

Article

An Information-Geometric Justification for Composite Coherence in Event-Based Narrative Extraction

Brian Keith-Norambuena 

Department of Computing & Systems Engineering, Universidad Católica del Norte, Antofagasta 1270398, Chile; brian.keith@ucn.cl

Abstract

Graph-based narrative extraction relies on a coherence function to score transitions between events, but the coherence metrics in current use are defined operationally and lack an information-theoretic foundation. We study the composite metric $C = \sqrt{A \cdot T}$, where A is the angular similarity of document embeddings and $T = 1 - d_{JS}$ is the topic proximity through the Jensen–Shannon distance of soft cluster memberships, and we provide an information-geometric reading of this metric together with an axiomatic characterization of the geometric-mean combinator. On the product manifold $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$, the negative log-coherence decomposes additively into an angular and a topic cost. Because the Riemannian metric tensor induced by the Jensen–Shannon distance on the simplex is proportional to the Fisher information matrix, the topic component is locally consistent with the Fisher–Rao metric singled out by Chentsov’s theorem. Within a parametric family of combinators (the compensability spectrum), the geometric mean is the unique combinator consistent with four natural axioms (a boundary/veto condition, symmetry, log-additivity, normalization), and the construction also motivates a proper product metric $d_\times$ that we use as a reference distance. Experiments on four corpora spanning news and academic domains (40 to 6000 documents), three general-purpose embedding families (GPT-4/ada-002, MP-Net, MiniLM-L6) plus citation-aware SPECTER2, and three alternative topic models (LDA, soft k -means, GMM) are consistent with the framework: the Fisher identity holds with $R \geq 0.99$, the geometric mean tracks $d_\times$ closely ($\rho = 0.999$), and a downstream LLM-as-judge consistency check shows that the geometric mean is not empirically dominated by any alternative combinator or single-channel baseline. Sweeping the compensability spectrum, the bottleneck-coherence gap between extracted storylines and random sequences splits into a symmetric component—maximized at the geometric mean on the four corpora above and a fifth, human-navigation corpus—and a displacement term; a cross-modal case study on a human-curated image narrative reproduces the same effect in a second modality. Together, these results provide an information-geometric justification for the composite coherence metric and articulate the conditions under which the geometric mean is the natural choice.



Academic Editor: Geert Verdoolaeghe

Received: 12 April 2026

Revised: 8 June 2026

Accepted: 24 June 2026

Published: 28 June 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: information geometry; Fisher–Rao metric; Jensen–Shannon divergence; narrative extraction; coherence metric; product manifold; Chentsov theorem; axiomatic characterization; scale complementarity; information theory

1. Introduction

Narratives are fundamental structures through which humans organize and communicate knowledge about sequences of events [1,2]. Computational narrative extraction seeks

to automatically discover these structures from document collections, and has applications in intelligence analysis [3], journalism [4], and social science [5].

At the core of graph-based narrative extraction methods lies a *coherence function* that measures how smoothly one event transitions into another. Following the standard event-based formulation, in which each document represents a single event [5]—a common and natural assumption for news corpora, whose scope we discuss in Section 2.1—a corpus of n documents induces a coherence graph $G = (V, E, C)$ with documents as nodes and edge weights $C(d_i, d_j)$ quantifying the transition quality between events d_i and d_j . Storylines, the basic unit of a narrative, are paths through this graph that maximize some aggregate of edge coherence [2,3,6,7].

A survey by Keith Norambuena et al. [5] documents the diverse, operational ways coherence has been defined: word-influence random walks [3], KL divergence with temporal decay [8], and, in multi-criteria path objectives, a coherence term weighted alongside auxiliary criteria such as relevance and coverage [9]. Each definition captures an intuitive aspect of narrative smoothness, but none of them is derived from an underlying geometry of the document space, none is shown to satisfy proper metric properties, and none is the consequence of an explicit set of axioms. These three concerns—absence of geometric structure, lack of true metric properties, and absence of an axiomatic foundation—are related but distinct, and the present paper addresses all three in a single framework.

Keith and Mitra [2] introduced, in their narrative maps framework, a composite coherence metric,

$$C(d_i, d_j) = \sqrt{A(e_i, e_j) \cdot T(\hat{e}_i, \hat{e}_j)} \quad (1)$$

that combines the angular similarity A of the document embeddings with the topic similarity $T = 1 - d_{JS}$ (where $d_{JS} = \sqrt{JSD}$ is the Jensen–Shannon distance) of the soft cluster membership distributions through the geometric mean. While this metric performs well empirically, an information-theoretic justification has not been formally established for it.

This paper makes four contributions:

Geometric contribution (Section 3). We give a product-manifold reading of the composite coherence metric on $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$, on which the negative log-coherence decomposes additively into an angular and a topic cost. Because the Riemannian tensor induced by the Jensen–Shannon distance on the simplex is proportional to the Fisher information matrix, the topic component is locally consistent with the Fisher–Rao geometry singled out by Chentsov’s theorem [10]. We also define a proper product metric $d_\times$ that we use as a reference distance in the empirical analysis.

Axiomatic contribution (Sections 4 and 5). Within a parametric compensability spectrum of power-mean combinatorial means (Min, HM, GM, AM, Max), we characterize the geometric mean as the unique combinatorial mean consistent with four natural axioms (a boundary/veto condition, symmetry, log-additivity, normalization). We discuss explicitly which axioms are load-bearing for this uniqueness statement and how dropping each one enlarges the admissible family. Adjoined to this spectrum is the Quad combinatorial mean $1 - d_\times^2$, which is not itself a power mean but arises directly as the similarity-domain counterpart of the product metric. A further proposition decomposes the bottleneck-coherence gap between coherent and incoherent storylines, swept across the compensability spectrum, into an even and an odd part: the even part is maximized at the geometric mean exactly when incoherent storylines are more channel-imbalanced—the design premise of the coherence metric—and the odd part only displaces the observed peak. Section 5 records the formal properties of this combinatorial family: the power-mean compensability ordering, the position of the Quad combinatorial mean, and the metric structure of the bounded dissimilarities (notably that $1 - C_{GM}$ is not a metric).

Information-theoretic contribution (Section 6). We connect each per-channel cost to an exact classical information-theoretic quantity: the angular cost is the surprisal of a random-hyperplane (SimHash) collision, and the topic cost is a strictly increasing function of mutual information. A data-processing argument further shows that the two channels capture complementary scales of abstraction (*scale complementarity*), so that the composite is genuinely more informative than either channel alone.

Empirical contribution (Section 7). We validate the framework on six corpora drawn from the narrative-extraction literature. The metric-level and downstream analyses use four event-based corpora—Cuba, COVID, VisPub, and AMiner (40 to 6000 documents) from the narrative maps [2] and narrative trails [7] repositories—together with three general-purpose embedding families plus citation-aware SPECTER2 and three alternative topic-model families (LDA, soft k -means, GMM). On these, the Fisher identity holds with $R \geq 0.99$, the geometric mean tracks $d_\times$ closely ($\rho = 0.999$), and a downstream LLM-as-judge consistency check shows that the geometric mean is not empirically dominated by any alternative combinator or single-channel baseline (Friedman $p \geq 0.099$). The bottleneck-gap analysis then adds the Wikispeedia human-navigation corpus [11] as a non-circular, human-grounded anchor and, as a cross-modal case study, the ROGER expedition-photograph corpus [12]: throughout, the metric’s design premise holds so that the symmetric component of the separation between extracted storylines and random sequences is maximized at the geometric mean. A secondary observation from the downstream comparison is that different combinators induce structurally different storylines at statistically indistinguishable aggregate qualities, so the choice of combinator is a modeling decision about which narrative structures to foreground rather than a free parameter to be tuned for quality (Section 8).

The remainder of the paper is organized as follows. Section 2 fixes notation and reviews related work. Section 3 develops the product-manifold reading and its local Chentsov compatibility; Section 4 gives the axiomatic characterization of the geometric mean together with the bottleneck-gap analysis; Section 5 records the metric’s formal properties; and Section 6 connects the per-channel costs to classical information theory. Section 7 reports the experimental validation; Section 8 discusses implications and limitations; and Section 9 concludes.

2. Preliminaries and Related Work

2.1. Event-Based Narrative Extraction

We consider a temporally ordered corpus $\mathcal{D} = \{d_1, \dots, d_n\}$ in which each document represents a single event. Operationalizing events at the document level is the standard assumption in event-based narrative extraction [5,7]: it is natural for news corpora, where an article—especially a breaking-news article—typically reports one main event, and it lets the narrative be studied at the level of inter-document transitions. Finer (sub-document or multi-event) resolutions are possible [5] but are orthogonal to the coherence question studied here. A *coherence graph* $G = (V, E, C)$ has $V = \mathcal{D}$, edges E connecting temporally compatible documents (i.e., $(d_i, d_j) \in E$ only if $i < j$, enforcing forward temporal ordering), and weights $C: V \times V \rightarrow [0, 1]$ measuring transition quality. A *storyline* is a path $\pi = (d_{i_1}, \dots, d_{i_L})$ through G —a temporally ordered sequence of L documents ($i_1 < \dots < i_L$), so that L is the number of events in the storyline. A *narrative map* is a collection of overlapping storylines forming a DAG [2], though a DAG structure is not a strict requirement [5,7]. The quality of π depends on the coherence of its edges: common objectives include maximizing the minimum edge coherence (maximin paths [7]) or the geometric mean of edge coherences.

The survey by Keith Norambuena et al. [5] catalogues nine distinct mathematical formulations of narrative coherence across three event resolution levels (clusters, docu-

ments, sentences), none of which is grounded in a principled mathematical framework. Existing metrics are either single-channel (e.g., word-influence random walks [3,6], minimum cosine similarity [13], n -gram overlap [14]) or combine multiple criteria via ad hoc rules (e.g., KL divergence with exponential temporal decay [9], sigmoid-transformed JSD [15], weighted multi-criteria scoring [16–18]). Keith and Mitra [2] introduced the composite metric $C = \sqrt{A \cdot T}$ studied in this paper, combining angular similarity of document embeddings with topic similarity via the Jensen–Shannon distance, which is the only formulation in the surveyed literature that operates on two distinct representation spaces with a multiplicative combinator. German et al. [7] adopted this metric in their subsequent work on narrative trails. Both works specify the metric operationally, without a theoretical justification. This paper provides the information-geometric justification for that specific combination.

2.2. The Composite Coherence Metric

The composite coherence of Equation (1) combines two components.

Angular similarity. Each document d_i has a dense embedding $e_i \in \mathbb{R}^d$ obtained from a language model. The angular similarity is

$$A(e_i, e_j) = 1 - \frac{\arccos(\cos(e_i, e_j))}{\pi}, \quad (2)$$

where $\cos(e_i, e_j) = e_i^\top e_j / (\|e_i\| \|e_j\|)$. Thus A ranges from 0 (antipodal) to 1 (identical direction), and the angular distance $d_{\text{ang}}(e_i, e_j) = \arccos(\cos(e_i, e_j))$ is the geodesic distance on the unit sphere $\mathbb{S}^{d-1}$ (see Section 2.4).

Topic similarity. Soft cluster membership distributions $\hat{e}_i \in \Delta^{K-1}$ are obtained by applying UMAP dimensionality reduction [19,20] (`n_neighbors` = 32, `n_components` = 48, `min_dist` = 0.0, cosine metric, fixed random seed) followed by HDBSCAN clustering [21] (`min_cluster_size` = 5, excess-of-mass cluster selection) with soft membership assignment, where $\Delta^{K-1} = \{p \in \mathbb{R}^K : p_k \geq 0, \sum_k p_k = 1\}$ denotes the probability simplex. The cluster count K is an emergent output of this configuration rather than a tuned parameter, with $K \geq 4$ the recommended operating regime (see Section 7.5, the Supplementary Materials for the cross-corpus and sensitivity analyses, and the Limitations in Section 8.4). The topic similarity is

$$T(\hat{e}_i, \hat{e}_j) = 1 - d_{\text{JS}}(\hat{e}_i, \hat{e}_j), \quad (3)$$

where $d_{\text{JS}}(p, q) = \sqrt{\text{JSD}(p, q)}$ is the Jensen–Shannon distance [22,23], the Kullback–Leibler divergence is $D_{\text{KL}}(p \| q) = \sum_k p_k \ln(p_k / q_k)$, and the Jensen–Shannon divergence is $\text{JSD}(p, q) = \frac{1}{2} D_{\text{KL}}(p \| m) + \frac{1}{2} D_{\text{KL}}(q \| m)$ with $m = \frac{1}{2}(p + q)$ [24]. Under the base-2 logarithm convention, $\text{JSD} \in [0, 1]$ and hence $d_{\text{JS}} \in [0, 1]$, so $T \in [0, 1]$. We use the JS distance (square root of the divergence) rather than the divergence itself for two related reasons. First, d_{JS} is a proper metric [22,23], which is needed for the product-metric construction of Theorem 2. Second, the topic similarity then has the same first-order structure as the angular similarity: $A = 1 - d_{\text{ang}} / \pi$ is linear in the spherical geodesic distance d_{ang} , and the relation $d_{\text{JS}} \approx d_{\text{FR}} / \sqrt{2 \ln 2}$ established in Remark 1 makes $T = 1 - d_{\text{JS}}$ linear, to first order, in the Fisher–Rao geodesic distance on the simplex. The two components thus enter the log-coherence decomposition (Proposition 1) at the same order in their respective geodesic distances, which is what we mean by “aligning” them; using the divergence JSD in place of d_{JS} would make the topic component second-order and break this symmetry (Remark 2).

Composite coherence. With the two components A and T now specified in detail, the composite coherence of Equation (1) is their geometric mean, $C = \sqrt{A \cdot T}$. Why this

particular combinator, applied to these particular components, is well-founded is the question the rest of this paper addresses.

2.3. Information Geometry

The Fisher–Rao metric was introduced by Rao [25] and studied systematically by Amari [26,27]. For textbook treatments, see Ay et al. [28] and Nielsen [29].

Information geometry in text analysis. Information-geometric methods have been applied to several NLP tasks, though not previously to narrative coherence. Lafferty and Lebanon [30] introduced diffusion kernels on the multinomial manifold for text classification, exploiting the Fisher–Rao geometry of word distributions. Lebanon [31] extended this to margin classifiers adapted to the simplex geometry. Colombo et al. [32] used the Fisher–Rao distance between masked language model output distributions as an evaluation metric for text generation, finding it outperforms standard n -gram metrics. Dhillon et al. [33] used the generalized Jensen–Shannon divergence for information-theoretic word clustering, minimizing within-cluster JSD. In a different geometric direction, Nickel and Kiela [34] embedded symbolic data in hyperbolic (constant negative curvature) spaces to capture hierarchical structure. Our work differs in that the geometry arises from the coherence metric’s structure rather than being imposed as a design choice: the product manifold $\mathbb{S}^{d-1} \times \Delta^{K-1}$ and the Fisher–Rao connection emerge from analyzing an existing metric, not from constructing a new one.

We now recall the key definitions used in the rest of the paper.

Jensen–Shannon divergence. Lin [24] popularized the JSD as a symmetrized, bounded variant of the KL divergence. It satisfies $\text{JSD}(p, q) \geq 0$ with equality iff $p = q$. Endres and Schindelin [22] and Österreicher and Vajda [23] independently proved that $\sqrt{\text{JSD}}$ is a metric. Fuglede and Topsøe [35] showed that $\sqrt{\text{JSD}}$ admits an isometric embedding into a Hilbert space.

Fisher information matrix. For a discrete distribution $p = (p_1, \dots, p_K)$ on a finite alphabet, the Fisher information matrix has entries

$$\mathbf{F}(p)_{kl} = \frac{\delta_{kl}}{p_k}, \quad (4)$$

where δ_{kl} is the Kronecker delta. This is a diagonal matrix with entries $1/p_k$.

Fisher–Rao manifold. The open probability simplex $\Delta_+^{K-1} = \{p \in \mathbb{R}^K : p_k > 0, \sum_k p_k = 1\}$, equipped with the Riemannian metric tensor $g_{\text{FR}}(p) = \mathbf{F}(p)$ induced by the Fisher information matrix, is the Fisher–Rao manifold $(\Delta_+^{K-1}, g_{\text{FR}})$. The Bhattacharyya angle is

$$d_{\text{FR}}(p, q) = \arccos\left(\sum_{k=1}^K \sqrt{p_k q_k}\right) \in [0, \pi/2]. \quad (5)$$

The Fisher–Rao manifold has constant positive sectional curvature $\kappa = 1/4$, and via the square-root parameterization $\xi_k = 2\sqrt{p_k}$, it maps isometrically onto a piece of the sphere $\mathbb{S}_+^{K-1}$ of radius 2 [27]. Note that d_{FR} as defined is the Bhattacharyya angle, i.e., the angular separation on the *unit* sphere under $\eta_k = \sqrt{p_k}$. Infinitesimally, $\delta p^\top \mathbf{F}(p) \delta p = 4 d_{\text{FR}}^2$. The factor of 4 arises because the metric (4) corresponds to the radius-2 sphere, whose geodesic distance is $2d_{\text{FR}}$.

Chentsov’s theorem. Chentsov’s uniqueness theorem [10] is a cornerstone of information geometry. Among all Riemannian metrics on the manifold of probability distributions on a finite sample space, the Fisher–Rao metric is the unique metric (up to a global scaling constant) that is invariant under sufficient statistics (equivalently, under the congruent Markov embeddings of the sample space that they induce). See Campbell [36] for an extension and Bauer, Bruveris, and Michor [37] for a modern proof in the infinite-dimensional

setting. The scope of this uniqueness statement is important: it characterizes the metric *tensor* (an infinitesimal object), not the global finite-distance distance function, so any monotone reparameterization of arc length yields a different scalar distance with the same underlying Riemannian geometry. We will use Chentsov’s theorem in this restricted sense throughout.

JSD–Fisher connection. The second-order expansion

$$\text{JSD}_{\text{nat}}(p, p + \delta p) = \frac{1}{8} \delta p^\top \mathbf{F}(p) \delta p + O(\|\delta p\|^3) \quad (6)$$

(equivalently $\frac{1}{2} d_{\text{FR}}^2$, since $\delta p^\top \mathbf{F} \delta p = 4 d_{\text{FR}}^2$ for the Bhattacharyya angle (5)) is a well-known identity in information geometry [29]. The novelty in the present paper is not the identity itself but its *application*: recognizing that the topic component of an existing narrative coherence metric inherits, at the level of the Riemannian metric tensor, the Fisher–Rao geometry that Chentsov’s theorem singles out among invariant metrics on the simplex (Section 3.3).

Logarithm base and JS distance. The identity (6) is stated for the natural-logarithm convention, written JSD_{nat} . Throughout the rest of the paper, JSD (unsubscripted) denotes the base-2 Jensen–Shannon divergence, $\text{JSD} = \text{JSD}_{\text{nat}} / \ln 2 \in [0, 1]$, for which the identity becomes

$$\text{JSD}(p, p + \delta p) = \frac{1}{8 \ln 2} \delta p^\top \mathbf{F}(p) \delta p + O(\|\delta p\|^3), \quad (7)$$

so that $\text{JSD} \approx d_{\text{FR}}^2 / (2 \ln 2)$ for nearby distributions. Since the topic similarity uses the Jensen–Shannon distance $d_{\text{JS}} = \sqrt{\text{JSD}}$, we obtain

$$d_{\text{JS}}(p, p + \delta p) \approx \frac{d_{\text{FR}}(p, p + \delta p)}{\sqrt{2 \ln 2}} + O(\|\delta p\|^2), \quad (8)$$

which is *first-order* in the Fisher–Rao distance, paralleling the first-order relationship $A = 1 - d_{\text{ang}} / \pi$ on the angular side. Two levels of this connection must be distinguished. At the *infinitesimal* level, the relationship is **exact**: the Riemannian metric tensor induced by d_{JS} on Δ_+^{K-1} is $(1 / (8 \ln 2)) \mathbf{F}(p)$, i.e., proportional to the Fisher information matrix, so d_{JS} induces the same Riemannian geometry as the Fisher–Rao metric (up to the global scale factor $1 / \sqrt{2 \ln 2}$). At *finite distances*, the relationship (8) is approximate, with higher-order corrections that vanish as $\|\delta p\| \rightarrow 0$. Empirically, the finite-distance approximation achieves a Pearson correlation $R \geq 0.99$ across all corpora (Section 7.5); throughout, R denotes the Pearson product-moment correlation coefficient (not its square) and ρ a rank correlation.

Finite-distance justification. Chentsov’s theorem guarantees uniqueness of the *metric tensor* (an infinitesimal object), but the coherence metric operates at finite distances. This is principled for three reasons. (a) The metric tensor induced by d_{JS} on Δ_+^{K-1} is *exactly* $(1 / (8 \ln 2)) \mathbf{F}(p)$, a constant multiple of the Fisher information matrix, so d_{JS} and $d_{\text{FR}} / \sqrt{2 \ln 2}$ have the same unparameterized geodesics (the global constant rescales arc length and sectional curvature). (b) The product decomposition (Proposition 1) uses d_{JS} directly, not its Fisher–Rao approximation. (c) Writing $\text{BC}(p, q) = \sum_k \sqrt{p_k q_k} = \cos d_{\text{FR}}$ for the Bhattacharyya coefficient (exact, by (5)), the known bound $\frac{1}{2}(1 - \text{BC}) \leq \text{JSD}_{\text{nat}} \leq 1 - \text{BC}$ [38] sandwiches JSD_{nat} between two explicit functions of the Fisher–Rao distance at every distance, not merely infinitesimally—equivalently $\frac{1}{4} H^2 \leq \text{JSD}_{\text{nat}} \leq \frac{1}{2} H^2$ with $H^2 = 2(1 - \text{BC})$ the squared Hellinger distance (Proposition 7)—and provides explicit non-asymptotic control, confirmed empirically by $R \geq 0.99$ across all four corpora (Section 7.5).

2.4. Spherical and Product-Manifold Geometry

The unit sphere in $\mathbb{R}^d$ is denoted $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. For unit vectors $u, v \in \mathbb{S}^{d-1}$, the angular distance $d_{\text{ang}}(u, v) = \arccos(u^\top v)$ is the geodesic distance of the round metric on $\mathbb{S}^{d-1}$. The sphere is a Riemannian manifold of constant positive sectional curvature 1, and the normalized angular similarity $A = 1 - d_{\text{ang}}/\pi$ maps $[0, \pi] \rightarrow [0, 1]$.

On the topic side, the relevant construction is the product Riemannian manifold: $M_1 \times M_2$ with the product metric $g = g_1 \oplus g_2$ is standard in Riemannian geometry [39]. In multi-view learning, combining similarity functions from different feature spaces is common [40], but the standard approach is additive (kernel sums) rather than multiplicative (geometric mean). The geometric mean of positive-definite matrices has been studied extensively in matrix analysis [41], but its use as a combinator of scalar similarities has been less often explored.

3. The Product Manifold Structure

We now show that the composite coherence metric admits a natural interpretation as a distance on a product Riemannian manifold. The tools we use (product Riemannian manifolds, the Fisher–Rao metric, and Chentsov’s theorem) are well-established in information geometry [27,28]. The contribution is recognizing that they apply to this specific coherence metric from the narrative extraction literature, yielding consequences for its justification and diagnostic use that were not previously apparent.

Formal results, interpretations, and empirical checks. Sections 3–6 mix three kinds of statements, and we keep them typographically distinct so that the logical status of each conclusion is clear. *Theorems, propositions, and corollaries* (with proofs) state what is formally established under explicitly stated hypotheses. Paragraphs labeled **Interpretation** give the conceptual or geometric reading of a formal result; they are not themselves proven and should be read as motivation. Paragraphs labeled **Empirical check**, and all of Section 7, report measurements on real corpora; these support the framework but do not extend the proofs. Where a result is exact and where it is approximate or corpus-dependent is stated explicitly at each occurrence.

3.1. Two Geometric Spaces

The composite coherence metric operates on two distinct Riemannian manifolds. The angular component $A = 1 - d_{\text{ang}}/\pi$ measures proximity on $\mathbb{S}^{d-1}$ via the geodesic distance of the round metric: after ℓ^2 -normalization, the embeddings $\bar{e}_i = e_i/\|e_i\|$ lie on $\mathbb{S}^{d-1}$ and $d_{\text{ang}}(\bar{e}_i, \bar{e}_j) = \arccos(\bar{e}_i^\top \bar{e}_j)$ is that geodesic distance. The topic component $T = 1 - d_{\text{JS}}$ measures proximity on Δ_+^{K-1} via the Jensen–Shannon distance $d_{\text{JS}} = \sqrt{\text{JSD}}$, which is a proper metric [22,23] and is locally Fisher–Rao-compatible in the sense recorded in Section 2.3.

3.2. Coherence as Product-Manifold Cost

The decomposition that we record next is algebraically immediate—it follows from $\log \sqrt{AT} = \frac{1}{2}(\log A + \log T)$ —and its value lies in the *geometric reading* it enables. Read as a cost on a product Riemannian manifold, the identity puts the angular and topic terms on the same footing, and this product-manifold picture is what subsequently motivates the product metric $d_\times$ (defined and shown to be a proper metric in Theorem 2), the diagnostic attribution it enables (Section 4.2), and the axiomatic characterization of the geometric mean (Section 4.3). Readers familiar with product spaces and combined distance functions may find Proposition 1 unsurprising on its own; its role here is structural as the pivot between the metric-construction and axiomatic strands of the framework.

Proposition 1 (Product decomposition). Let $d_{\text{ang}} \in [0, \pi]$ be the angular distance on $\mathbb{S}^{d-1}$ and $d_{\text{JS}} \in [0, 1]$ the base-2 Jensen–Shannon distance on Δ_+^{K-1} . For any documents d_i and d_j with $C(d_i, d_j) > 0$ (i.e., $d_{\text{ang}} < \pi$ and $d_{\text{JS}} < 1$),

$$-\log C(d_i, d_j) = \underbrace{\frac{1}{2} [-\log(1 - d_{\text{ang}}/\pi)]}_{\text{angular cost}} + \underbrace{\frac{1}{2} [-\log(1 - d_{\text{JS}})]}_{\text{topic cost}}. \quad (9)$$

Both cost functions $h(x) = -\log(1 - x)$ are monotonically increasing and convex on $[0, 1)$, so C is monotonically decreasing in each component distance separately; it is not, however, a monotonic function of the product metric $d_{\times}$ (Theorem 2).

Proof. Direct algebraic identity: $-\log C = -\log \sqrt{AT} = -\frac{1}{2}(\log A + \log T) = \frac{1}{2}[-\log(1 - d_{\text{ang}}/\pi)] + \frac{1}{2}[-\log(1 - d_{\text{JS}})]$. $\square$

Figure 1 illustrates the product-manifold reading: each document pair contributes one geodesic on the sphere factor and one geodesic on the simplex factor, and the log-coherence $-\log C$ adds the two costs with equal weight.

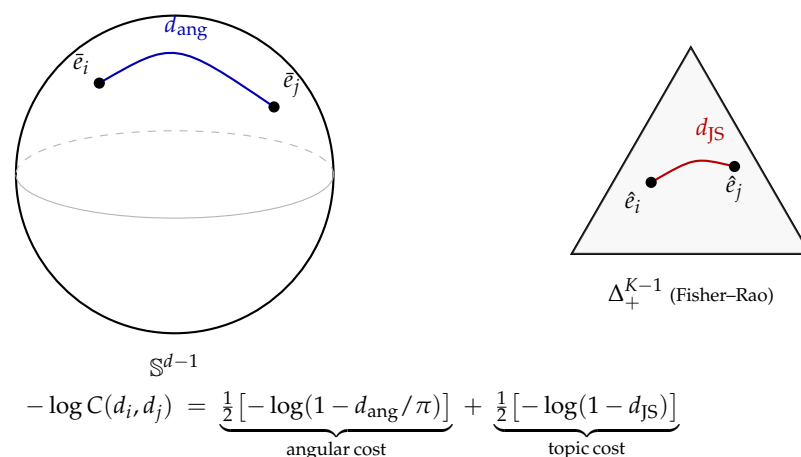


Figure 1. The product-manifold reading of the composite coherence metric. The angular component of the document representation lives on the unit sphere $\mathbb{S}^{d-1}$, and the topic component lives on the open simplex Δ_+^{K-1} equipped with the Jensen–Shannon (Fisher–Rao-compatible) geometry. Each document i is represented by an angular vector $\bar{e}_i$ on the sphere and a topic distribution $\hat{e}_i$ on the simplex; for a document pair, the geodesic on each factor has length d_{ang} and d_{JS} respectively. The log-coherence decomposes additively into an angular cost and a topic cost, one per factor of the product manifold.

Remark 1 (Fisher–Rao connection). The JSD–Fisher identity (8) gives $d_{\text{JS}} \approx d_{\text{FR}}/\sqrt{2 \ln 2}$ for nearby distributions. Substituting into the topic term of (9) yields

$$-\log C \approx \frac{1}{2} [-\log(1 - d_{\text{ang}}/\pi)] + \frac{1}{2} [-\log(1 - d_{\text{FR}}/\sqrt{2 \ln 2})], \quad (10)$$

connecting the coherence cost to the Fisher–Rao distance, which is the Riemannian metric on the statistical manifold singled out by Chentsov’s theorem (Theorem 1). The angular term is left exact. Only the $d_{\text{JS}} \rightarrow d_{\text{FR}}$ substitution is approximate, and it is only defined where $d_{\text{FR}}/\sqrt{2 \ln 2} < 1$; its finite-distance accuracy on the corpus is reported in the approximation-quality test (Section 7.2).

Remark 2 (Infinitesimal limit). In the limit of vanishing distances ($d_{\text{ang}} \rightarrow 0$, $d_{\text{JS}} \rightarrow 0$), Taylor-expanding each term of (9) via $-\log(1 - x) = x + x^2/2 + O(x^3)$ yields the exact expansion:

$$-\log C = \frac{1}{2}(d_{\text{ang}}/\pi + d_{\text{JS}}) + \frac{1}{4}((d_{\text{ang}}/\pi)^2 + d_{\text{JS}}^2) + O(\epsilon^3). \quad (11)$$

Both the angular and topic contributions enter at the same order: the first-order terms are $d_{\text{ang}}/(2\pi)$ and $d_{\text{JS}}/2$. This structural symmetry arises from using the Jensen–Shannon distance $d_{\text{JS}} = \sqrt{\text{JSD}}$ (which is first-order in d_{FR} by Equation (8)) rather than the divergence JSD (which is second-order in d_{FR}).

If one further substitutes $d_{\text{JS}} \approx d_{\text{FR}}/\sqrt{2\ln 2}$ (valid at first order), the leading terms become $\frac{1}{2}(d_{\text{ang}}/\pi + d_{\text{FR}}/\sqrt{2\ln 2})$. This first-order relation cannot, however, be substituted throughout while keeping the second-order expansion exact: since $d_{\text{JS}} = d_{\text{FR}}/\sqrt{2\ln 2} + O(d_{\text{FR}}^2)$, the first-order term $\frac{1}{2}d_{\text{JS}}$ already carries an $O(d_{\text{FR}}^2)$ correction—of the same order as the second-order terms—so re-expressing the whole expansion in d_{FR} would drop a second-order contribution. (Substituting inside the second-order term itself, $d_{\text{JS}}^2 \rightarrow d_{\text{FR}}^2/(2\ln 2)$, is by contrast harmless: it errs only at $O(d_{\text{FR}}^3)$.) The second-order expansion is therefore exact only in terms of d_{JS} .

At typical corpus distances ($d_{\text{ang}}/\pi \approx 0.35$, see Section 7.2), the exact formula (9) should be used.

Since $d_x^2 = \frac{1}{2}[(d_{\text{ang}}/\pi)^2 + d_{\text{JS}}^2]$ (Theorem 2), the expansion (11) can be written $-\log C = \frac{1}{2}(d_{\text{ang}}/\pi + d_{\text{JS}}) + \frac{1}{2}d_x^2 + O(\epsilon^3)$: the squared product metric appears exactly as the second-order correction to the additive first-order cost.

3.3. Local Chentsov Compatibility of the Topic Component

Theorem 1 (Chentsov). *Among all Riemannian metrics on the manifold Δ_+^{K-1} of probability distributions over a finite alphabet of size K , the Fisher–Rao metric is the unique metric (up to a positive multiplicative constant) that is invariant under sufficient statistics—equivalently, under the congruent Markov embeddings $\Delta^{K'-1} \hookrightarrow \Delta^{K-1}$ ($K' \leq K$) that they induce [10].*

This is a classical result of information geometry, not a contribution of the present paper; we restate it here, without proof, only so that the local compatibility statement below can refer to it. See Chentsov [10] for the original, Campbell [36] for an extension, and Bauer et al. [37] for a modern proof.

Interpretation. In the restricted (metric-tensor) sense of Section 2.3, Theorem 1 bears on the present setting through the following compatibility statement.

Corollary 1 (Local Fisher–Rao compatibility of the topic component). *The Riemannian metric tensor induced by the Jensen–Shannon distance d_{JS} on Δ_+^{K-1} is exactly $(1/(8\ln 2))\mathbf{F}(p)$ (identity (7)), that is, proportional to the Fisher information matrix. The topic component $T = 1 - d_{\text{JS}}$ therefore induces a constant multiple, $1/(8\ln 2)$, of the Fisher–Rao metric tensor on the open simplex: it has the same unparameterized geodesics and the same notion of invariance under sufficient statistics, with arc length rescaled by the square root of that constant and the sectional curvature by its reciprocal (the Fisher–Rao value $1/4$ becoming $2\ln 2$). Any other smooth Riemannian distance whose metric tensor is proportional to $\mathbf{F}(p)$ shares these geodesics; at finite distances, however, it need not be a fixed function of d_{JS} —which is itself not a function of the Fisher–Rao arc length, since two pairs with equal Bhattacharyya coefficients (hence equal d_{FR}) can have different d_{JS} (Section 2.3).*

The corollary is a *local* statement, in the restricted sense of Section 2.3: it identifies the metric tensor, and does not claim that d_{JS} is a monotone function of the Fisher–Rao distance at finite distances. The role of Chentsov’s theorem in our framework is therefore to single out the Fisher–Rao geometry as the invariant local geometry on the simplex, and to record that the Jensen–Shannon distance is locally consistent with it; this is what we mean by “local Chentsov compatibility” in what follows.

Empirical check. The corollary is infinitesimal, but the linear relation $d_{\text{JS}} \approx d_{\text{FR}}/\sqrt{2\ln 2}$ it rests on holds at the finite distances of real data (Section 2.3, quantified in Section 7.5), so the local result is usable in practice. It is not part of the corollary, and the

framework does not depend on the relation being exact, since the product decomposition of Proposition 1 uses d_{JS} directly.

Why d_{JS} rather than another Fisher-compatible distance. Once Theorem 1 singles out the Fisher–Rao geometry, the Hellinger distance, the Bhattacharyya distance, and the Jensen–Shannon distance are all candidate scalar distances that share that local geometry. The Jensen–Shannon distance $d_{JS} = \sqrt{JSD}$ has three operational advantages in the present setting, beyond the metric property already required for Theorem 2 (Section 2.2). First, its square $JSD(p, q) = H(m) - \frac{1}{2}[H(p) + H(q)]$ has a direct information-theoretic reading (mutual information between sample and label under a uniform mixture; Section 6.2), which connects naturally to the notion of a narrative transition cost. Second, d_{JS} is bounded and normalized under the base-2 logarithm ($d_{JS} \in [0, 1]$), so that the topic similarity $T = 1 - d_{JS}$ lies in $[0, 1]$ without additional normalization. Third, $JSD(p, q) = H(m) - \frac{1}{2}[H(p) + H(q)]$ with $m = (p + q)/2$ is a function of the mixture and the per-distribution Shannon entropies, so it is obtained from the same entropy primitive already used for the cluster-membership vectors, without the per-coordinate products $\sqrt{p_k q_k}$ on which the Hellinger and Bhattacharyya distances rely.

3.4. Unified Spherical Structure

Proposition 2 (Product of spheres). *Under the square-root parameterization $\xi_k = 2\sqrt{p_k}$, the Fisher–Rao manifold (Δ_+^{K-1}, g_{FR}) maps isometrically onto the positive orthant of a sphere of radius 2. Both components of the composite coherence metric therefore live on spheres, and the product manifold is $\mathbb{S}^{d-1} \times \mathbb{S}_+^{K-1}$.*

Proof. From (4), the squared Fisher–Rao line element is $ds_{FR}^2 = \sum_k dp_k^2 / p_k$. Under $\xi_k = 2\sqrt{p_k}$, we have $d\xi_k = dp_k / \sqrt{p_k}$, so $dp_k^2 / p_k = d\xi_k^2$. Thus $ds_{FR}^2 = \sum_k d\xi_k^2$, which is the Euclidean metric restricted to the positive orthant of the sphere $\|\xi\| = 2$ (since $\xi_k > 0$ and $\sum_k \xi_k^2 / 4 = \sum_k p_k = 1$, so $\|\xi\| = 2$). This is the round metric on a sphere of radius 2—the classical square-root isometry recalled in Section 2.3 [27], derived here to make the product structure explicit.

The embedding sphere $\mathbb{S}^{d-1}$ carries the standard round metric. On the product $\mathbb{S}^{d-1} \times \mathbb{S}_+^{K-1}$, the product metric is $ds^2 = ds_{\mathbb{S}^{d-1}}^2 + ds_{\mathbb{S}_+^{K-1}}^2$, which is the sum of two angular distances squared. $\square$

4. Why the Geometric Mean

This section makes the case for the geometric mean. We situate it on the compensability spectrum of candidate combinatorics (Section 4.1), establish its distinguishing structural properties—log-additivity and scale invariance (Section 4.2)—give an axiomatic characterization that singles it out under four stated conditions (Section 4.3), and close with a property of extraction output across the spectrum (Section 4.4).

4.1. Candidate Combinators

We consider the following combinatorics $F: [0, 1]^2 \rightarrow [0, 1]$:

$$C_{AM} = \frac{1}{2}(A + T) \quad (\text{arithmetic mean}), \quad (12)$$

$$C_{HM} = 2AT / (A + T) \quad (\text{harmonic mean}), \quad (13)$$

$$C_{GM} = \sqrt{AT} \quad (\text{geometric mean}), \quad (14)$$

$$C_{Quad} = 1 - d_x^2 = \frac{A(2-A) + T(2-T)}{2} \quad (\text{metric-squared complement}), \quad (15)$$

$$C_{min} = \min(A, T) \quad (\text{minimum}), \quad (16)$$

$$C_{max} = \max(A, T) \quad (\text{maximum}). \quad (17)$$

More generally, the *power mean* family $M_\alpha(A, T) = ((A^\alpha + T^\alpha)/2)^{1/\alpha}$ includes AM ($\alpha = 1$), HM ($\alpha = -1$), and GM ($\alpha \rightarrow 0$) as special cases, with Min ($\alpha \rightarrow -\infty$) and Max ($\alpha \rightarrow +\infty$) as limits. The parameter α controls a single axis, the *degree of compensability*, between the two channels. The Quad combinator $C_{\text{Quad}} = 1 - d_\times^2$ is *not* a power mean; it arises directly from the product metric and is treated separately below.

The compensability spectrum. Among the power-mean combinators, as α increases, a high score in one channel increasingly compensates for a low score in the other:

$$\underbrace{C_{\min}}_{\alpha \rightarrow -\infty} < \underbrace{C_{\text{HM}}}_{\alpha = -1} < \underbrace{C_{\text{GM}}}_{\alpha \rightarrow 0} < \underbrace{C_{\text{AM}}}_{\alpha = 1} < \underbrace{C_{\max}}_{\alpha \rightarrow +\infty}$$

for all $A \neq T$ in $(0, 1)$. At the extremes, Min enforces a *bottleneck* (both channels must be good, AND-semantics), while Max enforces *sufficiency* (either channel being good is enough, OR-semantics). Each intermediate value of α corresponds to a distinct ontological claim about the relationship between angular similarity and topic similarity.

- **HM** ($\alpha = -1$). Appropriate when A and T are *rates* (e.g., averaging speeds over equal distances). Neither coherence nor transition probability is naturally a rate. Moreover, HM exhibits pathological zero-dominance ($\text{HM}(0.9, 0.01) = 0.020$), collapsing to near-Min behavior for asymmetric scores.
- **GM** ($\alpha \rightarrow 0$). Appropriate when A and T are *multiplicatively independent*, so that the joint “pass-through” quality is $A \cdot T$. This is the natural model when both quantities behave like probabilities or signal transmittances.
- **AM** ($\alpha = 1$). Appropriate when A and T contribute *additively* to quality, with one unit of angular similarity perfectly substitutable for one unit of topic similarity. This implicitly assumes commensurable units, a strong claim that is hard to justify when the two channels measure different geometric quantities.

Quad as the product-metric complement. The Quad combinator $C_{\text{Quad}} = 1 - d_\times^2 = \frac{A(2-A)+T(2-T)}{2}$ is *not* a member of the power-mean family. It arises directly from the product metric: since $d_\times$ is the natural ℓ^2 -distance on the product manifold (Theorem 2), $C_{\text{Quad}} = 1 - d_\times^2$ is the metric-squared complement. Two properties distinguish Quad from the power-mean combinators. First, *functional equivalence with $d_\times$* : since $d_\times^2$ is a monotone transformation of $d_\times$ on $[0, 1]$, maximin extraction on C_{Quad} yields the same paths as minimax on $d_\times$ (maximin extraction is invariant under monotone transforms of the edge weights; Corollary 2). In other words, Quad is the product metric in similarity-domain clothing. Second, *additive separability*: $C_{\text{Quad}} = f(A) + f(T)$ with $f(S) = S(2 - S)/2$, so the total cost $\sum_e (1 - C_{\text{Quad},e})$ decomposes into per-channel contributions for sum-based algorithms such as Dijkstra shortest paths and linear programs (LPs), analogous to the log-additive decomposition of GM (Proposition 3) but in linear rather than log space.

Despite this geometric legitimacy, Quad operates in the Euclidean domain while the component similarities A and T are bounded in $[0, 1]$ and behave like probabilities, for which the natural geometry is log-Euclidean. The geometric mean is the natural mean in this geometry: $\log \text{GM}(A, T) = \frac{1}{2}(\log A + \log T)$, i.e., the arithmetic mean in log-space. Quad computes the Euclidean complement of a non-Euclidean quantity. This is reflected empirically: C_{GM} achieves the highest rank correlation with $d_\times$ among the power-mean combinators ($\rho = 0.999$; Quad attains $\rho = 1.000$ trivially, as a monotone function of $d_\times$), despite not being derived from $d_\times$, and the log-transform better captures the product-manifold geometry than the linear complement does.

Figure 2 illustrates how the spectrum behaves on a balanced and an imbalanced similarity pair. The geometric mean’s intermediate position is what permits the veto property at the imbalanced end while preserving meaningful gradation in the balanced regime.

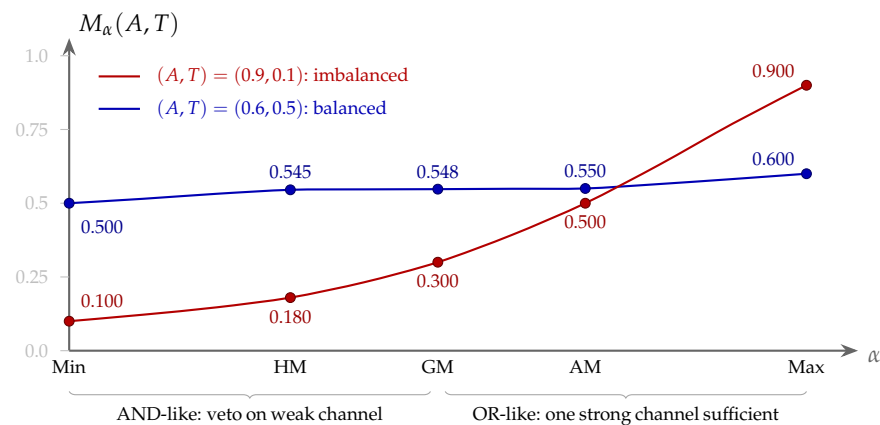


Figure 2. The compensability spectrum of power-mean combiners M_α evaluated at two example similarity pairs. The horizontal axis runs over the spectrum from $\text{Min} = \min(A, T)$ (as $\alpha \rightarrow -\infty$) through HM, GM, and AM to $\text{Max} = \max(A, T)$ (as $\alpha \rightarrow +\infty$), and each curve is drawn over the full coherence range it attains. For a balanced pair $(A, T) = (0.6, 0.5)$ (blue) the output stays within the narrow band $[0.50, 0.60]$ across the whole spectrum, so the choice of combinator hardly matters; for an imbalanced pair $(A, T) = (0.9, 0.1)$ (red) it sweeps the full range, from the minimum 0.10 through 0.18 (HM), 0.30 (GM), and 0.50 (AM) to the maximum 0.90, so the choice of combinator is operationally consequential. The geometric mean sits at $\alpha = 0$ and balances the two regimes.

4.2. Log-Additivity and Scale Invariance

Proposition 3 (Log-additivity). *The geometric mean is the unique combinator in the power-mean family such that $-\log C$ decomposes additively:*

$$-\log C_{\text{GM}} = -\frac{1}{2} \log A - \frac{1}{2} \log T. \quad (18)$$

This additive, cross-term-free structure parallels that of the Riemannian product metric $g = g_1 \oplus g_2$ on $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$, whose line element satisfies $ds^2 = ds_1^2 + ds_2^2$: in both, each factor contributes a separate term, with no cross-channel interaction. The correspondence is structural—each cost $-\log(1 - d_i)$ is first-order, not quadratic, in its geodesic distance (Remark 2)—and is precisely the geometric consistency between C_{GM} and $d_\times$ discussed after Theorem 2; the combinator derived from $d_\times$ itself is Quad (Section 4.1).

Proof. Throughout, $A, T \in (0, 1]$, so that all logarithms are defined. We first verify that the geometric mean is log-additive. The member at $\alpha = 0$ is defined by continuity, and the limit is classical [42]: applying L'Hôpital's rule to $\log M_\alpha$,

$$\lim_{\alpha \rightarrow 0} \frac{\log((A^\alpha + T^\alpha)/2)}{\alpha} = \lim_{\alpha \rightarrow 0} \frac{A^\alpha \log A + T^\alpha \log T}{A^\alpha + T^\alpha} = \frac{\log A + \log T}{2},$$

so $M_0 = \sqrt{AT} = C_{\text{GM}}$, and $-\log C_{\text{GM}} = -\frac{1}{2} \log A - \frac{1}{2} \log T$ is additive.

Next we show that no other member of the family is log-additive. For $\alpha \neq 0$,

$$-\log M_\alpha = -\frac{1}{\alpha} \log \frac{A^\alpha + T^\alpha}{2},$$

which decomposes into a sum of separate functions of $\log A$ and $\log T$ if and only if $A^\alpha + T^\alpha$ factors as $F(A)G(T)$ for some functions F, G . Suppose such a factorization holds for all $A, T \in (0, 1]$. Setting $T = 1$ and $A = 1$ in turn gives $F(A)G(1) = A^\alpha + 1$ and $F(1)G(T) = 1 + T^\alpha$, with $F(1)G(1) = 2 \neq 0$, so that

$$A^\alpha + T^\alpha = F(A)G(T) = \frac{(A^\alpha + 1)(T^\alpha + 1)}{2}.$$

Evaluating at $A = T$ gives $4A^\alpha = (A^\alpha + 1)^2$, i.e., $(A^\alpha - 1)^2 = 0$ for all $A \in (0, 1]$. Hence $A^\alpha \equiv 1$, forcing $\alpha = 0$, a contradiction. Thus no power mean with $\alpha \neq 0$ is log-additive.

Finally, the limiting members $M_{+\infty} = \max(A, T)$ and $M_{-\infty} = \min(A, T)$ of the spectrum (Section 4.1) are excluded as well. Writing $x = -\log A$ and $y = -\log T$ (so $x, y \geq 0$), we have $-\log M_{+\infty} = \min(x, y)$ and $-\log M_{-\infty} = \max(x, y)$. If $\min(x, y) = f(x) + g(y)$, then evaluating at $(x, y) = (0, 0), (1, 0), (0, 1)$ gives $f(0) + g(0) = 0$ and $f(1) + g(0) = f(0) + g(1) = 0$, so $f(1) + g(1) = 0$, contradicting $\min(1, 1) = 1$; the same four points rule out $\max(x, y)$, for which they would force $f(1) + g(1) = 2$ against $\max(1, 1) = 1$. The geometric mean is therefore the unique log-additive member of the power-mean family. $\square$

Theorem 2 (Associated product metric). Write $d_{\text{ang}} := d_{\text{ang}}(d_i, d_j)$ for the angular distance on $\mathbb{S}^{d-1}$ and $d_{\text{JS}} := d_{\text{JS}}(\hat{e}_i, \hat{e}_j)$ for the Jensen–Shannon distance on Δ_+^{K-1} . The product distance

$$d_{\times}(d_i, d_j) = \frac{1}{\sqrt{2}} \sqrt{(d_{\text{ang}}/\pi)^2 + d_{\text{JS}}^2} \quad (19)$$

is a proper metric on the document space (throughout the paper, “metric on the document space” means a metric on the representation space $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$; on documents it is a pseudometric, since two distinct documents with identical representations lie at a distance of zero) taking values in $[0, 1]$: it is the ℓ^2 -product of two proper metrics, normalized by the diameter $\sqrt{2}$ of the unit square $[0, 1]^2$. The coherence C_{GM} is monotone decreasing in d_{ang} and in d_{JS} separately. However, C_{GM} is not a function of $d_{\times}$ alone: pairs at the same $d_{\times}$ can have different coherences depending on how the distance is distributed between channels.

Proof. Both d_{ang}/π and d_{JS} [22,23] are proper metrics on their respective spaces with values in $[0, 1]$. The ℓ^2 -product of two metrics is a metric on the Cartesian product (a standard result in metric geometry [43]), so $\sqrt{(d_{\text{ang}}/\pi)^2 + d_{\text{JS}}^2}$ is a proper metric on the product space. Dividing by $\sqrt{2}$ (the diameter of $[0, 1]^2$ under the ℓ^2 norm) gives $d_{\times} \in [0, 1]$, which is also a proper metric (positive scaling preserves all metric axioms).

Monotonicity: $C = \sqrt{A \cdot T} = \sqrt{(1 - d_{\text{ang}}/\pi)(1 - d_{\text{JS}})}$. Both $A = 1 - d_{\text{ang}}/\pi$ and $T = 1 - d_{\text{JS}}$ are monotonically decreasing in their respective distances, so C is monotonically decreasing in each component distance separately. $\square$

Interpretation: three levels of relationship between C_{GM} and $d_{\times}$. The result above shows that $d_{\times}$ is a proper metric on the document space and that C_{GM} is monotonic in each component distance, but the relationship between C_{GM} and $d_{\times}$ is more nuanced than “ C_{GM} is derived from $d_{\times}$.” Three distinct levels of the relationship are worth separating, and we use this terminology consistently throughout the rest of the paper.

- **Monotone compatibility.** The rank correlation $\rho(1 - C_{\text{GM}}, d_{\times}) = 0.999$ reported in the combinator comparison (Section 7.3) is an empirical observation about the ordering of document pairs, not a derivation. Two scalar functions can produce essentially the same ranking while still being mathematically different objects.
- **Geometric consistency.** Both C_{GM} and $d_{\times}$ exist on the same product manifold $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$ and respond in compatible ways to changes in the two component distances: the log-additive decomposition of C_{GM} mirrors the additive Riemannian product metric that $d_{\times}$ instantiates (the correspondence stated with Proposition 3).
- **Metric derivation.** A metric derivation would be a stronger statement, namely that C_{GM} is a function of $d_{\times}$ alone. This is *not* the case in our setting, as Theorem 2 records. The Quad combinator $1 - d_{\times}^2$ is the combinator that *is* derived from $d_{\times}$, and it differs from C_{GM} in exactly this respect (Section 4.1).

The high rank correlation between C_{GM} and $d_\times$ supports monotone compatibility and geometric consistency; it does not establish metric derivation. We will use this distinction below when discussing why C_{GM} remains the recommended combinator despite $d_\times$ being a proper metric in its own right.

Empirical check. Theorem 2 is exact, so the quantity worth measuring is the one it does *not* fix: the rank correlation $\rho(1 - C_{GM}, d_\times) = 0.999$ (the combinator comparison, Section 7.3), which measures how closely C_{GM} tracks the product metric (monotone compatibility). This tracking is *not* implied by the theorem: the theorem holds on any document space, whereas the strength with which C_{GM} follows $d_\times$ —without being a function of it—is a property of the corpora studied.

Path cost and diagnostic attribution. The log-cost $-\log C$ is the natural edge weight for path extraction. The additive decomposition $-\log C = -\frac{1}{2} \log A - \frac{1}{2} \log T$ (Proposition 3) means that the total cost of any extracted path separates cleanly:

$$\sum_{e \in \text{path}} (-\log C_e) = \underbrace{\frac{1}{2} \sum_{e \in \text{path}} (-\log A_e)}_{\text{angular cost}} + \underbrace{\frac{1}{2} \sum_{e \in \text{path}} (-\log T_e)}_{\text{topic cost}}.$$

This enables *diagnostic attribution*. Given an extracted storyline, one can identify whether its total cost is dominated by semantic dissimilarity or topical divergence, and localize expensive edges to one channel or the other. The proper metric $d_\times$ is retained for the formal metric-space guarantees needed by triangle-inequality-dependent constructions.

The *operational* force of log-additivity depends on the extraction algorithm. For sum-based extractors (Dijkstra shortest paths, or the linear-programming formulations with coverage and connectivity constraints used in structured narrative extraction [2,3,6]), $-\log C$ serves as an additive edge cost and the per-channel budgets above directly shape which paths are optimal. For bottleneck-based extractors (maximin/minimax), any monotone transformation of C preserves path optimality (Corollary 2), so log-additivity is irrelevant to path *selection* but remains valuable as a post hoc diagnostic.

A heuristic information-theoretic intuition for why the log-additive structure and the geometric-mean summary fit together can be developed from a rate-distortion analogy under the assumptions of approximate channel independence and logarithmic per-channel rate-distortion functions; we record this analogy in the Supplementary Materials (Section S6) as a conceptual reading rather than a derivation, since the geometric mean is already characterized by Theorem 3 at the level of an exact algebraic identity.

Proposition 4 (Scale invariance). *For $\lambda > 0$, consider the common power rescaling $A' = A^\lambda$, $T' = T^\lambda$ of the similarity components—equivalently, a common multiplicative rescaling of the per-channel costs, $-\log A' = \lambda(-\log A)$ and $-\log T' = \lambda(-\log T)$. Call a combinator ranking-invariant if, for every $\lambda > 0$, the ranking of document pairs it induces on the rescaled components coincides with the ranking it induces on the originals. Within the power-mean spectrum of Section 4.1, the ranking-invariant members are exactly the geometric mean and the limits $M_{\pm\infty}$; the geometric mean is the unique ranking-invariant member with a finite exponent—equivalently, the unique one that is strictly increasing in both channels.*

Proof. *Invariance of the geometric mean.* $C'_{GM} = (A^\lambda T^\lambda)^{1/2} = (AT)^{\lambda/2}$, and $x \mapsto x^{\lambda/2}$ is strictly increasing for $\lambda > 0$, so the ranking by C'_{GM} coincides with the ranking by $C_{GM} = \sqrt{AT}$. In cost terms, $-\log C'_{GM} = \lambda(-\log C_{GM})$ by the log-additivity of Proposition 3.

Invariance of the limits. A strictly increasing map commutes with min and max, so $\min(A^\lambda, T^\lambda) = \min(A, T)^\lambda$ and $\max(A^\lambda, T^\lambda) = \max(A, T)^\lambda$: each rescaled combinator is a strictly increasing transform of the original, and the induced rankings are unchanged. We

note that $M_{\pm\infty}$ are ranking-invariant under any common strictly increasing reparameterization, being order statistics of the two channels; neither, however, is strictly increasing in both channels.

Non-invariance for finite $\alpha \neq 0$. For $\lambda > 0$ and $\alpha \neq 0$,

$$M_\alpha(A^\lambda, T^\lambda) = \left(\frac{A^{\lambda\alpha} + T^{\lambda\alpha}}{2}\right)^{1/\alpha} = \left[\left(\frac{A^{\lambda\alpha} + T^{\lambda\alpha}}{2}\right)^{1/(\lambda\alpha)}\right]^\lambda = M_{\lambda\alpha}(A, T)^\lambda, \quad (20)$$

so the ranking on rescaled components is the ranking induced by $M_{\lambda\alpha}$. Take $\lambda = 2$, so α and 2α are distinct, nonzero, and of the same sign. By the strict power-mean inequality [42], for fixed $x \neq y$ in $(0, 1)$ the map $q \mapsto M_q(x, y)$ is strictly increasing on $\mathbb{R}$. Fix q^* strictly between α and 2α and $t_- < t_+$ in $(0, 1)$, and set $a := M_{q^*}(t_-, t_+)$, $e_1 := (a, a)$, $e_2 := (t_-, t_+)$. Then $M_q(e_1) = a$ for every q , while $q \mapsto M_q(e_2)$ is strictly increasing with $M_{q^*}(e_2) = a$. Since exactly one of $\alpha, 2\alpha$ lies below q^* and the other above, M_α and $M_{2\alpha}$ rank the pair $\{e_1, e_2\}$ oppositely, so the ranking is not invariant under $\lambda = 2$. As a concrete illustration with the arithmetic mean ($\alpha = 1, \lambda = 2$): the pairs $(A, T) = (0.9, 0.1)$ and $(0.52, 0.52)$ satisfy $C_{AM} = 0.50 < 0.52$ before rescaling but $C'_{AM} = 0.41 > 0.27$ after—a strict reversal (for $\lambda > 1$, convexity of x^λ shifts the AM ranking toward imbalanced pairs; for $\lambda < 1$, concavity shifts it toward balanced ones). $\square$

Corollary 2 (Scale invariance of maximin extraction). *Consider the uniform rescaling $A' = A^\lambda$, $T' = T^\lambda$ with $\lambda > 0$ of Proposition 4. On every coherence graph, the maximin (widest-path) preorder on storylines induced by the geometric mean—together with its set of optimal storylines and the bottleneck edge of every path—is invariant under this rescaling for every $\lambda > 0$, and the same holds for the limits $M_{\pm\infty}$. Conversely, for every power mean M_α with finite $\alpha \neq 0$ there exist a coherence graph and an exponent $\lambda > 0$ for which the rescaling changes the extracted maximin storyline. The geometric mean is thus the unique finite-exponent power mean whose maximin extraction is scale-invariant in this sense—equivalently, the unique member of the spectrum that combines this invariance with strict monotonicity in both channels.*

Proof. *Step 1 (maximin extraction is ordinal).* Let G be a coherence graph with edge-coherence map $w: E \rightarrow [0, 1]$; the maximin score of a storyline π is $\sigma(\pi) = \min_{e \in \pi} w(e)$, and the extractor returns a storyline of maximal score between the prescribed endpoints. If $\varphi: [0, 1] \rightarrow [0, 1]$ is strictly increasing, then $\min_{e \in \pi} \varphi(w(e)) = \varphi(\min_{e \in \pi} w(e))$ for every path, since a strictly increasing map commutes with min. Hence $\sigma(\pi) \geq \sigma(\pi') \iff \varphi(\sigma(\pi)) \geq \varphi(\sigma(\pi'))$: replacing w by $\varphi \circ w$ leaves the maximin preorder on storylines, its set of maximizers, and the bottleneck edge $\arg \min_{e \in \pi} w(e)$ of every path unchanged.

Step 2 (invariance of M_0 and $M_{\pm\infty}$). For each of these three combinators the rescaling sends every edge coherence to its λ -th power: $M_0(A^\lambda, T^\lambda) = (A^\lambda T^\lambda)^{1/2} = M_0(A, T)^\lambda$, and, since a strictly increasing map commutes with order statistics, $\min(A^\lambda, T^\lambda) = \min(A, T)^\lambda$ and $\max(A^\lambda, T^\lambda) = \max(A, T)^\lambda$ (as in the proof of Proposition 4). In all three cases the rescaled weight map is $\varphi \circ w$ with $\varphi(x) = x^\lambda$ strictly increasing, so by Step 1 the maximin preorder, the set of optimal storylines, and every bottleneck edge are unchanged on every coherence graph.

Step 3 (non-invariance for finite $\alpha \neq 0$). Fix a finite $\alpha \neq 0$ and take $\lambda = 2$. By the proof of Proposition 4, there exist channel pairs $e_1 = (a, a)$ and $e_2 = (t_-, t_+)$ with components in $(0, 1)$ that M_α and $M_{2\alpha}$ rank oppositely; since their components lie in $(0, 1)$, $M_q(e_i) < 1$ for every exponent q . Build G on temporally ordered vertices $s < m_1 < m_2 < t$ with exactly two s – t storylines $\pi_i = (s, m_i, t)$, assigning the channel pair e_i to edge (s, m_i) and $(1, 1)$ to edge (m_i, t) . Because $M_q(1, 1) = 1 > M_q(e_i)$, the bottleneck of π_i is its first edge, so the extractor under M_α returns the storyline with the larger $M_\alpha(e_i)$. After the rescaling

$A' = A^2$, $T' = T^2$, the exponent-shift identity (20) gives edge coherences $M_{2\alpha}(\cdot)^2$, and by Step 1 (with $\varphi(x) = x^2$) the extraction is governed by the $M_{2\alpha}$ -ordering of $\{e_1, e_2\}$; since that ordering is opposite to the M_α -ordering, the extracted storyline differs before and after the rescaling.

Steps 2 and 3 together establish the corollary: within the spectrum, maximin extraction is rescaling-invariant exactly for M_0 and $M_{\pm\infty}$, and among these only the geometric mean is strictly increasing in both channels. $\square$

Operationally, Corollary 2 says that the storylines extracted under the geometric mean are insensitive to a uniform multiplicative rescaling of the per-channel costs—the otherwise arbitrary calibration of how strongly raw dissimilarity is penalized—and that it is the only power mean responsive to both channels with this property. The limits Min and Max share the insensitivity, but for the degenerate reason that order statistics commute with any common monotone recalibration, at the price of responding to only one channel at each comparison; for every other combinator the cost scale must be fixed before the extraction output is well-determined.

4.3. Conditional Axiomatic Characterization

The geometric mean is the unique combinator consistent with the four axioms stated below. The characterization is conditional on those axioms; the discussion after the proof identifies which axioms do the most work and what is admitted when each one is dropped.

Theorem 3 (Axiomatic characterization). *Let $F: [0, 1]^2 \rightarrow [0, 1]$ be a coherence combinator satisfying:*

- A1** (Boundary/veto) $F(0, t) = F(t, 0) = 0$ for all $t \in [0, 1]$, and $F(1, 1) = 1$.
- A2** (Symmetry) $F(a, t) = F(t, a)$ for all $a, t \in (0, 1]$.
- A3** (Log-additivity) *There exist functions $f, g: [0, \infty) \rightarrow [0, \infty)$ such that, for all $a, t \in (0, 1]$, $-\log F(a, t) = f(-\log a) + g(-\log t)$ (in particular $F > 0$ on $(0, 1]^2$, so the left-hand side is finite).*
- A4** (Normalization) $F(a, a) = a$ for all $a \in [0, 1]$.

Then $F(a, t) = \sqrt{a \cdot t}$.

Proof. From **A3**, letting $x = -\log a$ and $y = -\log t$ (so $x, y \geq 0$), we have $-\log F = f(x) + g(y)$ with $f, g: [0, \infty) \rightarrow [0, \infty)$, so in particular $f, g \geq 0$. The corner value $F(1, 1) = 1$ from **A1** gives $f(0) + g(0) = 0$, and since $f, g \geq 0$ this forces $f(0) = g(0) = 0$.

From **A2**, $f(x) + g(y) = f(y) + g(x)$ for all $x, y \geq 0$, which implies $f - g$ is constant. Since $f(0) - g(0) = 0$, we conclude $f = g$.

From **A4**, $F(a, a) = a$ for all $a \in (0, 1]$, so $-\log a = f(-\log a) + g(-\log a) = 2f(-\log a)$. Dividing both sides by 2: $f(x) = x/2$ for all $x \geq 0$.

Substituting $-\log F(a, t) = \frac{-\log a}{2} + \frac{-\log t}{2} = -\frac{1}{2} \log(at)$, so $F(a, t) = (at)^{1/2} = \sqrt{at}$ on the open square $(0, 1]^2$. The boundary axiom **A1** then fixes the axes, $F(0, t) = F(t, 0) = 0 = \sqrt{0 \cdot t}$, so $F(a, t) = \sqrt{at}$ on all of $[0, 1]^2$. $\square$

Scope of the characterization. Theorem 3 is stated for an arbitrary $F: [0, 1]^2 \rightarrow [0, 1]$, so axioms **A1**–**A4** single out the geometric mean among *all* coherence combinators—not only within the power-mean compensability spectrum used to introduce it. That spectrum (Min, HM, GM, AM, Max) situates the geometric mean among familiar combinators and grounds the compensability discussion of Section 4.1, but the uniqueness statement itself imposes no power-mean restriction.

No regularity is assumed. The characterization requires no continuity, monotonicity, or measurability of f and g . Functional-equation characterizations of means—of Cauchy,

Pexider, or bisymmetry type [44]—typically impose such hypotheses to exclude pathological (e.g., non-measurable) solutions; here the value $f(x)$ is pinned pointwise for every $x \geq 0$ (axiom **A4** on the diagonal, with $f = g$ from **A2**, gives $2f(-\log a) = -\log a$, and $a \mapsto -\log a$ maps $(0, 1]$ onto $[0, \infty)$) rather than left to satisfy a functional equation, so pathological solutions cannot arise.

Interpretation: each axiom and its role. Each axiom has independent motivation. The boundary/veto condition (**A1**) enforces that a zero in either component yields zero coherence while perfect similarity on both yields one. Note that the clause $F(1, 1) = 1$ is implied by axiom **A4** (at $a = 1$); it is retained in **A1** so that the boundary axiom remains self-contained when **A4** is ablated (Table 1)—a single failed channel cannot be compensated. Symmetry (**A2**) reflects that the two representation spaces play symmetric roles in our setting. Log-additivity (**A3**) is needed because sum-based narrative extraction algorithms (Dijkstra on $-\log C$ edge costs, LP formulations) require the total path cost to decompose into independent per-factor contributions. Normalization (**A4**) is a calibration requirement: when both components agree on a value a , the composite must equal a .

Which axioms are load-bearing. Each of the four axioms does independent work: dropping any one admits a combinator that is not the geometric mean, as Table 1 records. Log-additivity (**A3**) and normalization (**A4**) carry the most weight—**A3** alone cuts the entire compensability spectrum down to a single member, and **A4** fixes the calibration exponent—while symmetry (**A2**) fixes the equal-weight balance between the two channels and the boundary/veto condition (**A1**) fixes the behavior on the axes, without which the characterization pins F only on the open square $(0, 1]^2$.

Table 1. Axiom ablation for Theorem 3. For each row, the axiom in column 1 is dropped while the other three are retained, and column 2 gives at least one combinator that satisfies the remaining axioms but is not the geometric mean. The geometric mean is the unique combinator that satisfies all four.

Axiom Dropped	A Combinator Admitted by the Three Remaining Axioms	Role of the Dropped Axiom
A1: Boundary/veto	The surviving-channel extension $F(0, t) = F(t, 0) = t$, equal to $\sqrt{at}$ on the open square but returning the non-zero channel on the axes (only $F(0, 0) = 0$ is forced, by A4); the veto $F = 0$ at a zero channel is no longer enforced.	Fixes the axes (the veto); without it F is determined only on the open square.
A2: Symmetry	Weighted GM $F_w(a, t) = a^w t^{1-w}$, any $w \in (0, 1)$, $w \neq 1/2$.	Fixes equal weights ($w = 1/2$).
A3: Log-additivity	Veto-respecting non-GM power means, e.g., HM ($\alpha = -1$) and the minimum ($\alpha \rightarrow -\infty$); i.e., M_α with $\alpha < 0$.	Picks GM out of the veto-respecting compensability spectrum.
A4: Normalization	$F_\beta(a, t) = (at)^\beta$, any $\beta > 0$, $\beta \neq 1/2$.	Fixes the exponent to $\beta = 1/2$.

The empirical weight-sensitivity analysis in the Supplementary Materials (Section S2) corroborates the role of **A2**: across cluster-count configurations, the optimal angular-topic weight w^* (the exponent of the weighted geometric mean $A^w T^{1-w}$, distinct from the compensability exponent α) tracks the equal-weight value 0.5 closely (Cohen's $d \leq 0.25$ away from the degenerate $K = 2$ boundary), so the data are consistent with the symmetric

weighting that **A2** imposes. Theorem 3 is therefore best understood as a *conditional* characterization: it identifies the geometric mean as the unique combinator simultaneously consistent with log-additivity, calibration, symmetry, and the boundary/veto condition. The intent is not to argue that the geometric mean is universally optimal, only that it is the natural choice under the stated operational requirements.

4.4. The Bottleneck-Gap Profile

The arguments above are properties of the geometric mean as a combinator. We close Section 4 with a property of extraction *output*: across the compensability spectrum M_α , which exponent most sharply separates coherent storylines from incoherent ones, measured on the maximin (bottleneck) objective that narrative extractors optimize?

Fix $\varepsilon \in (0, 1)$ and let D_H, D_R be Borel probability measures on $[\varepsilon, 1]^2$ —the laws of the channel pair (A, T) of the *bottleneck edge* (the minimum-coherence edge under the geometric mean M_0 , held fixed across α) of, respectively, a coherent and an incoherent storyline. With $M_0 = \sqrt{AT}$ the geometric mean, we define the *bottleneck-gap profile*

$$g(\alpha) = \mathbb{E}_{D_H}[M_\alpha(A, T)] - \mathbb{E}_{D_R}[M_\alpha(A, T)], \quad \alpha \in \mathbb{R}. \quad (21)$$

Proposition 5 (Bottleneck-gap profile). *Let $g = E + O$ be the decomposition of (21) into an even and an odd function of α . Write $\mu_\bullet = \mathbb{E}_{D_\bullet}[M_0]$, $\Delta = \log A - \log T$, and $\varphi_\alpha = \log M_\alpha - \log M_0$. Then:*

- φ_α is odd in α , and $E(\alpha) = \mathbb{E}_H[M_0 \cosh \varphi_\alpha] - \mathbb{E}_R[M_0 \cosh \varphi_\alpha]$, $O(\alpha) = \mathbb{E}_H[M_0 \sinh \varphi_\alpha] - \mathbb{E}_R[M_0 \sinh \varphi_\alpha]$.
- $E(0) = \mu_H - \mu_R$, and $E(0) - E(\alpha) = \mathbb{E}_R[Y_\alpha] - \mathbb{E}_H[Y_\alpha]$ with $Y_\alpha = M_0(\cosh \varphi_\alpha - 1) \geq 0$. Hence the even part is maximized at $\alpha = 0$ (the geometric mean) if and only if $\mathbb{E}_R[Y_\alpha] \geq \mathbb{E}_H[Y_\alpha]$ for every α .
- g is smooth, with $g'(0) = \frac{1}{8}(\mathbb{E}_H[M_0 \Delta^2] - \mathbb{E}_R[M_0 \Delta^2])$ and $g''(0) = \frac{1}{64}(\mathbb{E}_H[M_0 \Delta^4] - \mathbb{E}_R[M_0 \Delta^4])$. Suppose $g''(0) < 0$, and fix (by continuity) $\delta > 0$ and $m > 0$ with $g'' \leq -m$ on $[-\delta, \delta]$. If $|g'(0)| < m\delta$, then g has a unique critical point α^* in $(-\delta, \delta)$; it is a strict local maximum and satisfies $|\alpha^*| \leq |g'(0)|/m$ together with

$$\left| \alpha^* + \frac{g'(0)}{g''(0)} \right| \leq \frac{L_3}{2m^3} g'(0)^2, \quad L_3 := \max_{[-\delta, \delta]} |g'''|.$$

In particular

$$\alpha^* = -\frac{g'(0)}{g''(0)} + O(g'(0)^2) = -\frac{8(\mathbb{E}_H[M_0 \Delta^2] - \mathbb{E}_R[M_0 \Delta^2])}{\mathbb{E}_H[M_0 \Delta^4] - \mathbb{E}_R[M_0 \Delta^4]} + O(g'(0)^2),$$

with implied constant $L_3/(2m^3)$.

Proof. A direct computation gives the reflection identity $M_\alpha M_{-\alpha} = AT = M_0^2$, so $\log M_\alpha + \log M_{-\alpha} = 2 \log M_0$ and $\varphi_{-\alpha} = -\varphi_\alpha$: φ_α is odd in α . With $u = \log A$, $v = \log T \in [\log \varepsilon, 0]$, one has $\log M_\alpha = K(\alpha)/\alpha$, where $K(\alpha) = \log(\frac{1}{2}(e^{\alpha u} + e^{\alpha v}))$ is the cumulant generating function of the uniform law on $\{u, v\}$; K is real-analytic on $\mathbb{R}$ with $K(0) = 0$, so $\log M_\alpha$ extends real-analytically across $\alpha = 0$ and M_α is real-analytic in α . On the compact domain $[\varepsilon, 1]^2$, M_α and all its α -derivatives are bounded uniformly on compact α -intervals, so g is C^∞ and its derivatives may be taken under the integral sign (dominated convergence).

(a) Since $M_{\pm\alpha} = M_0 e^{\pm\varphi_\alpha}$, Equation (21) gives $g(\alpha) = \mathbb{E}_H[M_0 e^{\varphi_\alpha}] - \mathbb{E}_R[M_0 e^{\varphi_\alpha}]$, and $\frac{1}{2}(g(\alpha) \pm g(-\alpha))$ yields the stated E and O ; $\cosh \varphi_\alpha$ is even and $\sinh \varphi_\alpha$ odd in α because φ_α is odd.

(b) $\cosh \varphi_0 = 1$ gives $E(0) = \mu_H - \mu_R$, and $E(0) - E(\alpha) = \mathbb{E}_R[M_0(\cosh \varphi_\alpha - 1)] - \mathbb{E}_H[M_0(\cosh \varphi_\alpha - 1)]$. As $\cosh \geq 1$, $Y_\alpha \geq 0$, so $E(\alpha) \leq E(0)$ for all α if and only if $\mathbb{E}_R[Y_\alpha] \geq \mathbb{E}_H[Y_\alpha]$ throughout.

(c) From $K(0) = 0$ and $K''(0) = \text{Var}\{u, v\} = \frac{1}{4}\Delta^2$, the analytic odd function φ_α satisfies $\varphi_\alpha = \dot{\varphi}\alpha + O(\alpha^3)$ with $\dot{\varphi} = \frac{1}{2}K''(0) = \frac{1}{8}\Delta^2$. Since E is even and O odd, $g'(0) = O'(0)$ and $g''(0) = E''(0)$; differentiating the closed forms of (a) at $\alpha = 0$ (using $\varphi_0 = 0$, $\varphi'_0 = \dot{\varphi}$, $\varphi''_0 = 0$) gives $O'(0) = \mathbb{E}_H[M_0\dot{\varphi}] - \mathbb{E}_R[M_0\dot{\varphi}]$ and $E''(0) = \mathbb{E}_H[M_0\dot{\varphi}^2] - \mathbb{E}_R[M_0\dot{\varphi}^2]$, which are the stated $g'(0)$, $g''(0)$ after substituting $\dot{\varphi} = \Delta^2/8$. If $g''(0) < 0$, continuity yields $\delta > 0$ and $m > 0$ with $g'' \leq -m$ on $[-\delta, \delta]$, so g' is strictly decreasing there. Assume $|g'(0)| < m\delta$. If $g'(0) \geq 0$ then $g'(\delta) \leq g'(0) - m\delta < 0$; if $g'(0) \leq 0$ then $g'(-\delta) \geq g'(0) + m\delta > 0$. In either case g' changes sign on $(-\delta, \delta)$ and, being strictly decreasing, has a unique zero α^* there, with $g' > 0$ to its left and $g' < 0$ to its right, so α^* is a strict local maximum of g . Integrating $g'' \leq -m$ between 0 and α^* gives $|g'(0)| \geq m|\alpha^*|$, i.e., $|\alpha^*| \leq |g'(0)|/m$. Finally, a second-order Taylor expansion of g' gives $0 = g'(\alpha^*) = g'(0) + g''(0)\alpha^* + \frac{1}{2}g'''(\xi)\alpha^{*2}$ for some ξ between 0 and α^* , rearranging gives $\alpha^* + g'(0)/g''(0) = -(g'''(\xi)/2g''(0))\alpha^{*2}$; using $|g''(0)| \geq m$ and $|\alpha^*| \leq |g'(0)|/m$ yields the stated bound. $\square$

Interpretation: the geometric mean and extraction output. Proposition 5 isolates what the geometric mean does for extraction *output*. The even part of the gap—its symmetric, combinator-balanced component—is maximized *exactly* at the geometric mean precisely when $\mathbb{E}_R[Y_\alpha] \geq \mathbb{E}_H[Y_\alpha]$, i.e., when incoherent storylines carry more channel imbalance than coherent ones. This is not an additional assumption: it is the design premise of the coherence metric itself—a coherent transition is good in *both* channels (balanced), an incoherent one fails at least one (imbalanced; a transition coherent in only one channel is vetoed, $C_{\text{GM}} = 0$, Section 5). The odd part only *displaces* the observed peak, by $-g'(0)/g''(0)$, an explicit ratio of imbalance moments of the data. Unlike the axiomatic characterization (Theorem 3), this argument refers to extracted storylines rather than to the combinator in isolation.

Empirical check. Experiment 7 (Section 7.8) measures $g(\alpha)$ on five corpora. The design-premise condition $\mathbb{E}_R[Y_\alpha] \geq \mathbb{E}_H[Y_\alpha]$ holds on every one; the even part is accordingly maximized at the geometric mean, and the observed peak of g itself is interior to the spectrum, displaced from $\alpha = 0$ by the odd part. The decomposition $g = E + O$ and the closed forms of Proposition 5 are exact; that the even part is maximized at $\alpha = 0$ is conditional on the design premise, which Experiment 7 verifies rather than assumes.

5. Formal Properties

The compensability ordering of Section 4.1 is the classical power-mean inequality: for all $A, T \in [0, 1]$ (with the convention $2AT/(A+T) := 0$ at $A = T = 0$),

$$\min(A, T) \leq \frac{2AT}{A+T} \leq \sqrt{AT} \leq \frac{A+T}{2} \leq \max(A, T), \quad (22)$$

with equality throughout iff $A = T$. This is the monotonicity of $q \mapsto M_q(A, T)$ in the exponent [42], evaluated at $q = -\infty, -1, 0, 1, +\infty$ —the same fact used in the proof of Corollary 2.

Interpretation. The geometric mean occupies a balanced middle ground: stricter than the arithmetic mean (penalizing imbalance between the components) but less harsh than the minimum (which ignores the stronger channel entirely).

Lemma 1 (Position of the Quad combinator). *The Quad combinator $C_{\text{Quad}} = 1 - d_\times^2$ is not a power mean and does not join the chain (22) at a fixed position. For all $A, T \in [0, 1]$,*

$$C_{\text{Quad}} - C_{\text{AM}} = \frac{1}{2}[A(1-A) + T(1-T)] \geq 0, \quad (23)$$

so $C_{GM} \leq C_{AM} \leq C_{Quad}$ everywhere, with $C_{AM} = C_{Quad}$ iff $A, T \in \{0, 1\}$. Its position relative to the maximum, by contrast, depends on the channel imbalance: writing $S_1 = \max(A, T)$ and $S_2 = \min(A, T)$,

$$C_{Quad} \leq C_{max} \iff S_1^2 + (1 - S_2)^2 \geq 1, \quad (24)$$

which holds for sufficiently imbalanced pairs (e.g., $(0.9, 0.1)$), where $C_{Quad} = 0.59 < 0.9$ and fails for every balanced interior pair: $A = T \in (0, 1)$ gives $C_{Quad} = A(2 - A) > A = C_{max}$.

Proof. With $C_{Quad} = \frac{1}{2}[A(2 - A) + T(2 - T)]$, equation (23) is immediate and vanishes iff each nonnegative term does. For (24), assume $A \geq T$, so $C_{max} = A$; then $C_{Quad} - A = \frac{1}{2}[T(2 - T) - A^2] = \frac{1}{2}[1 - (1 - T)^2 - A^2]$, which is ≤ 0 iff $A^2 + (1 - T)^2 \geq 1$. $\square$

Intuitively, squaring a value in $[0, 1]$ shrinks it, so the metric-squared complement $1 - d_x^2$ is the most generous combinator in the table over balanced pairs—and the same squaring, applied to the proper metric d_x , is what breaks the triangle inequality for $1 - C_{Quad} = d_x^2$ (the 8.21% violation rate in the combinator comparison, Section 7.3).

Proposition 6 (Metric structure of bounded dissimilarities). *Let $d_1 = d_{ang}/\pi$ and $d_2 = d_{JS}$ be the normalized component distances. Then:*

1. $1 - C_{AM} = (d_1 + d_2)/2$ is a proper metric (scaled ℓ^1 product);
2. $1 - C_{min} = \max(d_1, d_2)$ is a proper metric (ℓ^∞ product);
3. $1 - C_{GM} = 1 - \sqrt{(1 - d_1)(1 - d_2)}$ is not a metric in general.

Proof. For (1): $1 - \frac{A+T}{2} = \frac{(1-A)+(1-T)}{2} = \frac{d_1+d_2}{2}$; a sum of metrics is a metric, and positive scaling preserves all metric axioms. For (2): $1 - \min(A, T) = 1 - \min(1 - d_1, 1 - d_2) = \max(d_1, d_2)$, and the ℓ^∞ -product of metrics is a metric: for any x, y, z , $\max_i d_i(x, z) \leq \max_i (d_i(x, y) + d_i(y, z)) \leq \max_i d_i(x, y) + \max_i d_i(y, z)$.

For (3), consider first an abstract three-point space $\{x, y, z\}$ with $d_1(x, y) = d_1(y, z) = 0.01$, $d_1(x, z) = 0.02$ and $d_2(x, y) = 0.99$, $d_2(y, z) = 0.01$, $d_2(x, z) = 1.00$. Both components satisfy the triangle inequality (with equality in d_2 , which is permitted). Then $D_{GM}(x, z) = 1.000$ while $D_{GM}(x, y) + D_{GM}(y, z) = 0.901 + 0.010 = 0.911 < 1.000$, violating the triangle inequality; symmetry and the identity axiom hold, so only the triangle inequality fails. The configuration is realized on the actual component spaces: place the three embeddings along a great circle of S^{d-1} at normalized angular distances $d_{ang}/\pi = 0.01, 0.01, 0.02$ (so $A_{xy} = A_{yz} = 0.99$, $A_{xz} = 0.98$) and take $K = 2$ topic distributions $\hat{e}_x = (1, 0)$, $\hat{e}_y = (0.01, 0.99)$, $\hat{e}_z = (0, 1)$. The base-2 Jensen–Shannon distances are then $d_2(x, y) = 0.980$, $d_2(y, z) = 0.071$, $d_2(x, z) = 1$, giving $D_{GM}(x, y) = 0.858$, $D_{GM}(y, z) = 0.041$, $D_{GM}(x, z) = 1.000$ and the violation $1.000 > 0.858 + 0.041 = 0.899$. A sufficiently small perturbation of $\hat{e}_x$ and $\hat{e}_z$ into the open simplex Δ_+^1 preserves the strict violation by continuity (the margin at the vertex configuration is ≈ 0.10). In all cases the violation arises when the intermediate point y has extreme channel imbalance: close on one channel, far on the other. $\square$

Empirical check. In practice, HDBSCAN soft membership concentrates mass on one or two clusters but retains numerically small probabilities across the others, so membership vectors rarely approach the simplex vertices closely enough to produce the extreme channel imbalance the counterexample requires. Accordingly, no violations were detected among 10^5 sampled triplets in corpora with $K \geq 11$ (the cross-corpus validation, Section 7.5); violations appear only under degenerate clustering ($K = 2$, where membership vectors concentrate near the simplex vertices).

Figure 3 renders the configuration geometrically: the violation requires an intermediate point that is close on one channel and far on the other, which is exactly the channel-

imbalance regime that scale complementarity (Remark 3) and smooth clustering rule out in practice.

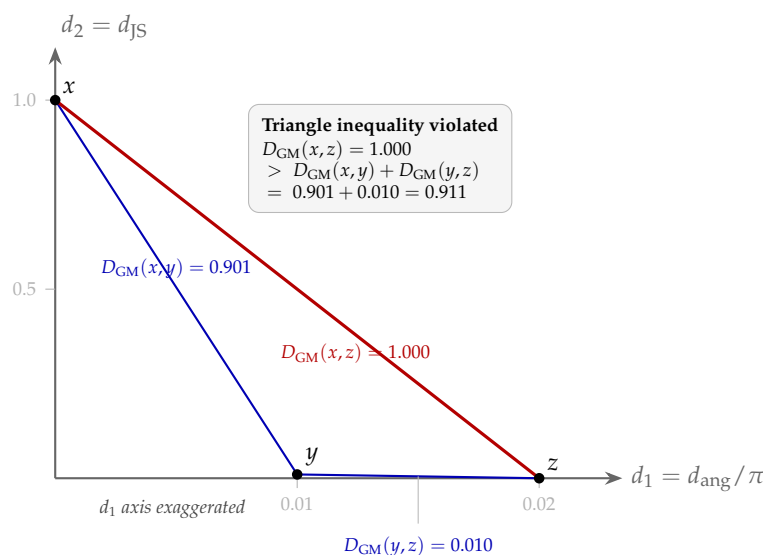


Figure 3. Geometry of the triangle-inequality counterexample for $1 - C_{GM}$. The three points x, y, z are plotted in the normalized component-distance plane $(d_1, d_2) = (d_{ang}/\pi, d_{JS})$; the d_1 axis is shown on a greatly exaggerated scale, since the three points span only $d_1 \in [0, 0.02]$. The intermediate point y is close to both endpoints on the angular channel ($d_1 \approx 0.01$ on both edges) and asymmetric on the topic channel ($d_2(x, y) = 0.99$ but $d_2(y, z) = 0.01$). This channel imbalance is what produces the violation $D_{GM}(x, z) = 1.000 > 0.911 = D_{GM}(x, y) + D_{GM}(y, z)$.

Proposition 7 (Connection to the Hellinger distance). Let $H(p, q) = (\sum_k (\sqrt{p_k} - \sqrt{q_k})^2)^{1/2}$ be the Hellinger distance and $BC(p, q) = \sum_k \sqrt{p_k q_k}$ the Bhattacharyya coefficient. Then, exactly and at all distances,

$$H^2 = 2(1 - BC) \quad \text{and} \quad H = 2 \sin(d_{FR}/2), \quad (25)$$

i.e., H is the chordal distance subtended by the Fisher–Rao arc under the square-root embedding. Moreover, for all $p, q \in \Delta^{K-1}$,

$$\frac{H}{2\sqrt{\ln 2}} \leq d_{JS} \leq \frac{H}{\sqrt{2 \ln 2}}, \quad (26)$$

with the upper bound tight to first order as $p \rightarrow q$. Consequently, wherever $H < \sqrt{2 \ln 2}$, the topic cost of the exact decomposition (9) satisfies

$$-\log\left(1 - \frac{H}{2\sqrt{\ln 2}}\right) \leq -\log(1 - d_{JS}) \leq -\log\left(1 - \frac{H}{\sqrt{2 \ln 2}}\right), \quad (27)$$

and to first order in d_{FR} , $-\log(1 - d_{JS}) \approx -\log(1 - H/\sqrt{2 \ln 2})$.

Proof. Expanding the square, $H^2 = \sum_k p_k - 2 \sum_k \sqrt{p_k q_k} + \sum_k q_k = 2(1 - BC)$. By the definition (5) of the Bhattacharyya angle, $BC = \cos d_{FR}$, so $H^2 = 2(1 - \cos d_{FR}) = 4 \sin^2(d_{FR}/2)$, proving (25); under $\eta_k = \sqrt{p_k}$ the distributions lie on the unit sphere, where d_{FR} is the arc and H the chord. For (26), the bound $\frac{1}{2}(1 - BC) \leq \text{JSD}_{\text{nat}} \leq 1 - BC$ [38] reads $\frac{1}{4}H^2 \leq \text{JSD}_{\text{nat}} \leq \frac{1}{2}H^2$; dividing by $\ln 2$ and taking square roots gives the claim. Near the diagonal, $\text{JSD}_{\text{nat}} = \frac{1}{2}d_{FR}^2 + O(d_{FR}^4)$ (identity (6), the cubic term vanishing) while $\frac{1}{2}H^2 = \frac{1}{2}d_{FR}^2 + O(d_{FR}^4)$, so the upper bound is tight to first order. The cost sandwich follows

from the monotonicity of $x \mapsto -\log(1-x)$; the lower expression is defined for all $H \leq \sqrt{2}$ since $H/(2\sqrt{\ln 2}) \leq \sqrt{2}/(2\sqrt{\ln 2}) < 1$, the upper for $H < \sqrt{2\ln 2}$. $\square$

Interpretation. Both d_{JS} and $H/\sqrt{2\ln 2}$ are first-order in the Fisher–Rao distance with the same constant, and (26) bounds their finite-distance discrepancy by a factor $\sqrt{2}$. They are not interchangeable as similarity measures, however: $1 - H/\sqrt{2\ln 2}$ becomes negative for near-disjoint supports ($H > \sqrt{2\ln 2}$), so the Hellinger substitute is not $[0, 1]$ -bounded—concretely the boundedness advantage that Section 3.3 cites in favour of d_{JS} . Combining (25) and (26) with $2\sin(d_{FR}/2) \leq d_{FR}$ also gives the pointwise exact bound $d_{JS} \leq d_{FR}/\sqrt{2\ln 2}$: the scatter of Figure 4 can only lie on or below the identity line, as observed.

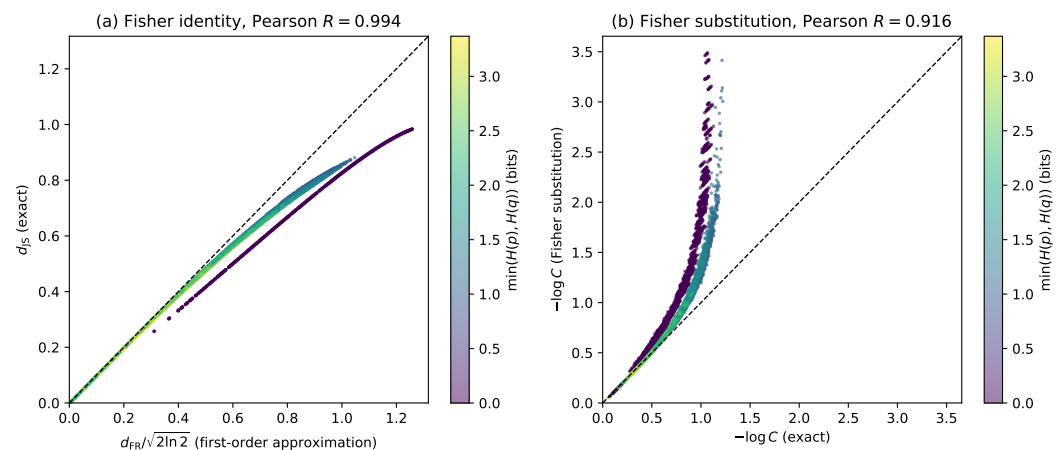


Figure 4. Approximation quality on Cuba, both panels colored by the minimum base-2 Shannon entropy of the two cluster-membership vectors in each pair. **(a)** First-order Fisher metric identity: exact d_{JS} vs. the linear approximation $d_{FR}/\sqrt{2\ln 2}$ on the 80,950 pairs with $C > 10^{-10}$. Pearson $R = 0.994$. The scatter resolves into two branches: an upper branch (yellow/green, high minimum entropy, diffuse distributions) hugging $y = x$ where the first-order identity is tight, and a lower branch (dark, low minimum entropy, at least one near-delta distribution) dipping below $y = x$ as the Fisher–Rao metric tensor degenerates near the simplex boundary (Remark 4). The horizontal hook at the top is where d_{JS} has saturated at its base-2 ceiling of 1. **(b)** Fisher substitution (10): exact angular term with d_{JS} replaced by $d_{FR}/\sqrt{2\ln 2}$, on the 51,342 pairs where the substitution is defined ($d_{FR}/\sqrt{2\ln 2} < 1$). Pearson $R = 0.916$. The same two-branch structure appears here: the diagonal agreement band at the lower left is the upper-branch population (high entropy, faithful substitution), and the vertical tail is the surviving $\approx 34\%$ of lower-branch pairs whose substituted cost is amplified toward the log-singularity by the $-\log(1 - \cdot)$ transform. The dashed line is $y = x$.

Behavior at extremes. If $A = 0$ (antipodal embeddings) or $T = 0$ ($d_{JS} = 1$, maximally different topics), then $C = 0$, so the geometric mean enforces that *both* channels must contribute non-trivially. This “veto” property is desirable for coherence because a transition that is semantically coherent but topically incoherent (or vice versa) should receive zero coherence. In practice, the veto is redundant for unconstrained maximin extraction, because such paths inherently avoid low-weight edges when alternative routes exist. It does, however, provide a hard safety guarantee for extraction algorithms that impose global constraints (the coverage, flow, and cardinality requirements of the narrative maps linear program) which may force otherwise-avoidable edges.

6. Connections to Classical Information Theory

This section connects the composite coherence metric to classical information-theoretic concepts: the data processing inequality, and the identification of each per-channel cost with an exact information-theoretic quantity. These connections are not required by the core

results of Sections 3–5, but they explain why the two channels, despite being structurally coupled (topics are derived from embeddings), capture genuinely different information.

6.1. Multi-Scale Complementarity

The pipeline for computing topic distributions involves successive information-reducing transformations:

$$e_i \in \mathbb{R}^d \xrightarrow{\text{UMAP}} \tilde{e}_i \in \mathbb{R}^p \xrightarrow{\text{HDBSCAN}} \hat{e}_i \in \Delta^{K-1}. \quad (28)$$

Since the topic channel is derived from the embedding channel via a lossy pipeline, statistical independence between A and T is structurally impossible. The relevant property is not independence but *complementarity*: the two channels should capture information at different *scales of abstraction*, from fine-grained semantic direction (A) to coarse-grained thematic membership (T), so that combining them yields a richer similarity than either alone.

Proposition 8 (Information hierarchy). *Let (e_i, e_j) be a pair of document embeddings with arbitrary joint distribution (for instance, a uniformly random pair from the corpus, or the endpoints of a random transition), and let $\tilde{e} = \text{UMAP}(e)$ and $\hat{e} = \text{HDBSCAN}(\tilde{e})$ be the reduced representations and topic distributions, the fitted pipeline being treated as a fixed, measurable, deterministic map (it is fit once on the corpus, with fixed seed, Section 2.2; the randomness refers only to the draw of the pair). Then*

$$I(e_i; e_j) \geq I(\tilde{e}_i; \tilde{e}_j) \geq I(\hat{e}_i; \hat{e}_j). \quad (29)$$

Proof. We use the following consequence of the data processing inequality [45]: if $X' = f(X)$ and $Y' = g(Y)$ are deterministic functions, then $X' - X - Y$ and $X' - Y - Y'$ are Markov chains, so applying the inequality twice,

$$I(X'; Y') \leq I(X'; Y) \leq I(X; Y)$$

(all quantities are well defined in $[0, \infty]$ and the inequalities hold there). Applying this with $(X, Y) = (e_i, e_j)$ and $f = g = \text{UMAP}$ gives $I(\tilde{e}_i; \tilde{e}_j) \leq I(e_i; e_j)$; applying it again with $(X, Y) = (\tilde{e}_i, \tilde{e}_j)$ and $f = g = \text{HDBSCAN}$ gives $I(\hat{e}_i; \hat{e}_j) \leq I(\tilde{e}_i; \tilde{e}_j)$. $\square$

Interpretation. The proposition fixes the *direction* of information flow: the topic channel computed from the K -dimensional clustered representation ($K \ll d$) cannot carry more pairwise information than the angular channel computed from the full d -dimensional embedding—it is a genuine compression. The inequality bounds the *amount* of information retained, not its content; that the discarded information is specifically fine-grained directional structure is an interpretation, while the load-bearing empirical fact is the separately measured non-redundancy of the two channels (the scale-complementarity test, Section 7.4).

Remark 3 (Scale complementarity as a design principle). *Two similarity channels A and T derived from the same underlying representations exhibit scale complementarity when three qualitative conditions hold.*

1. **Non-redundancy:** $\text{NMI}(A, T) \ll 1$, i.e., the channels share a small fraction of their information content despite the structural coupling. In our experiments, $\text{NMI} \approx 0.03$ (well below 0.05) across all corpora with $K \geq 11$ (Section 7.4).
2. **Scale separation:** T is obtained from A 's representation space via a many-to-one mapping (compression) that discards fine-grained structure while preserving coarse-grained structure, so the channels capture different levels of abstraction. This is verified structurally by the $\text{UMAP} \rightarrow \text{HDBSCAN}$ pipeline (Proposition 8).

3. **Composite informativeness:** The composite $C = F(A, T)$ is not a function of A alone, nor of T alone. For $C_{GM} = \sqrt{AT}$ this is immediate: two pairs with equal A but different T receive different C_{GM} .

Scale complementarity is a design principle, not a formal prerequisite: the axiomatic characterization (Theorem 3) and product metric (Theorem 2) hold at any correlation level. It guides practitioners toward pipelines where the composite is genuinely more informative than either channel alone; the scale-complementarity test (Section 7.4) confirms this for the present pipeline.

Condition 1 is verified empirically in Section 7.4, Condition 2 holds structurally because the pipeline is a deterministic compression (Proposition 8), and Condition 3 holds for C_{GM} by the elementary argument above; the joint (A, T) distribution (Figure 5) further shows the corpus pairs populate a genuinely two-dimensional region rather than collapsing onto a curve $T = f(A)$, so the non-redundancy is exercised in practice and not merely possible in principle.

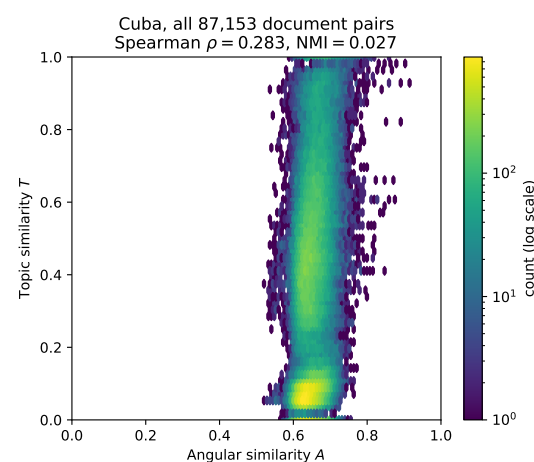


Figure 5. Joint distribution of angular similarity A and topic similarity T across all 87,153 document pairs. The Spearman rank correlation is $\rho = 0.283$ (NMI = 0.027): the channels are structurally coupled but capture different levels of abstraction (scale complementarity).

6.2. Information-Theoretic Content of the Per-Channel Costs

The log-coherence cost decomposition

$$-\log C_{GM} = \frac{1}{2}(-\log A) + \frac{1}{2}(-\log T) \quad (30)$$

is an algebraic identity (Proposition 3) that holds at any correlation level. Each of the two per-channel costs admits an exact information-theoretic identity. Both are classical results; we record them because together they pin down the information-theoretic content of (30), in place of an informal reading of $-\log A$ and $-\log T$ as “rates.”

The angular cost. The angular similarity is exactly a hash-collision probability.

Proposition 9 (Angular similarity as collision probability; 46,47). *Let $u, v \in \mathbb{R}^d \setminus \{0\}$ be two document embeddings separated by angle $\theta \in [0, \pi]$, so $A = 1 - \theta/\pi$. Let $g \in \mathbb{R}^d$ be isotropically random with $\Pr[g = 0] = 0$ (e.g., a standard Gaussian), inducing the one-bit hash $h_g(x) = \text{sign}\langle g, x \rangle$. Then*

$$A = \Pr_g[h_g(u) = h_g(v)], \quad (31)$$

the probability that the uniformly random hyperplane $g^\perp$ through the origin does not separate u and v .

Proof. Conditional on the radius of g , its law is a mixture of uniform measures on spheres, each of which assigns zero mass to any proper subspace; hence, it is almost certain that $\langle g, u \rangle \neq 0 \neq \langle g, v \rangle$ and the hashes take values in $\{\pm 1\}$. If $\theta \in \{0, \pi\}$ the identity is immediate: v is a positive (resp. negative) multiple of u , so the signs almost surely agree (resp. differ), matching $A = 1$ (resp. $A = 0$). Assume $0 < \theta < \pi$, so that $\text{span}(u, v)$ is a plane. The signs $\text{sign}\langle g, u \rangle$ and $\text{sign}\langle g, v \rangle$ depend on g only through its orthogonal projection onto this plane, which is almost surely nonzero and whose direction is uniform on the unit circle of the plane: in-plane rotations, extended by the identity on the orthogonal complement, commute with the projection and preserve the law of g . The hyperplane $g^\perp$ separates u and v exactly when that direction falls in the symmetric difference of the two half-circles $\{w : \langle w, u \rangle > 0\}$ and $\{w : \langle w, v \rangle > 0\}$, a union of two arcs of length θ each, with a total measure of 2θ out of 2π . Hence $\Pr[\text{separate}] = \theta/\pi$ and $\Pr[h_g(u) = h_g(v)] = 1 - \theta/\pi = A$. $\square$

This is the random-hyperplane (SimHash) identity. It gives the angular cost an exact reading: $-\log A$ is the Shannon surprisal of a hash collision—the ideal codeword length in the event that a random hyperplane fails to separate the two documents.

The topic cost. The Jensen–Shannon divergence underlying the topic channel is exactly the mutual information between a sample and the latent label indicating which of the two topic distributions produced it.

Proposition 10 (Topic dissimilarity as mutual information; 24). *Let $p, q \in \Delta^{K-1}$ be two topic distributions (with the convention $0 \log 0 = 0$). Introduce a latent label $Z \sim \text{Bernoulli}(\frac{1}{2})$ and a sample $X \in \{1, \dots, K\}$ drawn from p if $Z = 0$ and from q if $Z = 1$. Then, with base-2 logarithms,*

$$\text{JSD}(p, q) = I(X; Z), \quad (32)$$

the mutual information between the sample and the label, and $0 \leq I(X; Z) \leq 1$ bit, with $I(X; Z) = 0$ iff $p = q$ and $I(X; Z) = 1$ iff p and q have disjoint support.

Proof. The marginal law of X is the mixture $M = \frac{1}{2}p + \frac{1}{2}q$, so $I(X; Z) = H(X) - H(X | Z) = H(M) - \frac{1}{2}H(p) - \frac{1}{2}H(q)$, the uniformly weighted Jensen–Shannon divergence. The bounds follow from $0 \leq I(X; Z) \leq H(Z) = 1$. The lower bound is attained iff $X \perp Z$; since both values of Z have positive probability, this holds iff the two conditional laws coincide, i.e., $p = q$. The upper bound is attained iff $H(Z | X) = 0$, i.e., Z is almost surely determined by X ; the posterior $\Pr[Z = 0 | X = k] = p_k / (p_k + q_k)$ for k in the support of M lies in $\{0, 1\}$ for every such k iff no k has $p_k q_k > 0$, i.e., iff the supports are disjoint. $\square$

Corollary 3 (Topic cost as distinguishability). *On $I(X; Z) \in [0, 1)$, the topic cost satisfies $-\log T = -\log(1 - \sqrt{I(X; Z)})$, a strictly increasing function of $I(X; Z)$: it is an exact, strictly increasing function of the information a single observation carries about which of the two topic distributions generated it. As $I(X; Z) \rightarrow 1$ (disjoint supports), $T \rightarrow 0$ and the cost diverges—the information-theoretic face of the veto property (Section 5).*

Proof. $T = 1 - d_{\text{JS}} = 1 - \sqrt{\text{JSD}} = 1 - \sqrt{I(X; Z)}$ by Proposition 10, so $-\log T = (g \circ f)(I(X; Z))$ with $f(t) = \sqrt{t}$ and $g(s) = -\log(1 - s)$, a composition of the strictly increasing maps $f: [0, 1) \rightarrow [0, 1)$ and $g: [0, 1) \rightarrow [0, \infty)$. $\square$

Reading the decomposition. Equation (30) is thus an identity between three information-theoretic costs: the composite cost is the arithmetic mean of an angular collision-surprisal and a topic cost that is monotone with respect to the mutual information $I(X; Z)$. The arithmetic mean (not an unnormalized sum) is forced by the calibration axiom **A4**: $A = T = a$ must

give the composite cost $-\log a$, fixing the prefactor $\frac{1}{2}$; summation is instead the aggregation across edges of a path, $\sum_e -\log C_e$. That the two identities are of different types—collision probability and mutual information—is not incidental: the angular channel measures geometric agreement of directions and the topic channel distributional distinguishability, the two “levels of abstraction” of scale complementarity (Remark 3). Under scale complementarity (Section 7.4) the costs are also approximately uncorrelated, so they can be diagnostically separated along extracted paths (Section 4.2)—a contingent benefit of the corpus statistics, since (30) holds at any correlation level.

A heuristic rate-distortion reading. The decomposition (30) can also be read informally through a rate-distortion analogy, in which approximately independent channels with logarithmic rate-distortion functions make $(D_A D_T)^{1/2}$ a natural scalar distortion summary. That analogy rests on two non-factual assumptions (independence and a logarithmic rate-distortion form) and is not a derivation; we record it in the Supplementary Materials (Section S6). The identities of this subsection, by contrast, are exact.

6.3. Soft Membership and Topic Resolution

Remark 4 (Soft membership preserves manifold structure). *If $\hat{e}_i \in \Delta_+^{K-1}$ has full support (all $\hat{e}_{ik} > 0$), it lies in the interior of the simplex where the Fisher–Rao metric is well-defined and non-degenerate, T takes values in the full interval $(0, 1]$, and $d_\times$ inherits the full Riemannian structure. By contrast, hard assignment restricts membership vectors to the vertices of Δ^{K-1} , where the metric tensor degenerates ($1/p_k \rightarrow \infty$ as $p_k \rightarrow 0$) and T collapses to a binary same/different signal ($d_{JS} \in \{0, 1\}$) that provides no graded similarity information.*

HDBSCAN’s soft membership assignment [21] produces membership vectors with full support in practice, so the topic channel operates on the interior of the statistical manifold. This is a prerequisite for the product-manifold framework of Section 3: hard assignment would collapse the Fisher–Rao geometry entirely. A component ablation in the Supplementary Materials (Section S4) confirms this empirically: replacing the soft membership vectors with their hard one-hot counterparts collapses the rank correlation between C_{GM} and the product metric $d_\times$ from 0.999 to 0.65, and replacing the Jensen–Shannon distance with the divergence raises the triangle-inequality violation rate from 0% to 7.5%.

7. Experimental Validation

We validate the theoretical results with seven experiments on real-world corpora, with further supporting analyses in the Supplementary Materials (Sections S1–S8).

7.1. Datasets

The primary corpus is a subset of the Cuban news corpus from German et al. [7]: 418 documents with 1536-dimensional embeddings from a language model (GPT-4). Soft cluster membership distributions are computed via the UMAP→HDBSCAN pipeline described in Section 2.2. This yields K topic clusters with per-document membership vectors $\hat{e}_i \in \Delta^{K-1}$. Experiments 1 to 3 use this corpus. Experiment 4 (Section 7.5) extends the validation to three additional corpora from the narrative maps [2] and narrative trails [7] repositories, covering different domains, sizes, and embedding models. Experiment 7 (Section 7.8) additionally draws on the Wikispeedia navigation corpus [11], which supplies human storyline data absent from the other corpora.

7.2. Experiment 1: Approximation Quality

Proposition 1 establishes the exact decomposition $-\log C = \frac{1}{2}[-\log(1 - d_{\text{ang}}/\pi) - \log(1 - d_{\text{JS}})]$, where the topic term connects to the Fisher–Rao distance via Remark 1. We test the Fisher identity and the resulting coherence approximation on all $\binom{418}{2} = 87,153$ document pairs, restricted to the 80,950 pairs with $C > 10^{-10}$ (i.e., neither channel collapsed to zero). The excluded pairs correspond to documents in disjoint topic clusters where d_{JS} saturates at 1 and d_{FR} saturates at $\pi/2$. These are boundary artifacts of the metric range, and neither side of the approximation is finite there.

Figure 4a directly tests the first-order Fisher metric identity $d_{\text{JS}} \approx d_{\text{FR}}/\sqrt{2\ln 2}$ (Equation (8)). The Pearson correlation is $R = 0.994$, confirming a strong linear relationship. The scatter resolves into two visually distinct branches, which are separated by the minimum Shannon entropy of the two cluster-membership vectors in each pair (the Panel (a) colormap). The *upper branch*, consisting of pairs where both p and q are diffuse (high minimum entropy, mass spread across most of the $K = 11$ clusters), hugs $y = x$ closely: for these pairs, both distributions sit in the interior of the simplex where the Fisher–Rao metric tensor is well-conditioned, and the Taylor expansion of JSD around $p = q$ is accurate. The *lower branch*, where at least one of p, q peaks on a small number of clusters (low minimum entropy), dips systematically below $y = x$: here the distributions approach the simplex vertices, the Fisher–Rao metric tensor begins to degenerate ($1/p_k \rightarrow \infty$ as $p_k \rightarrow 0$; Remark 4), and the linear approximation $d_{\text{FR}}/\sqrt{2\ln 2}$ overshoots the truncated range of d_{JS} . The horizontal hook at the top, where d_{JS} has saturated at 1 while $d_{\text{FR}}/\sqrt{2\ln 2}$ continues to ≈ 1.33 , is the asymptotic limit of the lower branch: pairs whose topic distributions approach disjoint-support deltas. The two-branch structure is thus the empirical signature of the simplex-boundary degeneration the theory warns about, not an artifact. That the entire scatter lies on or below the diagonal—the upper branch tight against $y = x$, the lower branch below it, none above—is moreover mandated pointwise rather than merely observed: the first-order identity is in truth the one-sided bound $d_{\text{JS}} \leq d_{\text{FR}}/\sqrt{2\ln 2}$, exact at all distances (Proposition 7). It also explains why the overall Pearson correlation is high despite a visible bias, since the branches rejoin at both endpoints (origin and saturation) and only separate in the middle of the range.

Figure 4b evaluates the Fisher substitution (10), which replaces d_{JS} with $d_{\text{FR}}/\sqrt{2\ln 2}$ in the exact decomposition while keeping the angular term exact. The substitution is only mathematically defined when $d_{\text{FR}}/\sqrt{2\ln 2} < 1$ (otherwise $-\log(1 - d_{\text{FR}}/\sqrt{2\ln 2})$ is undefined), so we restrict Panel (b) to the 51,342 pairs that satisfy this condition in addition to the $C > 10^{-10}$ filter. The filter retains all 36,242 upper-branch pairs of Panel (a) but removes about two thirds of the lower-branch pairs, specifically the most extreme ones, where at least one cluster distribution is essentially a delta and d_{FR} is near $\pi/2$. The remaining 15,100 lower-branch pairs form the vertical tail seen at the upper left of Panel (b), where the $-\log(1 - \cdot)$ transform amplifies the Panel (a) bias toward the log-singularity as $d_{\text{FR}}/\sqrt{2\ln 2}$ approaches 1. The diagonal agreement band in the lower left is the upper-branch population where the substitution is faithful. The overall Pearson correlation on Panel (b) is $R = 0.916$, lower than Panel (a) because the nonlinear transform stretches the lower-branch residual vertically. The exact decomposition (9) remains perfect at all distances and is what the rest of the paper uses. Panel (b) is reported as a diagnostic of the first-order identity, not as an alternative computational path.

The exact decomposition (9) holds to machine precision ($<10^{-15}$), as it is an algebraic identity rather than an approximation. The infinitesimal expansion (11) requires d_{ang}/π and d_{JS} to be small, but in our corpus, typical angular distances are $d_{\text{ang}}/\pi \approx 0.35$, well outside this regime. The practical strength of the exact decomposition is that it provides the additive cost structure (angular cost plus topic cost) at *all* distances, not just infinitesimally.

7.3. Experiment 2: Combinator Comparison

We compare six combinators (GM, AM, HM, Quad, Min, Max) using the product-metric framework from Theorem 2. Two evaluation criteria align the experiment with the theoretical foundations: (a) triangle inequality violation rate for the bounded dissimilarity $1 - F$, and (b) rank correlation of $1 - F$ with the product metric $d_{\times}$.

Product metric sanity check. The product metric $d_{\times}$ (defined in Theorem 2) is a proper metric by construction (the normalized ℓ^2 -product of the angular and Jensen–Shannon distances, with $d_{\times} \in [0, 1]$), so it cannot violate the triangle inequality. The 0.00% violation rate it records on 10^5 random triplets is accordingly not a test of the theorem but a check on the implementation—confirming that the computed $d_{\times}$ matches the construction, not validating a result that is already exact.

Bounded triangle inequality. We measure violations on the bounded dissimilarity $D_F = 1 - F(A_{ij}, T_{ij}) \in [0, 1]$, which is defined for all triplets (the alternative $-\log F$ diverges at $F = 0$). Proposition 6 establishes that $1 - \text{AM}$ and $1 - \text{min}$ are proper metrics by construction (L^1 and L^∞ products, respectively), while $1 - \text{GM}$ is not a metric in general, though no violations are observed empirically.

Rank correlation with $d_{\times}$. For each combinator, we compute the Spearman rank correlation between $1 - F$ and $d_{\times}$ across all $\binom{n}{2}$ document pairs. This measures how faithfully each combinator’s ordering of document pairs matches the product-manifold geometry recorded in Theorem 2.

Table 2 reads as a contrast between the combinators whose metric statuses are settled and those whose statuses are not. AM and Min are proper metrics (Proposition 6), so their 0.00% violation rates are required rather than measured findings, and confirm only the implementation. The informative entries are the combinators with no such guarantee: GM shows 0.00% empirically despite not being a metric in general because the pathological configurations that violate the triangle inequality (extreme channel imbalance at an intermediate point) do not arise in this corpus. Quad shows 8.21% violations because $1 - C_{\text{Quad}} = d_{\times}^2$, and squaring a metric does not preserve the triangle inequality. Max shows 0.53% violations, reflecting its failure to penalize weakness in either channel.

Table 2. Combinator comparison on Cuba. Triangle inequality violations for the bounded dissimilarity $1 - F$ from 10^5 sampled triplets, and Spearman rank correlation with the product metric $d_{\times}$ over all $\binom{n}{2}$ pairs. Wilson 95% binomial intervals for the violation rates are within $\pm 0.06\%$ of the reported rates at the 0–1% regime and $\pm 0.18\%$ at the 8% regime (Quad); Spearman bootstrap CIs ($B = 200$, document-level resampling) are within ± 0.001 of the reported $\rho(d_{\times})$ values for GM, AM, HM, Quad, and Min and ± 0.01 for Max.

Combinator	Viol. (%)	$\rho(d_{\times})$
Geometric mean (GM)	0.00	0.999
Arithmetic mean (AM)	0.00	0.997
Harmonic mean (HM)	0.07	0.998
Metric-squared (Quad)	8.21	1.000
Minimum	0.00	0.995
Maximum	0.53	0.533

The rank correlations with $d_{\times}$ show that GM ($\rho = 0.999$), HM (0.998), AM (0.997), and Min (0.995) all track the product-manifold geometry closely, and the empirical gap among these four is narrow ($\Delta\rho < 0.005$). Quad achieves $\rho = 1.000$ by construction: since $1 - C_{\text{Quad}} = d_{\times}^2$ is a monotone transformation of $d_{\times}$, Spearman rank correlation is perfect. This confirms that Quad is functionally equivalent to the product metric for ranking purposes (Section 4.1). Only the maximum combinator ($\rho = 0.533$) is clearly inferior, losing substantial geometric information by ignoring the weaker channel. The

narrow gap among GM, HM, AM, and Min is itself informative: it shows that the product-manifold geometry is robust to the choice of combinator, and that the geometric structure of the document space, not the combinator, is the primary determinant of pair ordering. What distinguishes GM from the other three is not rank correlation but *boundary behavior*. GM enforces $C = 0$ whenever either channel is zero (the “veto” property), blocking transitions between topically unrelated events regardless of semantic similarity. AM permits nonzero coherence when one channel is zero and thus lacks the veto, whereas HM and Min, like GM, enforce it; HM, however, collapses pathologically toward zero for asymmetric inputs, and Min ignores the stronger channel entirely, discarding information. The GM thus occupies a unique position: it is the only combinator among the five that (a) enforces the veto property, (b) uses information from both channels, and (c) satisfies the axiomatic characterization of Theorem 3. A veto analysis on both the full (dense) graph and MST-pruned graph confirms that unconstrained maximin extraction never traverses low- T edges (0/100 paths on either graph for both corpora), making the veto redundant for that algorithm class; under the cardinality-constrained LP extraction regime it instead becomes binding (Supplementary Materials, Section S9).

GM vs. $d_{\times}$ for path extraction. Since $d_{\times} \in [0, 1]$, we define the coherence-space counterpart $C_{d_{\times}} = 1 - d_{\times} \in [0, 1]$, so that all six methods (five combinators plus $d_{\times}$) produce coherence matrices in $[0, 1]$ and use the same maximum-capacity-path algorithm. Maximin on $C_{d_{\times}}$ is equivalent to minimax on $d_{\times}$ because maximin extraction is invariant under monotone transformations of the edge weights (Corollary 2). The downstream evaluation in Section 7.6 compares all methods head-to-head using this unified framework.

7.4. Experiment 3: Scale Complementarity

Scale complementarity (Remark 3) posits that A and T capture different levels of abstraction. Since T is derived from the same embeddings as A via lossy compression, true statistical independence is structurally impossible, and a positive correlation is guaranteed. The question is whether the channels are nonetheless *non-redundant*: does the compression discard enough fine-grained information that T captures something genuinely different from A ? We test this by computing A_{ij} and T_{ij} for all $\binom{418}{2} = 87,153$ document pairs.

Figure 5 shows the joint distribution. The Spearman rank correlation is $\rho = 0.283$ ($p < 10^{-100}$), a weak positive association, consistent with the structural coupling but far from redundancy. Since Spearman captures only monotonic dependence, we also estimate the mutual information $I(A; T)$ via histogram binning (30 bins per axis). The normalized mutual information $\text{NMI} = 2I(A; T) / [H(A) + H(T)] = 0.027$ captures arbitrary nonlinear dependencies, confirming that A and T share less than 3% of their information content, even accounting for the nonlinear UMAP+HDBSCAN pipeline. Together with the multi-scale structure verified by the data processing inequality (Proposition 8), these results confirm scale complementarity (Remark 3): the channels are non-redundant (condition 1) and capture different levels of abstraction (condition 2).

Scale complementarity validates the pipeline design by confirming that the UMAP $\rightarrow$ HDBSCAN compression produces a topic channel distinct from the angular channel.

7.5. Experiment 4: Cross-Corpus Validation

Experiments 1 to 3 use a single news corpus. To verify that the information-geometric properties are not corpus-specific, we replicate the key findings across four corpora spanning different domains, sizes, and embedding models, and then test sensitivity to embedding model choice on a fixed corpus. The four corpora are: *Cuba* (418 news articles on Cuban politics; GPT-4 embeddings, $d = 1536$; the primary corpus), *COVID* (40 news articles on the COVID-19 pandemic; MiniLM embeddings, $d = 384$; from narrative maps [2]),

VisPub (3549 visualization research papers with abstracts; GPT-4 embeddings, $d = 1536$, for the metric-level analysis, and SPECTER2 embeddings, $d = 768$, over all 3620 papers for downstream evaluation; from narrative trails [7]), and *AMiner* (6000 ML/AI research papers; GPT-4 embeddings, $d = 1536$; from narrative trails [7]).

For each corpus, we compute: (1) the Fisher identity correlation $R(d_{JS}, d_{FR} / \sqrt{2 \ln 2})$ on pairs with $C > 10^{-10}$ (matching the filter of Experiment 1), (2) the GM rank correlation $\rho(1 - C_{GM}, d_{\times})$ over all pairs, and (3) the channel complementarity $\rho(A, T)$ (lower values indicate greater non-redundancy). We also test triangle inequality violations for $1 - C_{GM}$. Table 3 and Figure 6 summarize the results.

Table 3. Cross-corpus validation. Fisher identity $R \geq 0.99$, GM rank correlation $\rho \geq 0.989$, and channel correlation low (indicating scale complementarity). Bracketed values are 95% bootstrap confidence intervals from $B = 200$ document-level resampling iterations on the channel correlation $\rho_{A,T}$, excluding the same-document pairs created by resampling with replacement. The bootstrap half-widths for R_{Fisher} and ρ_{GM} are all within ± 0.01 of the point estimates (typically ± 0.002) and are omitted from the table for clarity [†].

Corpus	n	K	R_{Fisher}	ρ_{GM}	$\rho_{A,T}$ (95% CI)
Cuba	418	11	0.994	0.999	0.283 [0.231, 0.337]
COVID [†]	40	2	0.998	0.993	−0.014 [−0.140, 0.154]
VisPub	3549	113	0.997	0.992	0.234 [0.214, 0.255]
AMiner	6000	163	0.997	0.990	0.153 [0.140, 0.167]

[†] The COVID corpus ($n = 40$, $K = 2$) is below the framework's recommended operating regime ($K \geq 4$); it is included to illustrate boundary behavior. The wide $\rho_{A,T}$ confidence interval on COVID reflects the small sample size; the other corpora yield half-widths ≤ 0.06 . See text for discussion of its 7.6% triangle inequality violation rate.

The Fisher identity holds consistently ($R \geq 0.99$) across all corpora, indicating that the JSD–Fisher–Rao connection is not an artifact of a particular corpus or embedding model. The GM rank correlation with $d_{\times}$ is $\rho \geq 0.989$ in all cases (the smallest, 0.990 to three decimals, is AMiner's), consistent with the product-metric interpretation. Channel correlation remains low across all corpora ($|\rho(A, T)| \leq 0.283$ with 95% bootstrap intervals bracketed in Table 3), consistent with scale complementarity (Remark 3): despite the structural coupling, the channels capture different levels of abstraction. The COVID corpus shows a near-zero correlation ($\rho = -0.014$, CI $[-0.14, 0.15]$), consistent with its small size ($n = 40$) and minimal cluster structure ($K = 2$). With only two clusters, the coarse-grained thematic dimension is so impoverished that almost no monotonic structure couples it to the fine-grained angular dimension, and the wide confidence interval reflects the limited statistical leverage of $n = 40$ documents.

The COVID corpus ($K = 2$) exhibits 7.6% triangle inequality violations for the bounded dissimilarity $1 - C_{GM}$, while all other corpora ($K \geq 11$) show 0%. With only two clusters, membership vectors concentrate near the simplex vertices, producing $T \approx 0$ for documents in opposite clusters, exactly the pathological channel imbalance identified in the counterexample of Proposition 6. A hyperparameter sweep over HDBSCAN configurations on the COVID corpus confirms a monotonic relationship, with the violation rate dropping from 7.6% at $K = 2$ to $<0.5\%$ at $K = 4$ and $<0.1\%$ at $K \geq 5$. For corpora with sufficient cluster structure ($K \geq 4$), the empirical metric properties hold robustly.

Embedding model sensitivity. To disentangle the effect of the embedding model from the corpus, we re-embed the Cuba and VisPub corpora with two further models (MiniLM-L6, $d = 384$; MPNet, $d = 768$) and recompute the full UMAP→HDBSCAN→coherence pipeline. The metric-level properties are stable: GM rank correlation with $d_{\times}$ stays at $\rho \geq 0.99$ and triangle-inequality violations at 0% across all six corpus-embedding com-

binations, while the channel correlation $\rho_{A,T}$ varies within $[0.23, 0.29]$, remaining in the low-correlation regime expected by scale complementarity. The full results are in the Supplementary Materials (Section S3).

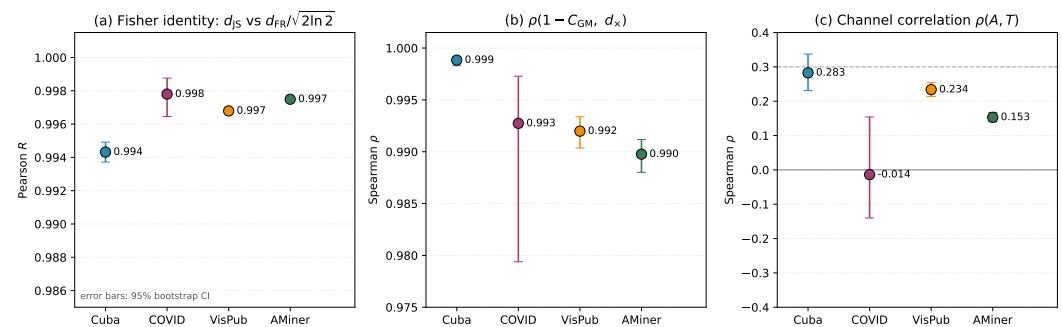


Figure 6. Cross-corpus validation across four corpora: point estimates with 95% bootstrap confidence intervals ($B = 200$ document-level resamples) as error bars. Panels (a,b) use zoomed ρ/R axes; the markers are points (not bars), so the zoom does not exaggerate the sub-percent spread. (a) Fisher identity holds consistently ($R \geq 0.99$). (b) GM rank correlation with d_x satisfies $\rho \geq 0.989$. (c) Channel correlation remains low, confirming scale complementarity.

7.6. Experiment 5: Downstream Consistency Check

Experiments 1 to 4 validate the information-geometric framework; the Supplementary Materials extend the validation to alternative topic models (Section S1) and cluster-count sensitivity (Section S2). Here we perform a *downstream consistency check* on the axiomatic recommendation of Section 4: we verify that the theoretically prescribed GM is not empirically dominated by any alternative combinator or single-channel baseline on the task the metric is designed for, following the narrative trails framework [7].

Setup. We evaluate on two corpora: the Cuba news corpus (418 documents) and the VisPub corpus (3620 papers). For VisPub, we use SPECTER2 embeddings [48,49] ($d = 768$) rather than the GPT-4 embeddings used in the metric-level experiments; SPECTER2 is trained on citation graphs, producing paths that follow research lineages rather than generic semantic similarity. The metric-level experiments use a 3549-paper subset—the papers carry abstracts, which the text-based GPT-4 embedding requires—whereas SPECTER2 represents all 3620 papers, including 71 abstract-less entries. Both corpora use a unified UMAP projection: both the angular similarity channel (A) and the topic similarity channel (T , via HDBSCAN clustering) are derived from the same UMAP(k^*) embedding space, where k^* is selected automatically via angular similarity saturation ($<10\%$ relative change)—the smallest dimension at which the angular structure stops changing. Deriving both channels from a single projection keeps them commensurable. Note that this configuration preserves the metric-level guarantees established for the raw-embedding configuration in Sections 3–5. On both downstream corpora the saturation criterion selects $k^* = 5$, and the product metric d_x (Theorem 2) remains free of triangle-inequality violations under triplet sampling; C_{GM} tracks d_x at Spearman $\rho = 0.98$ (Cuba) and 0.99 (VisPub), and the angular and topic channels remain complementary (Pearson correlation 0.26 and 0.23). These match the raw-embedding configuration, in which C_{GM} tracks d_x at $\rho = 0.999$ (Experiment 2, Section 7.3) and the angular and topic channels show a Pearson correlation of 0.29 on Cuba (the Pearson counterpart of the Spearman $\rho_{A,T} = 0.283$ of Experiment 3, Section 7.4). The shared UMAP coordinates are mean-centered before the angular channel is computed, placing the corpus centroid at the origin so that A measures relative direction among documents; because UMAP output sits at an arbitrary off-origin offset, omitting this step collapses the angular similarities toward 1. Every other angular channel in this paper—the metric-level Experiments 1–4 and the bottleneck-gap

Experiment 7—is instead computed from the raw embeddings, where the UMAP step feeds only the topic clustering and the reported relational summaries are insensitive to centering. Both corpora are temporally ordered, so we build *directed* sparse coherence graphs using a maximum spanning arborescence (Edmonds’ algorithm [50]) rooted at the earliest event, rather than the undirected MST used in the original narrative trails pipeline. For each of the six combinators (GM, AM, HM, Quad, Min, Max), we build a sparse graph using the arborescence-derived critical weight. Each combinator is tested with and without Dirichlet smoothing for channel balance (14 method variants total), plus two single-channel baselines (*A*-only and *T*-only).

For each corpus, we select 30 source-target pairs (source in the first temporal third, target in the last), requiring that *all methods* produce a valid path of length ≥ 3 for the same endpoints. This shared-endpoint design enables paired statistical comparisons. All methods use the same maximum-capacity-path (maximin) algorithm from narrative trails.

Structural pre-filtering. On Cuba, GM, AM, HM, and Quad form a single structural equivalence class ($>70\%$ path identity or $>85\%$ Jaccard node overlap), consistent with the narrow rank-correlation gap in Experiment 2. We use GM as the representative of this class and evaluate in three stages, with path-level deduplication across methods that produce identical paths for a given endpoint pair:

1. **Stage 1 (Balance):** GM vs. GM_{bal} (Dirichlet-smoothed) to determine whether channel balancing affects sequence quality.
2. **Stage 2 (Combinator):** GM, Quad, Min, Max (using the balance variant carried forward from Stage 1) to compare structurally distinct combinator families.
3. **Stage 3 (Channels):** The Stage 2 winner vs. single-channel controls (*A*-only, *T*-only) to test whether combining channels improves sequence quality over either alone.

Path-level deduplication further reduces cost, since when two methods produce *identical* paths for a given endpoint pair, the storyline is judged only once and the result is shared. This yields 458 (Cuba) and 470 (VisPub) LLM judge calls, roughly 35–45% fewer than judging all method variants on every endpoint pair.

Evaluation. Each extracted sequence is scored independently by two frontier LLM judges, GPT-5.4 and Claude Opus 4.6, using the multi-judge protocol of Keith [51]. The prompt presents the storyline in domain-agnostic terms (“document sequence coherence”), asks the judge whether each document follows logically from the previous one and whether the overall sequence forms a meaningful progression from source to target, and returns a single holistic coherence score on a 1 to 100 scale with a one- or two-sentence justification. The 1 to 100 scale, the absence of anchoring phrases, the single holistic score, and the brief justification are the prompt configurations that [51] finds to maximize discriminative power while controlling for the positivity bias of longer chain-of-thought protocols. Combined with the paired design (shared endpoints), this yields a score matrix of shape $30 \times k$ (k methods per stage) suitable for repeated-measures analysis.

Inter-judge agreement. The two judges show moderate rank agreement: on Cuba, Pearson $r = 0.57$ to 0.58 and Spearman $\rho = 0.59$ to 0.61 across the three stages; on VisPub, $r = 0.28$ to 0.44 and $\rho = 0.23$ to 0.49 . The Krippendorff α at the interval level is positive on Cuba ($\alpha \approx 0.20$) and negative on VisPub ($\alpha \approx -0.76$), but this is driven by a systematic offset between the two judges’ score levels on VisPub (Claude $\mu \approx 20$ versus GPT $\mu \approx 55$) rather than by lack of rank agreement, since the Spearman correlation remains positive. For the paired Friedman/Wilcoxon analysis below, the relevant signal is within-pair rank, not absolute level, so the offset does not affect the consistency-check conclusion; we treat the systematic offset as a known LLM-judge calibration effect and revisit it in Section 8.4.

Statistical analysis. Since all methods share the same endpoints, the comparison is a paired (within-subjects) design. We apply the Friedman test [52], the nonparametric

analogue of repeated-measures ANOVA, with the Nemenyi post hoc test [53] for pairwise comparisons. For two-method stages, we use the Wilcoxon signed-rank test. The Nemenyi critical difference is computed with a Bonferroni-normal approximation of the Studentized-range statistic; this is mildly conservative relative to the exact critical value (the critical difference is overstated by roughly 3%), which does not affect any conclusion below since no pairwise comparison is significant even under the slightly larger critical difference.

Results. Table 4 summarizes each method’s theoretical profile.

Table 4. Theoretical properties of each coherence model. Properties: **Veto** = $C = 0$ when either channel is zero; **Metric** = $1 - C$ satisfies triangle inequality; **Log-add.** = $-\log C$ decomposes additively; **Axioms** = satisfies all four axioms of Theorem 3; $\rho(d_{\times})$ = rank correlation with product metric.

Method	Veto	Metric	Log-Add.	Axioms	$\rho(d_{\times})$
GM	✓	empirical	✓	✓	0.999
AM	—	✓	—	—	0.997
HM	✓	—	—	—	0.998
Quad	—	—	—	—	1.000
Min	✓	✓	—	—	0.995
Max	—	—	—	—	0.533

Stage 1: Balance. On both corpora, the balanced variant (Dirichlet-smoothed cluster probabilities with UMAP dimensionality reduction) ties with or narrowly exceeds the unbalanced variant: Cuba mean scores 83.18 (GM_{bal}) vs. 83.15 (GM); VisPub 37.77 (GM_{bal}) vs. 37.38 (GM). Wilcoxon signed-rank tests yield $p = 0.71$ (Cuba) and $p = 0.53$ (VisPub), so the balance variant has at most a negligible effect. We carry the balanced configuration forward as the default (tie-break) choice in subsequent stages. One subtlety affects the notation: Dirichlet smoothing balances the two channels only when they are not already balanced. For combinators whose channels already satisfy the balance condition ($\gamma \approx 1$), the smoothing step is a no-op and the base combinator is the balanced configuration. This is the case for Min and Max on VisPub, where no smoothing is applied; we therefore write Min and Max (without the “bal” subscript) for VisPub and GM_{bal} , $Quad_{bal}$, Min_{bal} , Max_{bal} for Cuba, where all four receive smoothing ($\alpha \approx 0.016$ to 0.022).

Stage 2: Combinator comparison. Table 5 reports the Friedman test results. On neither corpus do the four combinators differ significantly: Cuba $\chi^2 = 0.81$, $p = 0.846$ (all pairwise Nemenyi $p = 1.0$), VisPub $\chi^2 = 6.26$, $p = 0.0996$ (no pairwise comparison significant after Nemenyi correction). $Quad_{bal}$ ranks first on Cuba (mean rank 2.40) with GM_{bal} indistinguishable at 2.43. On VisPub, Max ranks first (2.02) but no pairwise difference is significant.

The path-level overlap data explain these scores. On Cuba, GM_{bal} and $Quad_{bal}$ share 60% path identity and 86% Jaccard node overlap, producing nearly identical narratives, while Max_{bal} has 0% path identity with every other method. The functional equivalence of GM_{bal} and $Quad_{bal}$ is thus structural, as identical paths receive identical scores, and the 0.03 rank gap between them is within sampling noise. The more informative finding is that Min_{bal} and Max_{bal} , which produce structurally *different* narratives (Jaccard ≤ 0.30), still score equivalently. On VisPub, $GM_{bal}/Quad_{bal}$ overlap is high (Jaccard 0.82, identity 50%), and Max remains a complete structural outlier (0% identity with all methods). Despite this structural divergence, no significant quality differences emerge on either corpus.

Table 5. Stage 2: Combinator comparison (30 endpoint pairs per corpus). Mean scores (1 to 100 scale), Friedman mean ranks, and Friedman test statistics. Nemenyi critical difference $CD = 0.879$ at $\alpha = 0.05$ (computed with the Bonferroni-normal approximation of the Studentized-range statistic; see text). Each method uses the balance configuration carried forward from Stage 1: on Cuba all four combinators are Dirichlet-smoothed, while on VisPub GM and Quad are smoothed and Min and Max are already channel-balanced and use the base combinator.

Combinator	Cuba		VisPub	
	Mean	Rank	Mean	Rank
GM	83.4	2.43	38.2	2.68
Quad	83.5	2.40	37.6	2.72
Min	82.4	2.52	38.2	2.58
Max	83.2	2.65	40.1	2.02
Friedman χ^2	0.81		6.26	
Friedman p	0.846		0.0996	

Method column gives the combinator family; the balance configuration is corpus-dependent (Cuba: all Dirichlet-smoothed; VisPub: GM and Quad smoothed, Min and Max unsmoothed because already channel-balanced).

Stage 3: Single-channel controls. The Stage 2 winner ($Quad_{bal}$ on Cuba, Max on VisPub) is compared against single-channel baselines (A -only, T -only). On Cuba, $Quad_{bal}$ achieves the best mean rank (1.82), followed by A -only (2.07) and T -only (2.12), but the Friedman test is not significant ($\chi^2 = 1.69$, $p = 0.429$, all pairwise Nemenyi $p \geq 0.74$). On VisPub, A -only achieves the best mean rank (1.83), followed by Max (1.98) and T -only (2.18), with the differences again non-significant ($\chi^2 = 1.96$, $p = 0.374$, all pairwise Nemenyi $p \geq 0.53$). Unlike Stage 2, Stage 3 compares structurally *distinct* narratives: on Cuba, $Quad_{bal}$ shares only a 3% path identity and 32% Jaccard overlap with A -only, and a 7% identity/43% Jaccard with T -only; on VisPub, Max shares a 13% identity/28% Jaccard with A -only and 0% identity/16% Jaccard with T -only. The two single-channel controls share 0% identity with each other on both corpora. The combined metric therefore does not significantly outperform either single channel on holistic quality as measured by LLM judges, but its value lies in structural properties that a scalar quality score cannot capture, namely diagnostic decomposability (Proposition 3), the veto property (Table 4), and the principled reference metric $d_{\times}$.

The consistency check passes. Across all three stages, GM_{bal} is never significantly dominated by any alternative combinator or single-channel baseline (no stage significant; the smallest Friedman p is 0.0996, on VisPub). On Cuba it is statistically indistinguishable from $Quad_{bal}$ (rank 2.43 vs. 2.40), with which it is in the same structural equivalence class (60% path identity, 86% Jaccard node overlap). On VisPub the same holds relative to Max. Had GM significantly underperformed, the axiomatic recommendation of Section 4 would be called into question. The observed parity instead confirms that its distinctions from the other compensability-spectrum members (Section 4.1) are theoretical (metric structure, log-additivity, veto behavior) and structural (the bottleneck-gap profile of Experiment 7) rather than a matter of holistic downstream quality, and that the theoretically prescribed GM is safe to adopt in practice.

7.7. Experiment 6: Robustness to Embedding Perturbation

We test how the metric-level properties degrade when the document embeddings are perturbed by isotropic Gaussian noise. For each noise level $\sigma \in \{0.01, 0.05, 0.10\}$ (relative to the per-coordinate embedding standard deviation), we add fresh noise to the GPT-4 Cuba embeddings, run the full coherence pipeline, and compare against the clean pipeline along three axes: the metric-level correlations (R_{Fisher} , ρ_{GM} , $\rho_{A,T}$), the triangle-inequality

violation rate for $1 - C_{GM}$, and the rank stability $\rho_{stab} = \rho_{Spearman}(C_{GM}^{noisy}, C_{GM}^{clean})$ over all pairs. Three random seeds are averaged.

We report two modes, in line with the discussion of pipeline coupling in Section 6.1. In the *fixed-K* mode, only the angular channel is perturbed: we hold the HDBSCAN cluster labels fixed at the clean pipeline’s output, so the topic channel is unaffected and only A changes under the noise. This cleanly isolates the robustness of the metric structure itself. In the *floating-K* mode, the noise is fed through the full UMAP→HDBSCAN→coherence pipeline, so both the cluster count K and the membership vectors are allowed to shift; this is the more honest end-to-end robustness number, but it conflates metric-level effects with clustering instability. Table 6 summarizes the body-level results; the per-seed breakdown is in the Supplementary Materials (Section S5).

Table 6. Embedding perturbation results on Cuba ($n = 418$, GPT-4 embeddings, clean $K = 11$). Each cell is the mean over three seeds; “stab” is the Spearman correlation of the C_{GM} pair ranking under noise against the clean ranking. The fixed- K block holds HDBSCAN output fixed at the clean labels; the floating- K block re-runs the full pipeline.

Mode/ σ	K	R_{Fisher}	ρ_{GM}	$\rho_{A,T}$	stab C_{GM}	tri. viol. (%)
clean (reference)	11	0.994	0.999	0.283	—	0.00
fixed- K , $\sigma = 0.01$	11	0.994	0.999	0.283	1.000	0.00
fixed- K , $\sigma = 0.05$	11	0.994	0.999	0.283	1.000	0.00
fixed- K , $\sigma = 0.10$	11	0.994	0.999	0.283	1.000	0.00
floating- K , $\sigma = 0.01$	11.0	0.996	0.984	0.223	0.513	0.18
floating- K , $\sigma = 0.05$	10.0	0.996	0.987	0.197	0.488	0.19
floating- K , $\sigma = 0.10$	14.3	0.995	0.998	0.305	0.610	0.00

Takeaway. Under fixed- K , the metric-level structure is exactly preserved at the rank level ($\rho_{stab} = 1.000$ at all three noise levels and at the pair-Spearman precision of 10^{-4}): the angular channel is robust to embedding noise of up to 10% of per-coordinate standard deviation, and the channel correlation $\rho_{A,T}$, the Fisher identity correlation R_{Fisher} , and the triangle-inequality violation rate of $1 - C_{GM}$ are all unchanged. Under floating- K , the cluster count itself becomes the dominant source of variability (K takes values $\{2, 11–16\}$ across seeds and noise levels; the standard deviation of K across seeds at $\sigma = 0.01$ is 7.8), and ρ_{stab} drops to roughly 0.5 as documents are reassigned to different clusters. Even so, the within-run metric-level summaries (R_{Fisher} , ρ_{GM} , $\rho_{A,T}$, triangle violations) remain close to their clean values: the rankings change at the level of which documents are clustered together, but the geometric structure that the framework rests on is preserved at the same statistical level as in the clean pipeline.

The contrast between the two modes is informative. It shows that the robustness limit at small σ in the floating- K mode comes from clustering instability rather than from any fragility of the metric construction itself. This is consistent with the discussion in Section 7.5: the cluster count is the most impactful pipeline parameter, and any framework that relies on a soft-clustering pipeline inherits the variability of that pipeline as a downstream sensitivity. The metric-level claims of the present paper (Theorems 2 and 3) are conditional on a fixed clustering, and the fixed- K row of Table 6 verifies that, conditional on a fixed clustering, those claims are robust to embedding noise at the magnitudes tested.

7.8. Experiment 7: Bottleneck-Gap Profile Across Five Corpora

Experiment 7 tests Proposition 5 directly: we sweep the compensability exponent α of M_α and measure the bottleneck-gap profile $g(\alpha)$ of equation (21) between coherent and incoherent storylines on five corpora.

Setup. On Wikipedia [11]—3928 Wikipedia articles with 10,832 human navigation paths—the coherent population is comprised of human paths and the incoherent one of length-matched random path between the same endpoints. On the four narrative corpora (Cuba, COVID, VisPub, AMiner), which carry no human navigation data, the coherent population is comprised of the maximin storylines extracted as a temporal-DAG widest path (a *narrative trail* [7]) and the incoherent one a length-matched random chronological sequence between the same endpoints. For each storyline we take its bottleneck edge—the minimum-coherence edge—and score it under M_α across a grid of α ; D_H and D_R are the empirical laws of these bottleneck edges. From the $\alpha = 0$ bottleneck edges we also evaluate the leading-order peak prediction $\alpha^* = -g'(0)/g''(0)$ of Proposition 5(c).

Results. Figure 7 and Table 7 summarize the sweep; three predictions of Proposition 5 hold on every corpus.

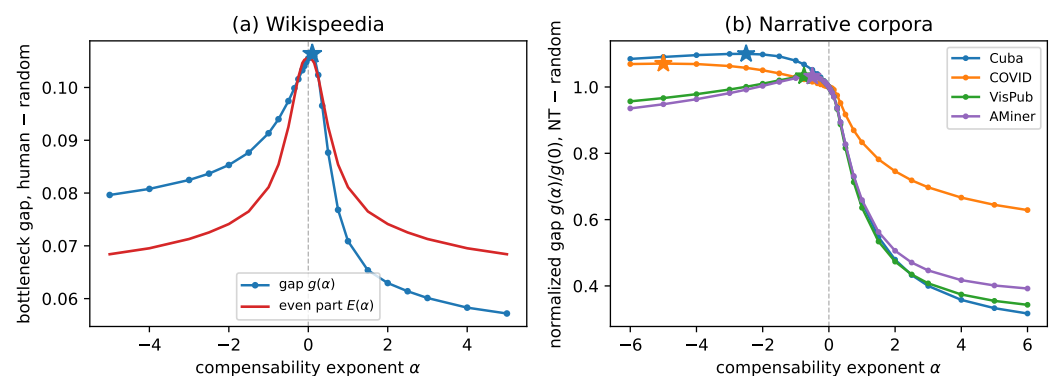


Figure 7. The bottleneck-gap profile $g(\alpha)$ (Experiment 7, Proposition 5). (a) Wikipedia, human versus random navigation: the gap $g(\alpha)$ (blue, star at its peak $\alpha = +0.10$) and its even part $E(\alpha)$ (red), maximized exactly at the geometric mean $\alpha = 0$. (b) The four narrative corpora, narrative-trail versus random chronological sequence; each gap normalized by its value $g(0)$ at the geometric mean; every profile peaks at $\alpha < 0$ (stars) modestly inside the compensatory side. Dashed line: $\alpha = 0$, the geometric mean.

Table 7. Bottleneck-gap profile across five corpora (Experiment 7). Wikipedia contrasts the human navigation path against a random path; the four narrative corpora contrast the narrative trail (the maximin storyline) against a random chronological sequence. n is the number of documents (articles, for Wikipedia) in the corpus; the gap is estimated over the corresponding endpoint-pair populations (10,832 human paths for Wikipedia; 169 to 2000 endpoint pairs for the narrative corpora). μ_{coh} and μ_R are the mean geometric-mean coherence of the coherent and random bottleneck edges, and their difference $g(0) = \mu_{\text{coh}} - \mu_R$ is the gap at the geometric mean. $g'(0)$ and $g''(0)$ are the first two Taylor coefficients of the profile at the geometric mean, given in closed form by Proposition 5(c); $\alpha_{\text{pred}}^* = -g'(0)/g''(0)$ is the leading-order predicted peak and α_{obs}^* the observed argmax of g over the swept grid. On every corpus, $g''(0) < 0$, so the profile is concave at the geometric mean and has an interior maximum, and the even part of g is maximized at $\alpha = 0$.

Corpus	n	μ_{coh}	μ_R	$g'(0)$	$g''(0)$	α_{pred}^*	α_{obs}^*
Wikipedia	3928	0.200	0.094	+0.0129	−0.0398	+0.32	+0.10
Cuba	418	0.702	0.177	−0.0695	−0.0757	−0.92	−2.5
COVID	40	0.774	0.588	−0.0062	−0.0056	−1.11	−5.0
VisPub	3549	0.651	0.108	−0.0775	−0.2186	−0.35	−0.75
AMiner	6000	0.666	0.116	−0.0800	−0.2367	−0.34	−0.5

(i) *The even part is maximized at the geometric mean.* On all five corpora the even part $E(\alpha) = \frac{1}{2}(g(\alpha) + g(-\alpha))$, measured directly from the sweep, decreases monotonically away from $\alpha = 0$ (plotted for Wikipedia in Figure 7a, red curve): the symmetric component

of the coherent-versus-incoherent separation is maximized *exactly* at the geometric mean. By Proposition 5(b), this is equivalent to the design-premise condition $\mathbb{E}_R[Y_\alpha] \geq \mathbb{E}_H[Y_\alpha]$, which therefore holds on every corpus.

(ii) *The full gap has an interior maximum.* On every corpus, $g(\alpha)$ has an interior maximum—never at the Min or Max endpoint (Table 7, α_{obs}^*). By Proposition 5(c) the peak sits at $\alpha^* = -g'(0)/g''(0)$. The curvature $g''(0)$ is negative on all five corpora (Table 7), so the sign of the displacement follows $g'(0)$: the peak lies above the geometric mean for Wikispeedia ($\alpha = +0.10$) and below it for the four narrative corpora. Here $g'(0) = \frac{1}{8}(\mathbb{E}_H[M_0\Delta^2] - \mathbb{E}_R[M_0\Delta^2])$ compares the M_0 -weighted channel imbalance of the coherent and random bottleneck edges (Proposition 5(c)): it is positive when the human navigation paths carry the more imbalanced bottleneck and negative when the narrative trail carries the less imbalanced one.

(iii) *The predicted peak.* The leading-order estimate $\alpha_{\text{pred}}^* = -g'(0)/g''(0)$ of Proposition 5(c) (Table 7) matches the observed peak α_{obs}^* in sign on all five corpora, and in magnitude within a small factor when the peak lies near 0 (a factor of 3.2, 2.1, and 1.5 for Wikispeedia, VisPub, and AMiner respectively); for Cuba and COVID the predicted and observed peak *locations* differ more, but this reflects the flatness of g on the compensatory side rather than a failure of the prediction: when the curvature at the geometric mean is small, the interior-critical-point guarantee of Proposition 5(c) lapses (its hypothesis $|g'(0)| < m\delta$ fails), so the leading-order *location* is no longer pinned down. There the profile forms a broad plateau—for both corpora g varies by under 4% across $\alpha \in [-8, -1]$ —so the point $\arg\max \alpha_{\text{obs}}^*$ is only weakly identified and is itself sensitive to sampling noise, which the path-level bootstrap confirms (Section S7). Measured instead by the quantity that matters for extraction, the gap *height*, the leading-order prediction is accurate on every corpus: the gap at the predicted peak $g(\alpha_{\text{pred}}^*)$ recovers between 92% and 99.7% of the maximum bottleneck gap (Wikispeedia 92.1%, Cuba 97.8%, COVID 96.4%, VisPub 99.7%, AMiner 99.7%). The closed-form peak therefore lands on the high-gap plateau in all cases, even where its precise location differs from the noisy empirical $\arg\max$.

(iv) *Separation at the geometric mean, and a non-circular anchor.* The mean gap $g(0)$ understates how cleanly coherent and incoherent storylines separate at the geometric mean. On Wikispeedia the human navigation paths separate from length-matched random paths with Cohen's $d = 0.65$ and ROC AUC = 0.77 on the maximin (bottleneck) objective, and more strongly ($d = 1.20$, AUC = 0.80) on the geometric-mean path-score (reliability) objective (both at $\alpha = 0$). On the four narrative corpora the separation is larger still ($d = 1.3$ to 4.1, AUC = 0.86 to 0.99), but these corpora are an *unsupervised* setting: lacking human-annotated storylines, we take the extractor's own maximin output as the coherent population, so the four narrative corpora serve as an internal consistency check rather than independent validation, and their separation magnitudes should be read in that light. The human-grounded validation of this experiment rests on Wikispeedia, where the coherent examples are actual human navigation paths rather than an output of the metric; it is the non-circular anchor, and it still separates the two populations at $d = 0.65$ (bottleneck) and $d = 1.20$ (reliability). The interior, compensatory-side peak is moreover not specific to the maximin objective that Proposition 5 analyzes: the geometric-mean reliability gap has an interior $\arg\max$ on every corpus too—negative on all four narrative corpora (-1.5 to -8) and -0.25 on Wikispeedia—so the same structure holds on both extraction objectives.

Robustness. These conclusions rest on replication across five independent corpora rather than on within-corpus resampling. For the smallest corpus, COVID (169 endpoint pairs), where sampling noise is most plausible, a path-level bootstrap (Supplementary Materials, Section S7) confirms that $g(0) > 0$, $g''(0) < 0$, $g'(0) < 0$ and $\alpha_{\text{obs}}^* < 0$ each hold across the full 95% interval; on the four larger corpora the corresponding intervals are far

tighter. As a cross-modal check, Supplementary Materials Section S8 applies the same analysis to a human-curated *image* narrative (the ROGER expedition-photograph corpus of German et al. [12]): the expert storyline scores above its random null and the even part of g are again maximized at the geometric mean, in a second modality.

The original narrative trails evaluation [7] reports a narrative-trail-versus-random bottleneck comparison at the single geometric-mean combinator; that measurement is the $\alpha = 0$ point of these profiles. Experiment 7 shows that point sits on the apex region of a curve whose structural separation—the even part—is, across five corpora, pinned to the geometric mean. This is an extraction-output counterpart to the axiomatic argument of Theorem 3: the geometric mean is not only the unique combinator consistent with the four axioms, it is also where the symmetric component of the storyline-versus-random separation is maximized.

8. Discussion

8.1. Interpretation

The product-manifold decomposition (Proposition 1) and Chentsov's theorem (Theorem 1) together provide an information-geometric reading of the composite coherence metric as a distance on the product space $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$ equipped with the round metric on the sphere factor and a Fisher-compatible metric on the simplex factor. Chentsov's theorem singles out the Fisher–Rao metric tensor (up to a positive scale) among Riemannian metrics on the simplex that are invariant under sufficient statistics; because $d_{\mathcal{J}}$ induces exactly this metric tensor (identity (7)), the topic component of the coherence metric is locally consistent with the Fisher–Rao geometry singled out by Chentsov's theorem. The axiomatic characterization (Theorem 3) is complementary: within the compensability spectrum, the geometric mean is the unique combinator consistent with the four stated axioms, and Section 4.3 discusses which axioms are load-bearing for this conclusion. Together, the geometric reading and the axiomatic characterization provide an information-theoretic justification for an existing operational construction [2,7].

8.2. Robustness in Practice

The mathematical pathologies of Section 5, in particular the non-metric status of $1 - C_{\text{GM}}$ from Proposition 6, require extreme channel imbalance ($A \approx 1$ with $T \approx 0$) to manifest. That configuration would need two documents to be nearly identical in embedding space yet assigned to completely disjoint topic clusters, which the UMAP→HDBSCAN coupling structurally prevents. Scale complementarity (Remark 3) is thus assigned a double duty: it justifies the product-manifold model (the channels are non-redundant) and it rules out the worst-case configurations that would break metricity.

8.3. Combinator Selection

Three properties favor C_{GM} as the operational tool despite $d_{\times}$ being a proper metric and Quad ($=1 - d_{\times}^2$) being its functionally equivalent similarity-domain counterpart. First, the *veto property* $C_{\text{GM}} = 0$ when $T = 0$ (Section 5) is redundant for the unconstrained maximin (narrative trails) extraction used in our experiments (Section 7.3), under which every combinator already avoids disjoint-topic edges, but becomes load-bearing under the cardinality-constrained linear-programming extraction of narrative maps [2], whose storyline-size requirement can force the path through otherwise-avoidable edges. Because $C_{\text{GM}} = \sqrt{A \cdot T}$ vanishes on disjoint-topic ($T = 0$) edges, those edges are pruned from the geometric mean's graph and its storylines can never traverse one, whereas the veto-less combinator ($d_{\times}$, AM) are routed through such transitions at a rate that grows with storyline length; the Supplementary Materials (Section S9) quantify this across an extraction-length

sweep. Because the composite metric is shared by both systems, this is a practical concern rather than a hypothetical one. Second, the log-additive decomposition (Proposition 3) keeps both channels diagnostically visible along extracted paths, whereas the squared product metric $d_{\times}^2$ concentrates 78 to 92% of the path budget in the angular channel and makes the topic channel nearly invisible. Third, the geometric mean is the established coherence metric in narrative maps [2] and trails [7]. Our framework justifies this existing choice rather than proposing a replacement.

Each combinator induces a distinct coherence landscape, a different weighting of the angular and topic channels, which extraction algorithms traverse differently, yielding structurally different narrative paths at statistically indistinguishable aggregate quality (Section 7.6). The choice of combinator is therefore a *modeling* decision about which narrative structures to foreground, not a free parameter to be tuned for quality. Table 4 makes this decision concrete: practitioners who want the axiomatic characterization and log-additive diagnostics should use GM; those who want a formal metric-space guarantee at the similarity level should use AM; those who want the product-manifold metric in similarity form should use Quad; and those who specifically want the veto property should use GM or Min. Our default recommendation of GM rests on its simultaneous combination of three properties: consistency with the four axioms of Theorem 3, the log-additive decomposition of Proposition 3, and the veto property. It is the only combinator in the compensability spectrum (Section 4.1) that satisfies all three together under the assumptions stated.

8.4. Limitations

LLM-as-judge evaluation, and the absence of a human anchor for the combinator comparison. The downstream evaluation in Section 7.6 uses two LLM judges (GPT-5.4 and Claude Opus 4.6) rather than human annotators. The evaluation protocol (brief justification with a single holistic score on a 1 to 100 scale, no anchoring) is designed to maximize discriminative power based on the findings of [51], and the judge prompt uses a domain-agnostic framing (“document sequence coherence”) to avoid privileging literary narrative criteria. Even so, LLM judges are known to exhibit biases that differ from human preferences (verbosity preference, positivity bias, sensitivity to surface form; [54]), and we did not run a human-evaluation study of combinator choices. Human grounding in this paper enters elsewhere: the bottleneck-gap analysis of Section 7.8 anchors the metric’s design premise to human navigation traces via the Wikipedia corpus.

Three mitigations partially address this gap, and one limitation remains. First, the multi-judge design [54] reduces single-model bias. Second, the role of the downstream evaluation in this paper is a *consistency check* on the axiomatic recommendation rather than a quality comparison: the relevant question is whether the geometric mean is empirically dominated, and a null result is harder for a judge’s bias to produce than a small positive effect would be, because systematic biases would have to cancel almost exactly across all combinators and baselines to yield a flat ranking. Establishing exact equivalence rather than non-domination would instead require an equivalence test such as the two one-sided tests (TOST) [55,56] against a pre-specified margin, which the consistency check does not attempt. Third, the inter-judge agreement statistics reported in Section 7.6 confirm that the two judges agree on the relative ordering of sequences even where their absolute score levels differ; the level offset that drives the negative Krippendorff α on VisPub does not affect the within-pair Friedman analysis, which is invariant to monotone rescaling of each judge’s scores.

The remaining limitation is that all three mitigations operate within the LLM-judge family: agreement between two strong LLM judges does not establish agreement with human readers, and the absolute level of LLM scores is not a calibrated quality measurement.

Two considerations bound the scope of this limitation. First, this paper is a *justification* of an existing coherence metric, not a proposal of a new one; the composite metric $C = \sqrt{A \cdot T}$ is already in use in the narrative maps [2] and narrative trails [7] systems, and the narrative maps representation has been assessed in human user studies, from the Amazon Mechanical Turk evaluation in the original work [2] to subsequent sensemaking studies [57,58], so the metric’s downstream usefulness has independent human-grounded support—although those studies evaluated the representation as a whole and did not compare combinator choices against one another. Second, the comparison most relevant here—between combinator within a single structural equivalence class—is one where the candidates are provably close: members of the class produce identical or near-identical paths (Section 7.6), so there is little mathematical room for a quality difference. Whether a human-perceptible difference nonetheless exists between mathematically near-equivalent combinator is a well-posed question in its own right, and a controlled human study of *that* question is the natural next step; a full human-evaluation study of narrative quality is otherwise beyond the scope of a paper whose contribution is the formalization of the metric rather than a new extraction method.

Prompt-domain trade-off. The domain-agnostic prompt (“document sequence coherence”) is a deliberate choice: it keeps Cuba and VisPub comparable under one criterion, which is essential for the cross-corpus consistency check, but may under-reward qualities specific to a domain (e.g., story-arc structure in news, logical-progression structure in research paper sequences). A domain-specific prompt per corpus would break this comparability, and since the metric itself is domain-agnostic ($C = \sqrt{A \cdot T}$), an agnostic yardstick matches what the metric is trying to optimize. A per-domain evaluation study is natural future work for benchmark-style comparisons.

Single coherence definition. We analyze the specific composite metric $C = \sqrt{A \cdot T}$ from the narrative maps framework [2]. The information-geometric framework applies whenever document similarity combines an embedding-space component with a probabilistic topic component, but we do not empirically compare against coherence definitions from other narrative extraction systems (e.g., word-influence [3], KL-based [9]).

Dependence on upstream representations. The theoretical results hold for any embedding and any soft topic assignment. The embedding sensitivity analysis (Supplementary Materials, Section S3) shows that GM rank correlation and triangle inequality violations are stable across three embedding architectures on two corpora, though channel correlation varies with embedding choice ($\rho_{A,T} \in [0.23, 0.29]$). The number of topic clusters has a stronger effect: the COVID corpus ($K = 2$) exhibits 7.6% triangle inequality violations for $1 - C_{GM}$, with the rate dropping monotonically as the cluster count increases. The practical recommendation is $K \geq 4$ clusters, which was sufficient in all other corpora tested. The clustering pipeline (UMAP hyperparameters, HDBSCAN configuration) was not varied independently of the cluster count.

Corpus scope. While the corpora used span news, academic, and encyclopedic texts (40 to 6000 documents) and three embedding architectures, all are English-language. Generalization to other languages or modalities was untested.

9. Conclusions

We have given an information-geometric reading of composite coherence metrics in event-based narrative extraction. On the product Riemannian manifold $\mathbb{S}^{d-1} \times \Delta_+^{K-1}$ the negative log-coherence decomposes additively into an angular and a topic cost; because the Jensen–Shannon distance induces a Riemannian tensor proportional to the Fisher information matrix, the topic component is locally consistent with the Fisher–Rao geometry that Chentsov’s theorem singles out, and the same construction motivates the

proper product metric $d_{\times}$ used as a reference distance. The central conclusion is a convergence: two arguments resting on different stated premises—the four operational axioms of Theorem 3 and the design premise that incoherent transitions are more channel-imbalanced (Proposition 5)—independently single out the geometric mean within the compensability spectrum, while the remaining combinators (AM, HM, Quad, Min, Max) each occupy a distinct, characterized position and the Quad combinator $1 - d_{\times}^2$ is recovered as the similarity-domain face of $d_{\times}$. The empirical study is consistent with this picture: across four corpora the metric-level experiments (Section 7) confirm the first-order Fisher approximation, the geometric mean's tracking of the product metric, and the non-redundancy of the two channels, while the supporting analyses establish robustness to the embedding model, the topic model, and the cluster count (Supplementary Materials). The downstream consistency check finds no alternative combinator or single-channel baseline that dominates the geometric mean on holistic quality, and the bottleneck-gap profile (Section 7.8)—extended to a human-navigation corpus and, in a cross-modal case study, to a human-curated image narrative (Supplementary Materials, Section S8)—shows the design premise holding throughout.

These results together provide an information-geometric justification for the composite coherence metric of [2,7] and articulate the conditions under which the geometric-mean combinator is the natural choice. The framework applies whenever document similarity combines a fine-grained embedding-space component with a coarse-grained probabilistic topic component under the design principle of *scale complementarity*, and it illustrates how information-geometric tools (Chentsov's theorem, the Fisher–Rao metric, product Riemannian manifolds, and the Jensen–Shannon distance) can clarify the structure of existing constructions in the information sciences without claiming optimality beyond what the stated assumptions support.

Several questions remain open. The downstream comparison shows that combinators within a single structural equivalence class are indistinguishable to current LLM judges; whether a human-perceptible quality difference nonetheless exists between mathematically near-equivalent combinators is a well-posed question that a controlled human study could settle. Generalization beyond English-language text—to other languages, to modalities beyond the single image-narrative case study, and to coherence definitions from other extraction systems—likewise remains untested. More broadly, scale complementarity as a design principle, and the bottleneck-gap profile as a diagnostic of where a combinator sits on the compensability spectrum, may transfer to other settings in which a similarity combines a fine-grained geometric component with a coarse-grained distributional one; characterizing that transfer is natural future work.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/e28070732/s1>: Section S1, validation on alternative topic models (LDA, soft k -means, GMM); Section S2, cluster-count sensitivity; Section S3, embedding-model sensitivity; Section S4, a component ablation isolating the contributions of soft cluster membership and of the Jensen–Shannon distance; Section S5, the per-seed breakdown of the embedding-perturbation experiment (Section 7.7); Section S6, a heuristic rate-distortion analogy for the geometric mean; Section S7, a path-level bootstrap quantifying the sampling uncertainty of the Experiment 7 bottleneck-gap profile on the smallest corpus; Section S8, a cross-modal case study applying the bottleneck-gap analysis to a human-curated image narrative (the ROGER corpus); and Section S9, a validation of the geometric-mean veto property under the cardinality-constrained narrative maps linear-programming extraction.

Funding: This research is funded by the ANID FONDECYT grant 11250039. The corresponding author is also supported by Universidad Católica del Norte grant 202311010033-VRIDT-UCN.

Data Availability Statement: The data and code used in the experiments is provided as part of the Supplementary Materials.

Acknowledgments: During the preparation of this work, the author used Claude to refine writing and support literature review activities. Additionally, Writefull integrated in Overleaf was used to improve writing quality and readability. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of the article.

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

Abbreviations

The following abbreviations are used in this manuscript:

JSD	Jensen–Shannon divergence
KL	Kullback–Leibler divergence
FR	Fisher–Rao
GM	Geometric mean
AM	Arithmetic mean
HM	Harmonic mean
NMI	Normalized mutual information
LLM	Large language model
LDA	Latent Dirichlet Allocation
GMM	Gaussian Mixture Model
UMAP	Uniform Manifold Approximation and Projection
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
DPI	Data processing inequality
LP	Linear program
MST	Minimum spanning tree

References

1. Halverson, J.R.; Goodall, H.L.; Corman, S.R. *Master Narratives of Islamist Extremism*; Palgrave Macmillan: New York, NY, USA, 2011. <https://doi.org/10.1007/978-0-230-11723-5>.
2. Keith, B.; Mitra, T. Narrative Maps: An Algorithmic Approach to Represent and Extract Information Narratives. *Proc. ACM Hum.-Comput. Interact.* **2021**, *4*, 1–33. <https://doi.org/10.1145/3432927>.
3. Shahaf, D.; Guestrin, C. Connecting the Dots Between News Articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2010; pp. 623–632. <https://doi.org/10.1145/1835804.1835884>.
4. Vossen, P.; Caselli, T.; Kontzopoulou, Y. Storylines for Structuring Massive Streams of News. In *Proceedings of the 1st Workshop on Computing News Storylines*; Association for Computational Linguistics: Beijing, China, 2015; pp. 40–49. <https://doi.org/10.18653/v1/W15-4507>.
5. Keith Norambuena, B.; Mitra, T.; North, C. A Survey on Event-Based News Narrative Extraction. *ACM Comput. Surv.* **2023**, *55*, 300. <https://doi.org/10.1145/3584741>.
6. Shahaf, D.; Guestrin, C.; Horvitz, E. Trains of Thought: Generating Information Maps. In *Proceedings of the 21st International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 899–908. <https://doi.org/10.1145/2187836.2187957>.
7. German, F.; Keith, B.; North, C. Narrative Trails: A Method for Coherent Storyline Extraction via Maximum Capacity Path Optimization. In *Proceedings of the 8th International Workshop on Narrative Extraction from Texts (Text2Story 2025)*, Lucca, Italy, 10 April 2025; CEUR Workshop Proceedings, Volume 3964, pp. 15–27. Available online: <https://ceur-ws.org/Vol-3964/paper2.pdf> (accessed on 10 March 2026).
8. Yan, R.; Wan, X.; Otterbacher, J.; Kong, L.; Li, X.; Zhang, Y. Evolutionary Timeline Summarization: A Balanced Optimization Framework via Iterative Substitution. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2011; pp. 745–754. <https://doi.org/10.1145/2009916.2010016>.

9. Li, J.; Li, S. Evolutionary Hierarchical Dirichlet Process for Timeline Summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Association for Computational Linguistics: Sofia, Bulgaria, 2013; pp. 556–560. Available online: <https://aclanthology.org/P13-2099> (accessed on 10 March 2026).
10. Chentsov, N.N. *Statistical Decision Rules and Optimal Inference*; Translations of Mathematical Monographs; AMS: Providence, RI, USA, 1982; Volume 53. <https://doi.org/10.1090/mmono/053>.
11. West, R.; Leskovec, J. Human Wayfinding in Information Networks. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 619–628. <https://doi.org/10.1145/2187836.2187920>.
12. German, F.; Keith, B.; Matus, M.; Urrutia, D.; Meneses, C. Semi-Supervised Image-Based Narrative Extraction: A Case Study with Historical Photographic Records. In *Proceedings of the Advances in Information Retrieval—47th European Conference on Information Retrieval (ECIR 2025)*, Cham, Switzerland; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2025; Volume 15573, pp. 248–262. https://doi.org/10.1007/978-3-031-88711-6_16.
13. Zhou, W.; Shen, C.; Li, T.; Chen, S.C.; Xie, N. Generating Textual Storyline to Improve Situation Awareness in Disaster Management. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration*; IEEE: New York, NY, USA, 2014; pp. 585–592. <https://doi.org/10.1109/IRI.2014.7051942>.
14. Tran, G.B.; Tran, T.; Tran, N.K.; Alrifai, M.; Kanhabua, N. Leveraging Learning to Rank in an Optimization Framework for Timeline Summarization. In *Proceedings of the SIGIR 2013 Workshop on Time-Aware Information Access (TAIA)*; Association for Computing Machinery: New York, NY, USA, 2013. Available online: <https://www.semanticscholar.org/paper/33d5e4161c1a9f8a94ee9ee7b537dc5a4202f6fc> (accessed on 10 March 2026).
15. Li, R.; Wang, T.; Wang, X. Tracking Events Using Time-Dependent Hierarchical Dirichlet Tree Model. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*; SIAM: Philadelphia, PA, USA, 2015; pp. 550–558. <https://doi.org/10.1137/1.9781611974010.62>.
16. Xu, N.; Tang, X. Generating Risk Maps for Evolution Analysis of Societal Risk Events. In *Proceedings of the Knowledge and Systems Sciences (KSS)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 115–128. https://doi.org/10.1007/978-981-13-3149-7_9.
17. Liu, B.; Niu, D.; Lai, K.; Kong, L.; Xu, Y. Growing Story Forest Online from Massive Breaking News. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM)*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 777–785. <https://doi.org/10.1145/3132847.3132852>.
18. Liu, B.; Han, F.X.; Niu, D.; Kong, L.; Lai, K.; Xu, Y. Story Forest: Extracting Events and Telling Stories from Breaking News. *ACM Trans. Knowl. Discov. Data* **2020**, *14*, 1–28. <https://doi.org/10.1145/3377939>.
19. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426. <https://doi.org/10.48550/arXiv.1802.03426>.
20. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861. <https://doi.org/10.21105/joss.00861>.
21. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.
22. Endres, D.M.; Schindelin, J.E. A New Metric for Probability Distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. <https://doi.org/10.1109/TIT.2003.813506>.
23. Österreicher, F.; Vajda, I. A New Class of Metric Divergences on Probability Spaces and Its Applicability in Statistics. *Ann. Inst. Stat. Math.* **2003**, *55*, 639–653. <https://doi.org/10.1007/BF02517812>.
24. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. <https://doi.org/10.1109/18.61115>.
25. Rao, C.R. Information and the Accuracy Attainable in the Estimation of Statistical Parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91. Reprinted in *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1992. https://doi.org/10.1007/978-1-4612-0919-5_16.
26. Amari, S.i. *Differential-Geometrical Methods in Statistics*; Lecture Notes in Statistics; Springer: Berlin, Germany, 1985; Volume 28. <https://doi.org/10.1007/978-1-4612-5056-2>.
27. Amari, S.i.; Nagaoka, H. *Methods of Information Geometry*; Translations of Mathematical Monographs; AMS: Providence, RI, USA, 2000; Volume 191. <https://doi.org/10.1090/mmono/191>.
28. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. *Information Geometry*; Ergebnisse der Mathematik und ihrer Grenzgebiete; Springer: Cham, Switzerland, 2017; Volume 64. <https://doi.org/10.1007/978-3-319-56478-4>.
29. Nielsen, F. An Elementary Introduction to Information Geometry. *Entropy* **2020**, *22*, 1100. <https://doi.org/10.3390/e22101100>.
30. Lafferty, J.; Lebanon, G. Diffusion Kernels on Statistical Manifolds. *J. Mach. Learn. Res.* **2005**, *6*, 129–163. Available online: <https://jmlr.org/papers/v6/lafferty05a.html> (accessed on 10 March 2026).
31. Lebanon, G. Riemannian Geometry and Statistical Machine Learning. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2005. Available online: <https://www.cs.cmu.edu/~lebanon/pub/thesis/thesis.pdf> (accessed on 2 February 2026).

32. Colombo, P.; Clavel, C.; Piantanida, P. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 10554–10562. <https://doi.org/10.1609/aaai.v36i10.21299>.
33. Dhillon, I.S.; Mallela, S.; Kumar, R. A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification. *J. Mach. Learn. Res.* **2003**, *3*, 1265–1287. Available online: <https://jmlr.org/papers/v3/dhillon03a.html> (accessed on 2 February 2026).
34. Nickel, M.; Kiela, D. Poincaré Embeddings for Learning Hierarchical Representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
35. Fuglede, B.; Topsøe, F. Jensen–Shannon Divergence and Hilbert Space Embedding. In *Proceedings of the IEEE International Symposium on Information Theory*; IEEE: New York, NY, USA, 2004; p. 30. <https://doi.org/10.1109/ISIT.2004.1365067>.
36. Campbell, L.L. An Extended Chentsov Characterization of the Information Metric. *Proc. Am. Math. Soc.* **1986**, *98*, 135–141. <https://doi.org/10.1090/S0002-9939-1986-0848890-5>.
37. Bauer, M.; Bruveris, M.; Michor, P.W. Uniqueness of the Fisher–Rao Metric on the Space of Smooth Densities. *Bull. Lond. Math. Soc.* **2016**, *48*, 499–506. <https://doi.org/10.1112/blms/bdw020>.
38. Topsøe, F. Some Inequalities for Information Divergence and Related Measures of Discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609. <https://doi.org/10.1109/18.850703>.
39. Lee, J.M. *Introduction to Riemannian Manifolds*, 2nd ed.; Graduate Texts in Mathematics; Springer: Cham, Switzerland, 2018; Volume 176. <https://doi.org/10.1007/978-3-319-91755-9>.
40. Xu, C.; Tao, D.; Xu, C. A Survey on Multi-View Learning. *arXiv* **2013**, arXiv:1304.5634. <https://doi.org/10.48550/arXiv.1304.5634>.
41. Bhatia, R. *Positive Definite Matrices*; Princeton Series in Applied Mathematics; Princeton University Press: Princeton, NJ, USA, 2009. <https://doi.org/10.1515/9781400827787>.
42. Bullen, P.S. *Handbook of Means and Their Inequalities*; Mathematics and Its Applications; Springer: Dordrecht, The Netherlands, 2003; Volume 560. <https://doi.org/10.1007/978-94-017-0399-4>.
43. Burago, D.; Burago, Y.; Ivanov, S. *A Course in Metric Geometry*; Graduate Studies in Mathematics; American Mathematical Society: Providence, RI, USA, 2001; Volume 33. <https://doi.org/10.1090/gsm/033>.
44. Aczél, J.; Dhombres, J. *Functional Equations in Several Variables*; Encyclopedia of Mathematics and its Applications; Cambridge University Press: Cambridge, UK, 1989; Volume 31. <https://doi.org/10.1017/CBO9781139086578>.
45. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006. <https://doi.org/10.1002/047174882X>.
46. Goemans, M.X.; Williamson, D.P. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J. ACM* **1995**, *42*, 1115–1145. <https://doi.org/10.1145/227683.227684>.
47. Charikar, M.S. Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*; Association for Computing Machinery: New York, NY, USA, 2002; pp. 380–388. <https://doi.org/10.1145/509907.509965>.
48. Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; Weld, D.S. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Online, 2020; pp. 2270–2282. <https://doi.org/10.18653/v1/2020.acl-main.207>.
49. Singh, A.; D’Arcy, M.; Cohan, A.; Downey, D.; Feldman, S. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Singapore, 2023; pp. 5548–5566. <https://doi.org/10.18653/v1/2023.emnlp-main.338>.
50. Edmonds, J. Optimum Branchings. *J. Res. Natl. Bur. Stand. B* **1967**, *71B*, 233–240. <https://doi.org/10.6028/jres.071B.032>.
51. Keith, B. LLM-as-a-Judge Approaches as Proxies for Mathematical Coherence in Narrative Extraction. *Electronics* **2025**, *14*, 2735. <https://doi.org/10.3390/electronics14132735>.
52. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. <https://doi.org/10.1080/01621459.1937.10503522>.
53. Nemenyi, P. Distribution-Free Multiple Comparisons. Ph.D. thesis, Princeton University, Princeton, NJ, USA, 1963. Available online: <https://catalog.princeton.edu/catalog/9920813653506421> (accessed on 26 February 2026).
54. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
55. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680. <https://doi.org/10.1007/BF01068419>.
56. Lakens, D. Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **2017**, *8*, 355–362. <https://doi.org/10.1177/1948550617697177>.

57. Keith Norambuena, B.; Mitra, T.; North, C. Design Guidelines for Narrative Maps in Sensemaking Tasks. *Inf. Vis.* **2022**, *21*, 220–245. <https://doi.org/10.1177/14738716221079593>.
58. Keith Norambuena, B.; German, F.; Krokos, E.; Joseph, S.; North, C. Semantic Interaction for Narrative Map Sensemaking: An Insight-Based Evaluation. In Proceedings of the 9th International Workshop on Narrative Extraction from Texts (Text2Story 2026), 2026, Delft, The Netherlands, 29 March 2026 ; CEUR Workshop Proceedings, Volume 4202. Available online: <https://ceur-ws.org/Vol-4202/paper4.pdf> (accessed on 5 May 2026).

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.