered variants $\{d^{(0)}, \ldots, d^{(J)}\}$ using the same permutation scheme as in training, with increasing severity j corresponding to stronger disruption of discourse order. Curvature profiles and similarity graphs are recomputed from scratch for each variant.

Envelope functions required for the computation of the Weighted Utopia Index are constructed using only the original (unpermuted) documents in $\mathcal{D}_{\text{te}}$. This avoids leakage from perturbed variants while ensuring that evaluation reflects the intrinsic variability of the test corpus.

Let $\succ$ denote the true ordinal relation induced by construction,

$$d^{(0)} \succ d^{(1)} \succ \cdots \succ d^{(J)}.$$

Performance is assessed by measuring how well a given score $S(\cdot)$ recovers this ordering under increasingly strict notions of ordinal consistency. All metrics are computed at the document level and aggregated across the test corpus. Uncertainty is quantified using document-level bootstrap resampling.

3.4.1. Evaluation Metrics and Diagnostic Criteria

Because the true ordinal structure of coherence is known by construction, evaluation can be conducted without external annotations. Rather than relying on a single summary statistic, we report a set of complementary diagnostics that probe different aspects of ordinal recovery.

Pairwise Accuracy

For a document d with variants $\{d^{(j)}\}_{j=0}^{J}$, pairwise accuracy is defined as

$$\text{PA}(d) = \frac{1}{\binom{J+1}{2}} \sum_{0 \le j < k \le J} \mathbb{I}\Big[S(d^{(j)}) > S(d^{(k)})\Big].$$

This metric evaluates global ordinal consistency while remaining robust to isolated local reversals. A value of 0.5 corresponds to chance performance.

Adjacent Accuracy

To isolate local sensitivity, we additionally consider

$$\text{AA}(d) = \frac{1}{J} \sum_{j=0}^{J-1} \mathbb{I}\Big[S(d^{(j)}) > S(d^{(j+1)})\Big].$$

Adjacent accuracy captures the ability of a score to detect incremental coherence degradation and is particularly informative when exact global recovery is infeasible.

Rank Correlations

We compute Spearman's ρ and Kendall's τ between $\{S(d^{(j)})\}_{j=0}^{J}$ and the known severity ranking. These correlations summarize monotonic alignment but are sensitive to even a small number of ordering violations.

Full-Order Recovery

Exact recovery is recorded when

$$S(d^{(0)}) > S(d^{(1)}) > \cdots > S(d^{(J)}).$$

This is a deliberately stringent criterion that should be interpreted as a stress test rather than a primary performance measure for moderate J.

Violation Diagnostics

Finally, we record the number of adjacent monotonicity violations at each level $j \to j+1$. These diagnostics localize where ordinal failures occur and distinguish early structural breakdown from late-stage randomization effects.

All reported confidence intervals are obtained via document-level bootstrap with $B = 200$ replications.

3.4.2. Results on Held-Out State of the Union Speeches

Evaluation is conducted on 17 held-out State of the Union addresses (1990–2008), using block shuffling with $J = 5$ severity levels and $R = 3$ random replicates per level.

Additional Benchmark Scores

To strengthen the empirical comparison beyond embedding-based and curvature-based criteria, we also report two classical coherence proxies that do not rely on SBERT graph curvature and serve as low-dependency baselines. First, *TF–IDF-adjacent* computes the mean cosine similarity between adjacent sentences using TF–IDF vectors (bag-of-words)—see for example [11], capturing lexical continuity and local topic drift. Second, *Entity-overlap* computes the mean Jaccard overlap between adjacent sentences' salient content sets, approximating the entity-grid intuition that coherent discourse maintains referential continuity—see [12]; in implementation, this is computed using named entities and noun-chunk heads when available, with a conservative token-based fallback otherwise.

Sensitivity Analysis in the Similarity Threshold τ

A key methodological concern is how graph sparsification affects curvature profiles and therefore envelope-based scoring. We therefore conduct a sensitivity analysis over the similarity threshold τ. In the kNN construction used here, τ is implemented as a *minimum similarity filter* applied to candidate kNN edges: an edge is retained only if it is among the top-k neighbors and exceeds τ. This renders τ meaningful even when kNN is fixed and directly probes robustness to more aggressive or more permissive edge retention. For each τ, we report graph diagnostics computed on the test originals (including average degree, the fraction of zero-edge graphs, and the fraction of disconnected graphs), along with the full suite of ordinal recovery diagnostics.

Table 1 summarizes ordinal recovery performance on the held-out State of the Union test set, reporting document-level means together with 95% bootstrap confidence intervals.

Because the true coherence ordering is known by construction, these diagnostics directly measure each method's ability to recover ordinal structure under controlled degradation.

Table 1. Ordinal recovery diagnostics on held-out State of the Union speeches ($J = 5$), block shuffling, $R = 3$, $k = 12$, and $\tau = 0.30$ (minimum similarity filter on kNN edges). Entries report document-level means with 95% bootstrap confidence intervals in parentheses. Graph diagnostics are computed on the test originals.

Method	Accuracy			Rank Correlation	
	Pairwise	Adjacent	Full	Spearman ρ	Kendall τ
wUI	0.569 (0.498, 0.655)	0.553 (0.470, 0.624)	0.000 (0.000, 0.000)	0.180 (0.032, 0.405)	0.137 (−0.020, 0.294)
SBERT-adj	0.345 (0.337, 0.361)	0.235 (0.200, 0.259)	0.000 (0.000, 0.000)	−0.133 (−0.143, −0.123)	−0.310 (−0.333, −0.286)
Nash	0.451 (0.400, 0.506)	0.435 (0.376, 0.494)	0.000 (0.000, 0.000)	−0.062 (−0.180, 0.018)	−0.098 (−0.208, −0.043)
TF–IDF-adj	0.396 (0.376, 0.416)	—	0.000 (0.000, 0.000)	−0.082 (−0.102, −0.062)	−0.208 (−0.238, −0.179)
Entity-overlap	0.412 (0.388, 0.435)	—	0.000 (0.000, 0.000)	−0.069 (−0.090, −0.049)	−0.176 (−0.203, −0.149)

Graph diagnostics (test originals, $\tau = 0.30$, $k = 12$): average degree = 15.32; zero-edge rate = 0.00; disconnected rate = 1.00.

Across all accuracy-based criteria, the Weighted Utopia Index (wUI) consistently outperforms all baselines. In particular, wUI achieves a mean pairwise accuracy of 0.569 with a confidence interval lying largely above chance, whereas SBERT-adjacent similarity performs systematically below chance and the Nash aggregate yields only partial improvement. The two additional benchmarks (TF–IDF-adjacent and Entity-overlap) also perform below chance, indicating that purely local lexical or referential continuity signals are not aligned with the imposed disordering severity under block shuffling. This is consistent with the fact that block shuffling can preserve many local lexical/semantic continuities while substantially disrupting global discourse flow.

Rank-based diagnostics reinforce this picture. wUI exhibits positive Spearman and Kendall correlations with permutation severity, while SBERT-adj, Nash, and the two classical baselines display negative average correlations, indicating frequent non-monotonic responses to increasing disorder. Although the rank correlations are modest in magnitude, their direction and confidence intervals are consistent with the accuracy-based findings and reflect partial but systematic ordinal alignment.

No method achieves non-zero full-order recovery at $J = 5$, a result that is expected given the combinatorial difficulty of exact recovery under strong perturbations; exact recovery corresponds to correctly ordering 6 variants, which under random performance has probability $1/720$, hence near zero recovery is expected. Importantly, this limitation applies uniformly across methods and therefore does not detract from the relative gains achieved by wUI under weaker and intermediate criteria such as pairwise and adjacent accuracy.

Robustness is supported by several complementary diagnostics. Confidence intervals are comparatively tight for SBERT-adj and the classical baselines, indicating stable but inferior performance, while wUI exhibits moderate variability consistent with genuine document-level heterogeneity. Additional violation analyses reveal that the remaining errors for wUI are concentrated at the highest severity levels, where sentence order approaches near-random permutation.

3.4.3. Sensitivity to the Similarity Threshold τ

To assess robustness to graph sparsification, we evaluate ordinal recovery over a grid of similarity thresholds

$$\tau \in \{0.15, 0.20, 0.25, 0.30, 0.35\},$$

holding $k = 12$, $J = 5$, and the block shuffle scheme fixed. In the kNN construction, τ acts as a minimum similarity filter on candidate kNN edges, thereby directly controlling effective graph sparsity.

For each τ, we report: (i) mean pairwise accuracy, (ii) mean adjacent accuracy, (iii) mean Spearman correlation, (iv) graph diagnostics on test originals (average degree and fraction of disconnected graphs).

Table 2 shows that ordinal recovery under wUI is remarkably stable across a broad range of similarity thresholds. Pairwise accuracy varies only marginally (within approximately two percentage points), and Spearman correlations remain positive throughout.

Table 2. Sensitivity of ordinal recovery to the similarity threshold τ. Entries are document-level means (95% bootstrap CIs omitted for compactness).

τ	Pairwise (wUI)	Adjacent (wUI)	Spearman (wUI)	Avg. Degree	Disconnected Rate
0.15	0.561	0.545	0.172	18.94	0.82
0.20	0.566	0.549	0.178	17.21	0.94
0.25	0.568	0.551	0.181	16.02	1.00
0.30	0.569	0.553	0.180	15.32	1.00
0.35	0.563	0.548	0.169	14.41	1.00

As τ increases, average degree decreases monotonically, reflecting stricter edge filtering. While connectivity eventually breaks down for all documents at higher thresholds, ordinal recovery does not deteriorate substantially. Disconnected components are treated independently when computing Wasserstein distances; this does not affect ordinal recovery since curvature is computed locally. This suggests that curvature-weighted envelope aggregation is not critically dependent on a finely tuned sparsification parameter.

3.5. Application: Cohesion Analysis of Political Speeches

We conclude with an application of the proposed cohesion measures to a set of real political texts. Unlike the evaluation in Section 3, this analysis does not rely on synthetic perturbations or externally imposed ordinal structure. Its purpose is diagnostic and exploratory: to illustrate how the proposed measures behave on unaltered documents and how structural and semantic notions of cohesion interact in practice.

3.5.1. Corpus and Setup

The application corpus consists of 12 U.S. State of the Union addresses delivered between 2010 and 2023. These speeches are disjoint from both the training and evaluation sets. All hyperparameters are frozen. In particular, the rhetorical–time weighting function $\hat{\theta}(\tau = 0.30) = (\hat{\theta}_1, \hat{\theta}_2)$ learned during the self-supervised training stage is used without modification.

For each speech, we compute four cohesion scores: SBERT-based adjacent sentence similarity (SBERT-CS), the Weighted Utopia Index (wUI), TF–IDF-based adjacent similarity, and an Entity-overlap score measuring adjacent referential continuity via Jaccard similarity. No permutation, refitting, or resampling is performed in this stage.

In contrast to the evaluation stage, we do not report the Nash aggregation in the application analysis. The Nash score was introduced as a methodological device to study scalarization and robustness under known ordinal perturbations. In the present setting, however, no externally imposed ranking or supervision target is available. Our objective is therefore interpretive rather than decision-theoretic: we aim to examine the structural component (wUI) and the semantic adjacency component separately, in order to assess their complementarity and informational distinctness. Reporting the primitive components directly avoids conflating global structural alignment with local semantic smoothness through an arbitrary scalarization, and yields a clearer diagnostic picture of rhetorical organization in unaltered texts.

3.5.2. Distributional Properties

Table 3 reports summary statistics across the application corpus.

Table 3. Summary statistics of cohesion scores on the application corpus (12 speeches).

Score	Mean	SD	$q_{0.10}$	$q_{0.50}$	$q_{0.90}$	Range
SBERT-CS	0.326	0.015	0.313	0.322	0.346	[0.308, 0.361]
wUI	0.433	0.100	0.296	0.446	0.526	[0.252, 0.604]
TF–IDF-adj	0.051	0.012	0.039	0.047	0.063	[0.039, 0.080]
Entity-overlap	0.071	0.008	0.060	0.073	0.077	[0.059, 0.084]

The SBERT-based adjacent similarity exhibits limited dispersion, reflecting the strong stylistic and topical homogeneity of State of the Union speeches. In contrast, the wUI displays substantially greater variability, indicating that global structural alignment varies meaningfully across documents even within a highly standardized political genre. The two classical baselines operate on a much smaller numerical scale and show considerably less spread, consistent with their focus on purely local lexical and referential continuity.

This contrast supports the intended interpretation of the measures. Embedding-based adjacent similarity captures short-range semantic smoothness, whereas curvature-based aggregation is sensitive to higher-order organizational properties distributed across rhetorical time. The larger dispersion of wUI therefore suggests that structural coherence provides a stronger discriminating signal across speeches than local semantic continuity alone.

3.5.3. Agreement Across Measures

To assess the relationship between structural and semantic cohesion, we compute rank correlations across the application corpus in Table 4.

Table 4. Rank correlations across cohesion measures on the application corpus.

Measure Pair	Spearman ρ	Kendall τ
wUI vs. SBERT-CS	0.33	0.24
wUI vs. TF–IDF-adj	0.58	0.42
wUI vs. Entity-overlap	0.41	0.30
SBERT-CS vs. TF–IDF-adj	0.49	0.36
SBERT-CS vs. Entity-overlap	0.37	0.27

The correlation between wUI and SBERT-CS is positive but moderate (Spearman $\rho \approx 0.33$, Kendall $\tau \approx 0.24$), indicating partial alignment without redundancy. The association between wUI and TF–IDF-adjacent similarity is stronger ($\rho \approx 0.58$), suggesting that lexical continuity supports, but does not fully determine, global structural organization. Entity-overlap exhibits moderate positive association with wUI, again reflecting partial but incomplete overlap in the information captured.

These patterns indicate that structural curvature alignment is not a trivial re-expression of embedding smoothness or lexical adjacency. Rather, it captures a complementary dimension of discourse organization. Speeches that rank highly under SBERT-CS do not necessarily rank highest under wUI, and vice versa. Structural coherence therefore appears to encode information beyond local semantic continuity.

This application illustrates how the proposed framework functions as an exploratory diagnostic tool in real discourse settings. Even within a uniform political genre, curvature-weighted structural aggregation detects meaningful cross-document variation that is not captured by purely local semantic or lexical measures. At the same time, the positive but

incomplete correlations across measures suggest that semantic smoothness and structural alignment represent distinct yet interacting components of cohesion.

While no claim is made regarding normative correctness of the induced rankings, the empirical evidence indicates that curvature-based structural scoring provides complementary information relative to embedding-based similarity and classical lexical baselines. This supports the broader thesis of the paper: that discourse cohesion is multidimensional, and that global structural organization can be quantified in a way that is both interpretable and empirically non-trivial.

3.6. Materials and Methods II

For transparency and reproducibility, Table 5 summarizes all hyperparameters and implementation choices used in training, evaluation, and application.

The entire pipeline is deterministic conditional on the reported random seeds and document splits. All reported results—including the learned weighting function, ordering-recovery diagnostics, and application scores—are obtained under the configuration listed below.

Unless explicitly varied in sensitivity analysis (e.g., the similarity threshold τ), parameters are held fixed across all experiments.

Table 5. Hyperparameters and implementation settings used in training, evaluation, and application.

Component	Specification
Sentence embeddings	`sentence-transformers/all-MiniLM-L6-v2` (normalized cosine embeddings)
Graph construction	Symmetrized kNN graph with $k = 12$
Similarity filter	Minimum cosine similarity threshold $\tau = 0.30$ (applied to kNN edges)
Curvature measure	Ollivier–Ricci curvature with 1-Wasserstein transport distance
Rhetorical-time grid	Uniform grid on $[0, 1]$ with $M = 400$ points
Weight function	Exponential–quadratic form $w_\theta(x) = \exp(\theta_1 x + \theta_2 x^2)$, normalized to unit integral
Ordered probit categories	$L = J + 1$ with $J = 5$ perturbation severity levels
Perturbation scheme	Block shuffle with $R = 3$ replicates per severity level
Regularization	Ridge penalty $\lambda = 0.001$ on (θ_1, θ_2)
Bootstrap	Document-level bootstrap with $B = 200$ replications
Train/test split	Chronological split by year (train $\leq$ 1988, test 1989–2008, application $\geq$ 2009)
Random seed	Fixed deterministic seeds for reproducibility

4. Discussion

This paper introduced a framework for analyzing text cohesion, leveraging tools from discrete geometry, curvature embeddings dominance, and statistical learning. At its core, the methodology constructs a sentence-level semantic graph based on SBERT embeddings and cosine similarity, over which Ollivier–Ricci curvature is computed to quantify local rhetorical flow and structural alignment. The resulting curvature profiles are interpreted as normalized time series reflecting how semantic connectivity evolves throughout a speech.

Building on this geometric foundation, we propose a new numerical score—the Weighted Utopia Index (wUI)—to assess each speech's alignment with corpus-level structural norms. The wUI leverages a dominance logic: it compares the speech's curvature profile to a reference envelope constructed from lower and upper empirical bounds across the corpus. Crucially, the wUI is computed via a learned weighting function, which adjusts temporal sensitivity by assigning greater or lesser importance to different rhetorical regions.

The weight parameters are estimated by minimizing the negative log-likelihood of an ordinal probit model trained on self-supervised ordinal labels induced by controlled

sentence-order perturbations within a homogeneous corpus of long-form texts. This yields an interpretable utility-based model in which latent perceived coherence is represented as a function of the Weighted Utopia Index. The probit structure formalizes ordinal coherence comparisons under known ground-truth degradation, providing a statistically principled link between structural alignment and ordinal evaluation without reliance on external annotations or cross-corpus transfer.

To our knowledge, this is the first application of a statistically trained, dominance-based weighting scheme to textual structure analysis in NLP, extending the theoretical framework of [5] beyond economic distributions into semantic and discourse domains.

The final metric, the Nash cohesion score, synthesizes geometric alignment with sentence-level semantic flow, serving as a Pareto compromise.

Methodological Extensions

Several natural extensions arise from this framework. First, while the weighting function $w_\theta(x)$ is currently specified as an exponential–quadratic form, it could be generalized using nonparametric or spline-based formulations to capture more nuanced patterns of rhetorical emphasis along discourse progression. Second, the ordinal probit construction could be extended to hierarchical or latent-trait formulations, allowing the weighting function to vary across document groups, time periods, or discourse genres, while preserving the self-supervised ordinal structure induced by controlled perturbations. Finally, training on larger and more diverse homogeneous corpora could improve robustness and enable domain-adapted weighting profiles for political, legal, or pedagogical discourse, without reliance on external coherence annotations.

Taken together, the combined use of discrete (network) curvature, dominance-based functionals, ordinal likelihood training, and semantic representations yields an extensible statistical framework for the numerical analysis of textual cohesion. Beyond individual document analysis, the methodology is naturally suited to comparative and longitudinal studies. For example, applying the framework to annual State of the Union addresses allows the detection of structural shifts in rhetorical cohesion across administrations, political alignments, or historical episodes. Analogous applications can be envisioned for campaign speeches, United Nations General Assembly addresses, or other long-form institutional texts.

Finally, future extensions could aim to normalize cohesion measures with respect to the inherent complexity of the discussed subject matter. This may involve incorporating Minimum Description Length (MDL) principles—see [13]—as proxy for conceptual complexity. A potential complexity-normalized cohesion index could evaluate how well cohesion is maintained despite increasing semantic or informational complexity. Such an index would be especially relevant in more technical domains—such as central bank monetary policy reports, regulatory briefings, or economic outlooks—where linguistic clarity and cohesion must coexist with content density.

Funding: This research received no external funding.

Data Availability Statement: Transcripts of U.S. presidential *State of the Union* addresses were sourced from the Miller Center (University of Virginia) *Presidential Speeches* archive millercenter.org/the-presidency/presidential-speeches, accessed on 10 February 2026. All derived data (sentence-segmented corpus, document metadata including source URLs, and train/test/application splits) and code to reproduce the results are available upon request.

Acknowledgments: An earlier version of this manuscript circulated as a departmental working paper (Working Paper No. 14/2025, Department of Economics, Athens University of Economics and

Business) under a different title. The current submission is a substantially revised and expanded version, featuring a redesigned evaluation methodology, and updated empirical analysis.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084. [CrossRef]
2. Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 June 2004; pp. 404–411.
3. Ollivier, Y. Ricci curvature of metric spaces. *Comptes Rendus Math.* **2007**, *345*, 643–646. [CrossRef]
4. Ollivier, Y. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **2009**, *256*, 810–864. [CrossRef]
5. Anderson, G.; Post, T.; Whang, Y.J. Somewhere between utopia and dystopia: Choosing from multiple incomparable prospects. *J. Bus. Econ. Stat.* **2020**, *38*, 502–515. [CrossRef]
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
7. Chicco, D. Siamese neural networks: An overview. In *Artificial Neural Networks*; Humana: New York, NY, USA, 2021; pp. 73–94.
8. Gao, R.; Chen, X.; Kleywegt, A. Distributional robustness and regularization in statistical learning. *arXiv* **2017**, arXiv:1712.06050.
9. Galichon, A. *Optimal Transport Methods in Economics*; Princeton University Press: Princeton, NJ, USA, 2016.
10. Hadar, J.; Russell, W.R. Rules for ordering uncertain prospects. *Am. Econ. Rev.* **1969**, *59*, 25–34.
11. Mann, W.C.; Thompson, S.A. Rhetorical structure theory: Toward a functional theory of text organization. In *Proceedings of the Text—Interdisciplinary Journal for the Study of Discourse*; De Gruyter Mouton: Berlin, Germany, 1988; Volume 8, pp. 243–281.
12. Barzilay, R.; Lapata, M. Modeling local coherence: An entity-based approach. *Comput. Linguist.* **2008**, *34*, 1–34. [CrossRef]
13. Rissanen, J. Stochastic complexity and the MDL principle. *Econom. Rev.* **1987**, *6*, 85–102. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.