electronics

MDPI

Article

LLM-as-a-Judge Approaches as Proxies for Mathematical Coherence in Narrative Extraction

Brian Keith

Department of Systems and Computing Engineering, Universidad Católica del Norte, Antofagasta 1270709, Chile; brian.keith@ucn.cl

Abstract

Evaluating the coherence of narrative sequences extracted from large document collections is crucial for applications in information retrieval and knowledge discovery. While mathematical coherence metrics based on embedding similarities provide objective measures, they require substantial computational resources and domain expertise to interpret. We propose using large language models (LLMs) as judges to evaluate narrative coherence, demonstrating that their assessments correlate with mathematical coherence metrics. Through experiments on two data sets—news articles about Cuban protests and scientific papers from visualization conferences—we show that the LLM judges achieve Pearson correlations up to 0.65 with mathematical coherence while maintaining high inter-rater reliability (ICC > 0.92). The simplest evaluation approach achieves a comparable performance to the more complex approaches, even outperforming them for focused data sets while achieving over 90% of their performance for the more diverse data sets while using less computational resources. Our findings indicate that LLM-as-a-judge approaches are effective as a proxy for mathematical coherence in the context of narrative extraction evaluation.

Keywords: narrative evaluation; narrative extraction; large language models; coherence metrics; LLM-as-a-judge

Academic Editors: Hao Fei, Fei Li and Wei Ji

Received: 5 June 2025
Revised: 29 June 2025
Accepted: 4 July 2025
Published: 7 July 2025

Citation: 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

1. Introduction

The extraction and evaluation of coherent narrative sequences from large document collections has become increasingly important in knowledge discovery and information retrieval applications [1–3]. Narrative coherence [4]—the degree to which a sequence of documents forms a logical, thematically consistent story—is traditionally measured using mathematical metrics based on embedding similarities and topic distributions [5].

However, these metrics require significant computational resources and domain expertise for interpretation and may not capture the nuanced aspects of coherence that human readers perceive. Thus, these metrics create significant barriers to practical adoption. When a coherence score is 0.73, what does this actually mean for narrative quality? Domain experts such as journalists, intelligence analysts, and researchers often lack the mathematical background to interpret angular similarities in embedding spaces or Jensen–Shannon divergence calculations [6], limiting their ability to assess and improve extracted narratives.

A fundamental distinction exists between narrative extraction and narrative generation that shapes all evaluation considerations. Narrative generation creates new fictional text from prompts or outlines, producing content that can be evaluated against reference stories or quality metrics. In contrast, narrative extraction identifies and sequences existing documents from collections to reveal underlying stories.

We note that **narrative extraction** as a task differs fundamentally from **narrative generation**: while generation seeks to create new stories from scratch, extraction seeks to discover the underlying stories by selecting and ordering documents from document collections [1]. This distinction is crucial: while generation benefits from established metrics like BLEU [7] or ROUGE [8] that compare against reference texts, extraction faces a unique challenge—there are no "correct" narratives to serve as ground truth. Multiple valid narratives can be extracted from any document collection, making traditional evaluation paradigms inapplicable. This is not an oversight but a fundamental characteristic of the extraction task. Thus, evaluating extracted narratives requires assessing document selection and ordering quality, not generated text quality.

The field has seen significant advances in timeline summarization [9–11] and event-based narrative extraction [12], with coherence measurement remaining a central challenge. Recent surveys have analyzed the landscape of narrative extraction methods [1–3] establishing relevant taxonomies for representations, extraction methods, and evaluation approaches. However, there are several challenges that remain to be solved, such as having effective evaluation metrics and approaches that do not require extensive human evaluation or expensive computations [1].

Building upon the graph-based approaches pioneered by Shahaf and Guestrin's **Metro Maps** [13] and Keith and Mitra's **Narrative Maps** [4] methods, the **Narrative Trails** algorithm [5] addresses narrative extraction by formulating it as a **maximum capacity path problem**, where paths maximize the minimum coherence between consecutive documents. This coherence metric [4] combines angular similarity in high-dimensional embedding spaces with topic similarity based on cluster distributions, which requires a deep understanding of information theory and embedding spaces for interpretation. Although this approach produces mathematically optimal narratives, evaluating their quality remains a challenge [1]. The field currently has no standardized evaluation methods beyond mathematical coherence metrics. Surveys of narrative extraction [1] reveal that existing works either use these opaque mathematical metrics or create custom evaluation approaches that cannot be compared across studies. This evaluation gap severely limits the practical adoption of narrative extraction tools.

Recent advances in large language models (LLMs) have demonstrated their capability to perform complex evaluation tasks. The emergence of **LLM-as-a-judge** approaches has shown promise in various domains, from creative writing evaluation [14] to narrative messaging analysis [15]. Multi-agent evaluation frameworks have shown superior performance over single-agent approaches [16], while reliability studies have revealed both the strengths and limitations of LLM-based evaluation [17,18].

In this article, we investigate whether LLMs can serve as effective proxies for mathematical coherence metrics in the evaluation of narratives defined as sequences of events. This approach could democratize narrative evaluation by providing *interpretable assessments* without requiring deep technical knowledge of embedding spaces or coherence formulations. In particular, our work makes three primary contributions to the field of narrative extraction evaluation:

1. We demonstrate that LLM-as-a-judge approaches can serve as effective proxies for mathematical coherence metrics in narrative extraction, achieving meaningful correlations while maintaining high inter-rater reliability across different evaluation modes.
2. We show that simple evaluation prompts achieve a comparable performance to complex approaches, reaching 85–90% of their effectiveness while using substantially fewer computational resources, providing practical guidance for implementation.

3. We provide the first scalable evaluation method for narrative extraction that does not require technical expertise in embedding spaces or information theory, making narrative quality assessments accessible to domain experts.

More specifically, we conducted experiments on two contrasting data sets: (1) a focused collection of news articles about Cuban protests and (2) a diverse collection of visualization research papers that span three decades. Our analysis reveals that LLM-based evaluations achieve stronger correlations with mathematical coherence for the heterogeneous scientific paper data set (up to $r = 0.65$) compared to the topically focused news data set (up to $r = 0.46$). Furthermore, our multi-agent design addresses known biases in LLM evaluation, achieving reliability levels that exceed typical human annotation studies. These findings suggest that simple, well-designed LLM evaluation can bridge the gap between mathematical optimization and practical narrative assessment. Finally, we demonstrate that a simple evaluation achieves a performance comparable to complex approaches, even outperforming them for focused data sets while achieving more than 90% of their performance for diverse data sets. Thus, simple evaluation prompts provide the best cost–performance trade-off.

2. Related Work

2.1. Narrative Extraction and Coherence

Narrative extraction methods aim to identify meaningful sequences of documents that tell coherent stories. Early work by Shahaf and Guestrin introduced Connect the Dots [13], which uses linear programming to find coherent chains between documents based on word-level features. This approach was extended in Metro Maps [19,20] to create multithreaded narrative structures reminiscent of public transportation maps.

The field has since expanded significantly, with approaches ranging from event chain extraction [12] to dynamic attention-based methods for joint event extraction [21]. Timeline summarization has become a particularly active area, with recent work introducing incremental approaches leveraging LLMs [9], constrained summarization with self-reflection [10], and date-first paradigms [11].

The Narrative Maps algorithm [4] and its derived approaches advanced the field by incorporating user interactions [22] and coverage constraints based on the embedding representations generated by deep learning models, allowing more complex narrative structures. However, these methods rely heavily on relatively slow linear programming approaches and heuristics to introduce user feedback, making them difficult to scale and generalize across domains.

Coherence measurement has evolved from word-level features [13] to more sophisticated approaches [1]. Castricato et al. [23] introduced the concept of *Fabula Entropy Indexing*, an objective measure based on entropy for story coherence in the context of narrative generation. For multi-document scenarios, the SEM-F1 metric [24] has shown that traditional ROUGE metrics are insufficient to evaluate semantic overlap in narrative summaries. The *Narrative Trails* algorithm [5] takes a different approach by leveraging the semantic-level information embedded in the latent space of deep learning models. Based on previous work on narrative extraction [1,5], we take the definition of narrative coherence between two documents d_u and d_v to be

$$\theta(d_u, d_v) = \sqrt{S(z_u, z_v) \cdot T(\hat{z}_u, \hat{z}_v)}, \tag{1}$$

where $S(z_u, z_v) = 1 - \arccos(\text{cos\_sim}(z_u, z_v))/\pi$ represents the angular similarity in the embedding space and $T(\hat{z}_u, \hat{z}_v) = 1 - \text{JSD}(\hat{z}_u, \hat{z}_v)$ represents the topic similarity based on the Jensen–Shannon divergence between the distributions of cluster membership.

2.1.1. Narrative Generation Versus Extraction

The field of computational narratives encompasses two fundamentally distinct tasks: narrative generation and narrative extraction. Narrative generation seeks to create new stories based on prompts, outlines, or more complex structural definitions, producing fictional or creative text that can be evaluated against reference stories or quality metrics [25,26]. In contrast, narrative extraction identifies and sequences existing documents from collections to reveal underlying stories [1]. This distinction is crucial yet often overlooked due to the overloaded terminology in the field of computational narratives. Furthermore, this key distinction shapes all subsequent evaluation considerations, as we discuss below.

The evaluation paradigms for these tasks differ fundamentally. Generation tasks benefit from established metrics like BLEU [7] and ROUGE [8] that compare generated text against references. Extraction tasks lack such ground truth, as multiple valid narratives can emerge from any document collection, and the task involves selecting and ordering existing documents rather than creating new text. Recent surveys [1,3] confirm that narrative extraction relies primarily on mathematical coherence metrics, as traditional generation metrics cannot assess document selection and ordering quality. This evaluation gap motivates our exploration of alternative assessment approaches based on LLMs.

2.1.2. Coherence Evaluation Methods

While narrative generation has developed sophisticated coherence metrics, extraction relies on mathematical measures that require technical expertise. Generation approaches like SCORE [27] and CoUDA [28] leverage neural architectures to assess story quality, but assume generated text rather than document sequences or more complex structures. These methods cannot directly evaluate whether selected documents form coherent narratives or assess the quality of document ordering.

For extraction tasks, coherence measurement typically involves angular similarity in embedding spaces and topic distribution divergence [4,5]. The text2story toolkit [29] exemplifies this challenge, as despite providing extraction capabilities, it lacks dedicated evaluation beyond experimental modules. This gap between extraction capabilities and evaluation methods presents a barrier to practical adoption, as domain experts often lack the mathematical background to interpret coherence scores.

2.2. LLM-Based Evaluation

The emergence of large language models has opened new possibilities for automated evaluation on various NLP tasks [30]. LLMs have been successfully applied to evaluate machine translation quality [31], summarization effectiveness [32], and dialogue system performance [33]. The LLM-as-a-judge paradigm leverages these models' ability to assess quality based on learned representations of human preferences.

Recent work has explored the reliability and consistency of LLM evaluators. G-Eval [34] introduced a framework for using GPT-4 to evaluate natural language generation with *Chain-of-Thought* (CoT) reasoning [35]. GPTScore [36] extended this by testing 19 pre-trained models on 22 evaluation aspects, demonstrating a higher correlation with human judgments for the assessment of summarization coherence. Multi-agent approaches such as ChatEval [16] have shown that diverse role prompts in multi-agent referee teams achieve better accuracy and correlation with human assessment than single-agent evaluators.

Reliability and bias concerns have been systematically addressed in recent studies. Doostmohammadi et al. [17] revealed that while simple metrics like ROUGE-L correlate well for short answer tasks, they fail in free-form generation, and the effectiveness of GPT-4 diminishes without reference answers. Chen et al. [18] investigated multiple types of biases, including misinformation oversight, gender, authority, and beauty biases in both

human and LLM judges, demonstrating successful bias exploitation attacks and proposing evaluation frameworks free from ground truth dependency.

The intersection of narrative extraction and LLM evaluation has produced innovative approaches. HEART-felt Narratives [37] exemplifies this bridge by using large language models for both narrative element extraction and the evaluation of the impact of narrative quality on empathy. Gómez and Williams [14] evaluated 12 LLMs in creative writing in multiple dimensions, finding that state-of-the-art commercial LLMs match or outperform human writers in most dimensions except creativity. Domain-specific applications have emerged, such as the use of GPT-4 to extract and analyze narrative messaging differences in climate change coverage across cultures [15]. However, concerns about bias, cost, and reproducibility have limited widespread adoption of systematic evaluation tasks. Our work addresses these concerns by employing multiple agents with different parameters and demonstrating high inter-rater reliability across evaluation modes.

2.2.1. Mathematical Reasoning in NLP

Given the absence of traditional evaluation metrics for narrative extraction, we also examine whether assessment capabilities from other domains might transfer. Mathematical reasoning in NLP provides insights into the distinction between computation and evaluation. GSM8K [38] and MathQA [39] test computational problem solving, while MR-GSM8K [40] introduces meta-reasoning tasks where models assess solution correctness without computing answers.

The performance gap between these tasks is striking, as models like DeepSeek-v2 achieve 88% on computation but only 18% on assessment. ReasonEval [41] further demonstrates that models can evaluate mathematical reasoning quality independently of computational accuracy, analogous to how experts assess proofs without re-deriving steps. This separation between computation and assessment suggests that language models might similarly evaluate narrative coherence without computing embedding similarities.

2.2.2. Biases in LLM Evaluation

Using LLMs to evaluate narrative extraction introduces specific bias concerns beyond those in generation tasks. The CALM framework [42] identifies twelve bias types affecting LLM judges, including position bias and self-enhancement bias. These biases particularly impact extraction evaluation, where document order and selection must be assessed objectively.

Multi-agent approaches effectively mitigate evaluation biases. PoLL [43] shows that ensemble methods outperform single judges, while position bias mitigation [44] improves consistency. For extraction evaluation, where no ground truth exists, these bias controls become essential. Randomizing document presentation, controlling for length effects, and using multiple evaluators help ensure reliable assessments of extracted narratives.

2.2.3. Hallucination in Evaluation Tasks

Hallucination mitigation strategies differ significantly between generation and evaluation contexts. Comprehensive surveys [45,46] focus on preventing models from fabricating content during text generation. However, evaluation tasks present different challenges, as models must assess existing content rather than create new information.

Constrained output formats show promise for reducing evaluation errors. Béchard and Ayala [47] achieve low hallucination rates through structured outputs, while Kollias et al. [48] demonstrate that generation constraints improve factual accuracy. For evaluation tasks specifically, constraining outputs to numerical scores or structured assessments reduces opportunities for fabrication. However, format bias studies [49] warn that rigid constraints may introduce artifacts, suggesting balanced approaches that maintain assess-

ment validity. Nevertheless, we follow a structured assessment approach as it provides a simple baseline for our LLM-as-a-judge approach.

2.2.4. Evaluation Complexity Trade-Offs

The computational demands of extraction evaluation motivate examinations of complexity trade-offs. Advanced techniques like self-consistency [50] and chain-of-verification [51] improve evaluation quality but require substantial computational overhead. For large document collections typical in extraction tasks, these costs multiply quickly.

Recent evidence favors simpler approaches without sacrificing quality. Multi-agent systems show diminishing returns beyond 2–5 agents, with performance improvements plateauing while computational costs scale linearly [52]. Confidence-informed methods similarly reduce computational requirements while maintaining accuracy [53]. For narrative extraction evaluation, where scalability matters due to large document collections and multiple narrative paths, these findings suggest starting with minimal agent configurations before adding complexity. The challenge lies in balancing evaluation quality with practical computational constraints, particularly when processing thousands of potential narrative sequences.

These considerations inform our approach to narrative extraction evaluation, as detailed in the following section. In particular, building on these insights, we propose an LLM-as-judge approach that addresses the specific challenges of narrative extraction evaluation.

3. Methodology

3.1. Narrative Extraction

We extract narrative sequences using two contrasting methods to establish a clear quality gradient for evaluation. Max capacity paths represent optimal narratives that maximize the minimum coherence between consecutive documents, extracted using the Narrative Trails algorithm [5]. These paths solve the maximum capacity path problem on a sparse coherence graph where edge weights represent document-to-document coherence.

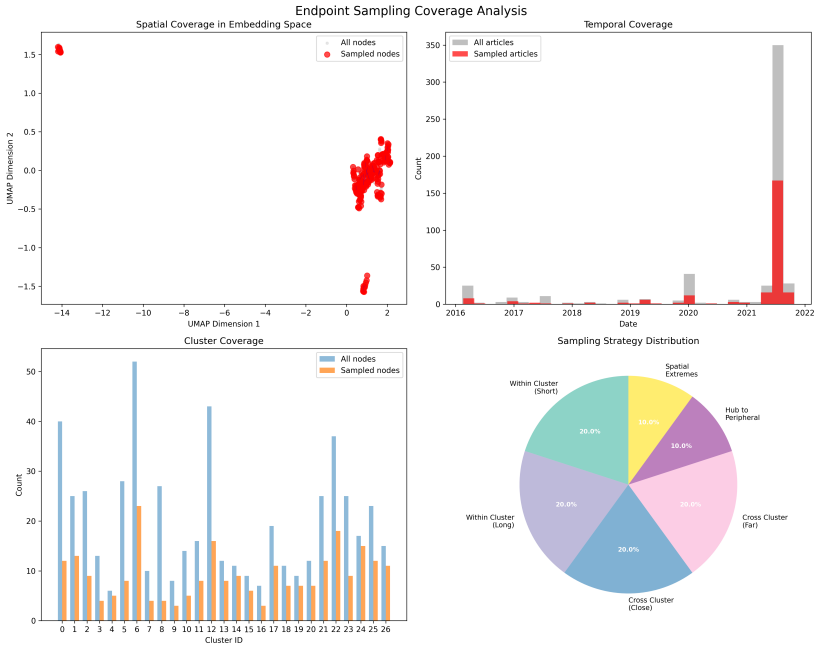
Random chronological paths serve as baselines, created by randomly sampling documents in chronological order between the source and target endpoints. For each source–target pair, we sample intermediate documents from those that occur temporally between the endpoints, maintaining chronological order but without coherence optimization. The path length follows a normal distribution that matches the mean and standard deviation of the length of the max capacity paths to ensure a fair comparison.

3.2. Endpoint Sampling Strategy

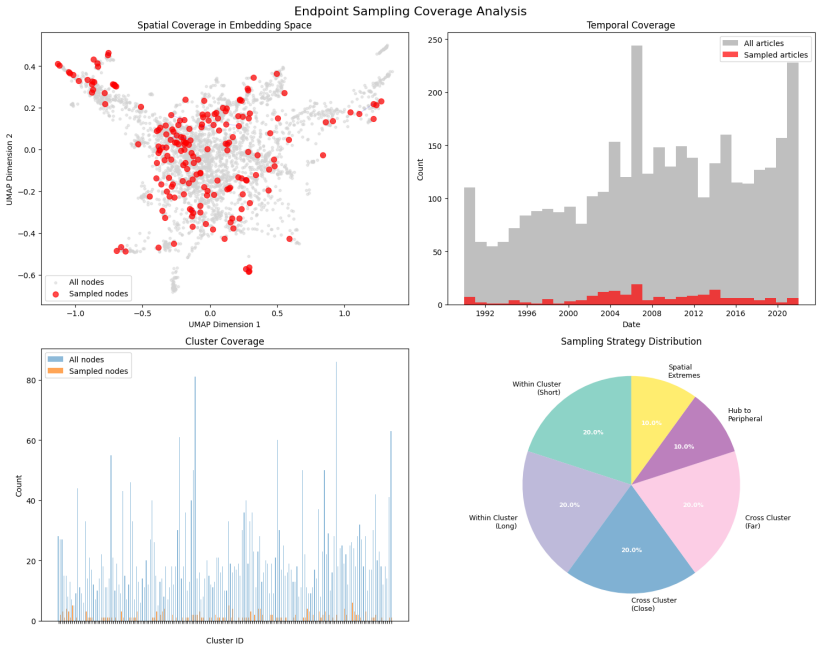
To ensure a diverse and representative evaluation, we employ a stratified endpoint sampling strategy. This strategy considers multiple approaches to select endpoints with the goal of generating diverse narratives to test our evaluation approaches. Figure 1 illustrates the coverage achieved by our sampling approach in both data sets.

The sampling strategy balances six different approaches to capture various types of narratives. **Within-cluster short temporal paths** (20%) select endpoints from the same topical cluster with minimal temporal distance, testing coherence for focused narratives. **Within-cluster long temporal paths** (20%) choose endpoints from the same cluster but separated by significant time, evaluating how well the system handles topic evolution. **Cross-cluster close temporal paths** (20%) select endpoints from different clusters within similar time periods, assessing thematic transitions. **Cross-cluster far temporal paths** (20%) combine different clusters and time periods, testing the system's ability to bridge diverse content. **Hub-to-peripheral paths** (10%) connect highly connected documents to sparsely connected ones, evaluating navigation from central to niche topics. **Spatial extremes** (10%)

select endpoints that are maximally separated in the embedding space, testing coherence across semantic distances.



(**a**) News data set: endpoint sampling coverage showing concentrated temporal distribution around 2016 Cuban protests.



(**b**) VisPub data set: endpoint sampling coverage spanning 30 years of visualization research.

Figure 1. Endpoint sampling coverage analysis for both data sets. The sampling strategy ensures diverse coverage across spatial embedding dimensions, temporal spans, and topical clusters while maintaining connectivity within the sparse coherence graph.

The sampling process also ensures connectivity within the sparse coherence graph, rejecting endpoint pairs without valid paths. We generated 200 endpoint pairs for the news data set and 100 endpoint pairs for the scientific papers data set.

3.3. LLM-as-a-Judge Framework Overview

We evaluate narratives using GPT-4.1 with four distinct evaluation modes, each designed to capture different aspects of narrative coherence. To ensure reliability and measure consistency, we employ K independent LLM agents ($K = 5$ for news and $K = 3$ for scientific articles) with variable temperature settings (0.3 to 0.7) and different random seeds. We note that we initially evaluated the news data set using $K = 5$ agents to ensure that the results were reliable and consistent between the different agents. We note that the resulting high inter-rater reliability (ICC(2,1) > 0.96) demonstrated that even single agents provide consistent evaluations. Based on this empirical evidence, we reduced the value of K to simply three agents for the VisPub data set to manage computational costs while maintaining statistical reliability. This decision was validated by similarly high ICC values (>0.92) for the scientific papers data set, which confirmed that three agents provide sufficient coverage for a reliable evaluation.

We note that our election of OpenAI's GPT-4.1 is based on empirical evidence demonstrating its effectiveness in narrative and coherence assessment tasks. Recent studies have shown that GPT-4 achieves human-level performance in evaluating discourse coherence [54], with agreement rates exceeding 80% with human judgments [55]. The G-Eval framework demonstrated that GPT-4 achieves Spearman correlations of 0.514 with human evaluations on creative text tasks [34], while maintaining high inter-rater reliability (ICC > 0.94) across evaluation periods [56]. While newer models may offer incremental improvements, comparative studies show that GPT-4 already operates within the range of human inter-rater reliability for adjacent narrative assessment tasks [14], making further improvements marginal for our purposes.

Furthermore, the model's training data (through 2021) encompasses our evaluation period, ensuring familiarity with both news events (2016–2021) and scientific literature evolution (1990–2022). Critically, OpenAI's stable API availability addresses reproducibility concerns that plague rapidly evolving language model research [57]. By establishing our methodology with GPT-4, we enable direct comparison with future work, as newer models can be evaluated against this baseline. Finally, we note that our objective is not to use the latest state-of-the-art models to maximize evaluation performance, but to demonstrate that LLM-as-a-judge approaches can serve as effective proxies for mathematical coherence metrics in narrative extraction. In this context, GPT-4's documented capabilities are more than sufficient to achieve this goal [58,59].

3.4. Evaluation Modes and Prompt Design Rationale

Our four evaluation modes were designed following principles derived from both narrative theory and LLM evaluation best practices. Each mode tests different hypotheses about what constitutes effective narrative evaluation.

3.4.1. Simple Evaluation

The **simple evaluation** tests whether minimal instruction suffices for narrative assessment. By requesting only a coherence score without additional guidance, we establish a baseline that relies entirely on the model's pre-trained understanding of narrative quality. This approach minimizes prompt engineering effects and computational costs while testing whether sophisticated evaluation frameworks are necessary.

In particular, the simple mode requests a single coherence score (1 to 10) based on the way the articles flow together. The prompt emphasizes the general flow of the narrative without requiring detailed analysis, making it the most efficient approach. We show the prompt used in this case in Figure 2.

Simple Evaluation Prompt

```
Rate the coherence of this narrative sequence on a scale of 1-10.

Narrative:
{chr(10).join(f'{j+1}. {title}' for j, title in enumerate(narrative))}

Provide a single coherence score (1-10) based on how well the articles flow
    together.
```

Figure 2. Simple evaluation prompt requesting a single coherence score.

3.4.2. Chain-of-Thought (CoT) Evaluation

The **Chain-of-Thought** approach applies reasoning transparency principles to narrative evaluation. By explicitly decomposing the evaluation into steps that examine consecutive connections, overall flow, and jarring transitions, we test whether structured reasoning improves correlation with mathematical coherence. This design follows evidence that step-by-step analysis improves LLM performance on complex tasks while providing interpretable reasoning traces.

In particular, the CoT mode guides the model through systematic analysis by examining connections between consecutive articles, assessing the general narrative flow, identifying jarring transitions, and synthesizing these observations into a final score. This structured approach aims to improve the quality of the evaluation through explicit reasoning. We show the prompt used in this case in Figure 3.

Chain-of-Thought (CoT) Evaluation Prompt

```
Rate the coherence of this narrative sequence on a scale of 1-10.

Narrative:
{chr(10).join(f'{j+1}. {title}' for j, title in enumerate(narrative))}

Let's evaluate this step by step:

1. First, examine how each article connects to the next one. Are there
    clear logical or thematic links between consecutive articles?

2. Next, consider the overall flow. Does the sequence tell a coherent story
    or follow a logical progression from beginning to end?

3. Then, check for any jarring transitions or articles that seem out of
    place in the sequence.

4. Finally, based on your analysis above, provide a single coherence score
    (1-10) where:
   - 1-3 = Very poor coherence (random/disconnected articles)
   - 4-6 = Moderate coherence (some connections but gaps exist)
   - 7-9 = Good coherence (clear narrative flow with minor issues)
   - 10 = Excellent coherence (seamless narrative progression)

Think through each step before providing your final score.
```

Figure 3. Chain-of-Thought evaluation prompt with structured reasoning steps.

3.4.3. Standard Evaluation

The **standard evaluation** operationalizes narrative quality through four fundamental dimensions drawn from narrative theory: logical flow (causal connections), thematic consistency (topical coherence), temporal coherence (chronological ordering), and narrative completeness (story arc). These dimensions map to key components of the mathematical coherence metric while remaining interpretable to non-technical users.

In particular, the standard mode evaluates four specific dimensions: logical flow (how well articles connect logically), thematic consistency (the consistency of themes throughout), temporal coherence (whether events follow a reasonable timeline), and narrative completeness (whether the sequence tells a complete story). The overall score averages these dimensions. We show the prompt used in this case in Figure 4.

Standard Evaluation Prompt

```
You are evaluating the coherence of a sequence of news articles.

Rate the following narrative on these criteria (1-10 scale):
1. Logical Flow: How well do the articles connect logically?
2. Thematic Consistency: How consistent are the themes throughout?
3. Temporal Coherence: Do the events follow a reasonable timeline?
4. Narrative Completeness: Does the sequence tell a complete story?

Narrative sequence:
{chr(10).join(f'{j+1}. {title}' for j, title in enumerate(narrative))}

BE FAIR BUT CRITICAL. A score of 5 means "adequate", 7 means "good", 10
    means "exceptional".
Most narratives should score between 4-7.
```

Figure 4. Standard evaluation prompt assessing four specific dimensions. Note that for the VisPub data set, "news articles" is replaced with "scientific articles".

3.4.4. Complex Evaluation

The **complex evaluation** extends assessments to six fine-grained criteria, testing whether additional granularity improves evaluation quality. The criteria—semantic coherence, topic evolution, information density, causal relationships, narrative arc, and contextual relevance—capture nuanced aspects of narrative quality that might be overlooked in simpler evaluations.

In particular, the complex mode assesses six detailed criteria: semantic coherence (conceptual flow between articles), topic evolution (the smoothness of topic transitions), information density (the appropriateness of pacing), causal relationships (the clarity of cause–effect connections), narrative arc (the completeness of story structure), and contextual relevance (the relevance of each article to the overall narrative). We show the prompt used in this case in Figure 5.

Complex Evaluation Prompt

```
Evaluate this narrative sequence on multiple detailed criteria.

Narrative:
{chr(10).join(f'{j+1}. {title}' for j, title in enumerate(narrative))}

Rate each criterion (1-10):
1. Semantic Coherence: How well do concepts flow between articles?
2. Topic Evolution: How smoothly do topics transition and evolve?
3. Information Density: Is the progression appropriately paced?
4. Causal Relationships: Do articles show clear cause-effect relationships?
5. Narrative Arc: Does the sequence form a complete story arc?
6. Contextual Relevance: How relevant is each article to the overall
    narrative?

BE CRITICAL. Random sequences should score low (2-4), coherent sequences
    moderate (5-7), exceptional sequences high (8-10).
```

Figure 5. Complex evaluation prompt with six detailed criteria for assessment.

3.4.5. Experimental Design

Our experimental design inherently provides both ablation analysis and robustness testing. The four evaluation modes (simple, Chain-of-Thought, standard, and complex) constitute a systematic ablation study where each mode progressively adds evaluation complexity. This allows us to isolate the contribution of different prompt components: from minimal instruction (simple) to structured reasoning (CoT) to multi-dimensional assessment (standard and complex). Additionally, our multi-agent approach with K independent evaluators using different temperature settings (0.3 to 0.7) and random seeds provides robustness testing across model parameters, ensuring our results are not artifacts of specific sampling configurations. This progression from simple to complex allows us to empirically determine the optimal trade-off between evaluation complexity and practical efficiency.

3.4.6. Chain-of-Thought Usage

We note that Chain-of-Thought reasoning could be applied to the standard and complex evaluation modes, requiring step-by-step analysis for each dimension. However, we decided against this combination for several reasons. First, the standard and complex modes already provide structured decomposition through their multiple evaluation criteria, achieving similar benefits to Chain-of-Thought by guiding the model through specific aspects of narrative quality. Adding explicit reasoning steps for each dimension would introduce redundancy rather than additional insight. Second, the computational cost would increase substantially, requiring 4–6 times more tokens for complete reasoning traces across all dimensions. Third, maintaining these as separate evaluation modes provides clearer experimental comparisons, allowing us to isolate the effects of structured reasoning (CoT) versus multi-dimensional assessment (standard/complex). Our results validate this decision, as the standard and complex modes achieve strong performance without explicit reasoning steps, suggesting that dimensional decomposition alone provides a sufficient evaluation structure.

3.4.7. Score Calibration

We note that we intentionally varied the score anchoring guidance across evaluation modes to test the robustness of LLM evaluation to different scoring frameworks. The Chain-of-Thought mode defines moderate coherence as scores 4–6, while the complex mode

shifts this range to 5–7. Similarly, what constitutes "very poor" versus "exceptional" narratives receives different numerical mappings across modes. This variation tests whether LLM evaluators can maintain consistent quality assessments despite different calibration instructions, which is a critical concern for practical deployment where users might provide varying scoring guidelines.

Our results validate this approach: despite different score anchoring, all evaluation modes achieve high inter-rater reliability (ICC > 0.92) and significant correlations with mathematical coherence. The multi-agent design further demonstrates that independent evaluators converge on consistent quality gradients regardless of specific numerical guidance. This robustness to prompt variations suggests that LLMs assess narrative quality based on learned representations rather than merely following numerical instructions, making the approach resilient to variations in evaluation setup.

Furthermore, we note that the evaluation modes intentionally employ different scoring guidance strategies, which serve as a test of the robustness of LLM evaluation to different point anchoring strategies:

- **Simple evaluation**: provides no explicit score definitions, relying entirely on the model's pre-trained understanding of a 1–10 scale.
- **Chain-of-Thought**: defines explicit ranges (1–3 for very poor, 4–6 for moderate, 7–9 for good, and 10 for excellent).
- **Standard evaluation**: offers point anchors (5 = adequate, 7 = good, and 10 = exceptional) without defining ranges.
- **Complex evaluation**: uses behavioral descriptions (e.g., "random sequences should score low").

This diversity, which includes cases with no explicit guidance, ranges, and point estimates, tests whether a coherent evaluation emerges from the model's understanding rather than rigid adherence to numerical instructions. Despite these variations, all modes achieve high inter-rater reliability (ICC > 0.92) and significant correlations with mathematical coherence, demonstrating that LLMs can maintain consistent quality assessment across different calibration approaches. This robustness is crucial for practical deployment, where users may provide varying levels of scoring guidance.

4. Experimental Setup

4.1. Data Sets

We evaluated our methods on two data sets with contrasting characteristics to test the generalizability of LLM-as-a-judge approaches. The news data set [5,22] contains 540 articles spanning 2016 to 2021, mainly focused on Cuban protests and US–Cuba relations. Although the data set covers a 5-year period, the majority of articles concentrate on Cuban protest events that occurred in 2021, creating a temporally focused narrative around a specific political event. This provides a thematically constrained evaluation scenario with a clear temporal progression. Additionally, the data set includes 40 articles about COVID-19, which, while unrelated to the Cuban protests, introduce controlled topical diversity that allows us to test the robustness of our approach when narratives must bridge between related political events and pandemic coverage.

The VisPub data set [60] contains 3549 scientific papers from IEEE VIS, EuroVis, and other visualization conferences that span 1990–2022. After removing papers without abstracts, we analyze 3549 papers across 171 automatically discovered topics. This data set represents diverse research areas that include scientific visualization, information visualization, visual analytics, and human–computer interaction. The 30-year span captures the evolution of the field from early volume rendering work to modern machine learning applications in visualization.

4.2. Implementation Details

Document embeddings are generated using OpenAI's `text embedding-3-small` model, which provides 1536-dimensional representations. For narrative landscape construction, we use UMAP for dimensionality reduction (48 components) followed by HDBSCAN for topic discovery. The sparse coherence graph is constructed with the critical coherence (minimum spanning tree bottleneck weight) as the threshold, ensuring graph connectivity while removing weak connections.

For the news data set, we impose temporal constraints ensuring that edges only connect earlier to later documents. The scientific paper data set uses the publication year for temporal ordering. All experiments use consistent random seeds for reproducibility, and the results are cached to enable follow-up analyses.

To ensure structured and consistent evaluations, we employed OpenAI's function-calling feature with predefined schemas. Figure 6 shows the schema for the *simple* evaluation, while Figure 7 presents the *standard* evaluation schema. The *Chain-of-Thought* schema includes special fields for step-by-step reasoning (`step1_connections`, `step2_overall_flow`, `step3_transitions`) before the final score, but is otherwise similar to the simple approach. The *complex* schema extends the standard evaluation to six evaluation criteria, as detailed in the corresponding prompt.

Simple Evaluation Function Schema

```
{
    "name": "evaluate_simple_coherence",
    "description": "Evaluate narrative coherence with a simple overall
    score",
    "parameters": {
        "type": "object",
        "properties": {
            "coherence_score": {
                "type": "integer",
                "description": "Overall coherence score (1-10)",
                "minimum": 1,
                "maximum": 10
            },
            "explanation": {
                "type": "string",
                "description": "Brief explanation of the score"
            }
        },
        "required": ["coherence_score"]  // "explanation" added when
    required
    }
}
```

Figure 6. Function schema for simple evaluation mode. For our experiments, `require_explanations` was set to `False` to reduce computational costs.

Standard Evaluation Function Schema

```
{
    "name": "evaluate_standard_coherence",
    "description": "Evaluate narrative coherence on multiple dimensions",
    "parameters": {
        "type": "object",
        "properties": {
            "logical_flow": {
                "type": "integer",
                "description": "Score for logical flow (1-10)",
                "minimum": 1,
                "maximum": 10
            },
            "thematic_consistency": {
                "type": "integer",
                "description": "Score for thematic consistency (1-10)",
                "minimum": 1,
                "maximum": 10
            },
            "temporal_coherence": {
                "type": "integer",
                "description": "Score for temporal coherence (1-10)",
                "minimum": 1,
                "maximum": 10
            },
            "narrative_completeness": {
                "type": "integer",
                "description": "Score for narrative completeness (1-10)",
                "minimum": 1,
                "maximum": 10
            },
            "overall_coherence": {
                "type": "number",
                "description": "Average coherence score"
            },
            "explanation": {
                "type": "string",
                "description": "Brief explanation of the ratings"
            }
        },
        "required": ["logical_flow", "thematic_consistency",
                     "temporal_coherence", "narrative_completeness",
                     "overall_coherence"]  // "explanation" added when
    required
    }
}
```

Figure 7. Function schema for standard evaluation mode. For our experiments, `require_explanations` was set to `False` to reduce computational costs.

5. Results

5.1. Extracted Narrative Lengths

The narratives extracted exhibit distinct characteristics that reflect both the underlying document collections and the extraction methods used. Table 1 presents the path length statistics for both data sets and extraction methods.

Table 1. Path length statistics by data set and extraction method.

Data Set	Method	Path Length			
		Mean	**Std**	**Min**	**Max**
News	Max capacity	4.70	2.03	2	12
	Random chronological	3.97	1.80	2	8
VisPub	Max capacity	5.91	4.15	2	20
	Random chronological	5.64	3.56	2	16

Although we designed random chronological path generation to match the length distribution of maximum capacity paths following a normal distribution, in practice, this goal was only achieved for the VisPub data set (difference of 0.27, $p = 0.62$). For the news data set, the random paths were significantly shorter (difference of 0.73, $p < 0.001$), though this represents less than one document on average. Given that path lengths are discrete positive integers, this means that random paths in the news data set typically contain one fewer document than their maximum capacity counterparts.

Despite this length discrepancy, the comparison remains informative. In fact, the shorter random paths in the news data set could theoretically have an advantage in maintaining coherence, yet maximum capacity paths still achieve substantially higher coherence scores. VisPub narratives are on average 25 to 44% longer than news narratives depending on the method, reflecting the greater temporal span and topical diversity of the scientific literature collection.

5.2. Overall Performance Comparison

Figure 8 presents the analysis of the LLM judge performance in both data sets. Several key patterns emerge from this comparison. In general, both data sets show clear discrimination between maximum capacity and random paths, with stronger correlations observed in the diverse VisPub data set. In detail, Figure 8 shows the following key results:

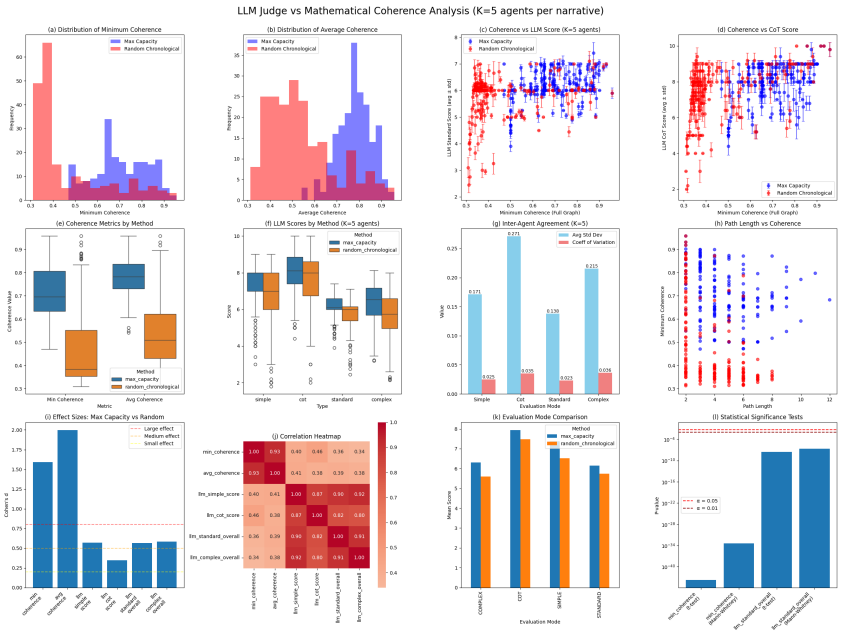
(a) The distribution of minimum coherence values comparing the maximum capacity paths and random chronological paths, showing the frequency of minimum coherence scores for each path extraction method.
(b) The distribution of the average coherence values comparing the maximum capacity paths and random chronological paths, showing the frequency of mean coherence scores along entire paths for each path extraction method.
(c) Scatter plots with error bars showing the relationship between minimum coherence and LLM judge evaluation scores using the simple approach, with error bars representing the standard deviation in the evaluation of the K agents.
(d) Scatter plots with error bars displaying the relationship between minimum coherence and LLM judge evaluation scores using the Chain-of-Thought approach, with error bars representing the standard deviation in the evaluation of the K agents.
(e) Box plots comparing the distribution of minimum and average coherence metrics between max capacity and random chronological path extraction methods.
(f) Box plots showing the distribution of LLM evaluation scores across different evaluation modes (simple, CoT, standard, and complex) for both path extraction methods.
(g) Bar charts displaying inter-agent agreement metrics, showing the average standard deviation and coefficient of variation across K agents for each evaluation mode.
(h) Scatter plots examining the relationship between path length (the number of documents) and minimum coherence, color-coded by path extraction method.

(i) Bar charts showing Cohen's d effect sizes for each metric, comparing the ability to discriminate between the maximum capacity and random paths, with reference lines for small, medium, and large effects.

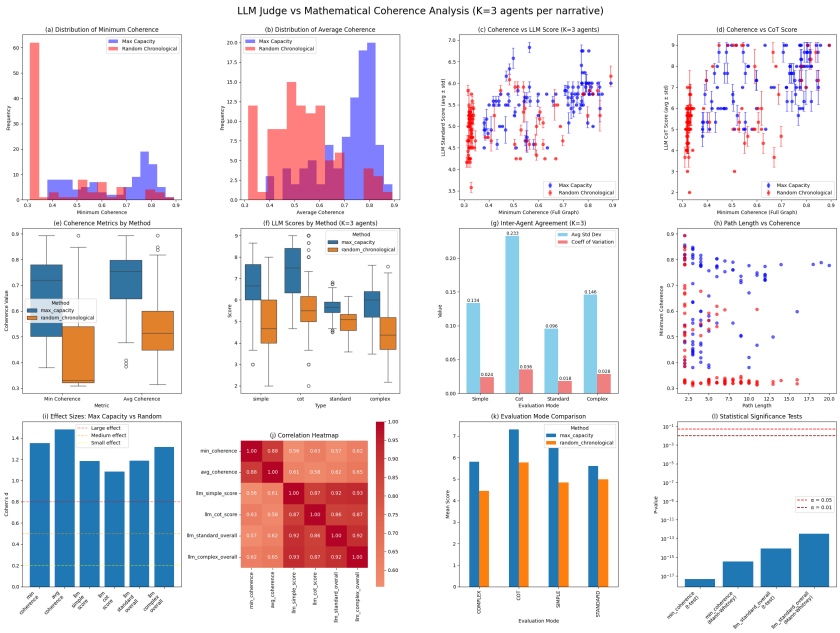
(j) Heat maps showing Pearson correlation coefficients between mathematical coherence metrics and LLM evaluation scores in different modes.

(k) Grouped bar charts comparing mean LLM evaluation scores between path extraction methods across different evaluation modes (simple, CoT, standard, and complex).

(l) Bar graphs showing p-values of the statistical significance tests (t-test and Mann–Whitney U) for each metric, with reference lines at significance levels $\alpha = 0.05$ and $\alpha = 0.01$.



A News data set results.



B VisPub data set results.

Figure 8. Analysis of LLM judge performance versus mathematical coherence.

The distribution of mathematical coherence values differs substantially between data sets. The coherence distribution of the news data set shows that maximum capacity paths are concentrated around 0.70–0.75 minimum coherence, while random paths cluster around 0.47. The VisPub data set exhibits broader distributions, with maximum capacity paths ranging from 0.4 to 0.9 and random paths showing much lower minimum coherence values in general, as shown by the high frequency of 0.30 minimum coherence paths. This difference reflects the focused nature of the news content versus the topical diversity of scientific papers.

The average coherence metrics reveal consistent patterns in both data sets. Max capacity paths achieve mean minimum coherence of 0.705 (news) and 0.648 (papers), compared to 0.470 and 0.433 for random paths, respectively. The effect sizes (Cohen's d) for mathematical coherence are 1.59 (news) and 1.35 (papers), indicating strong discrimination between path types.

5.3. Correlation Analysis

Table 2 reveals the underlying differences between the data sets. The VisPub data set consistently shows stronger correlations between LLM evaluations and mathematical coherence, with the complex evaluation mode achieving $r = 0.65$ with average coherence. The news data set peaks at $r = 0.46$ using the Chain-of-Thought evaluation approach with minimum coherence.

Table 2. Pearson correlation between LLM evaluation modes and mathematical coherence metrics across data sets. Note that for n we count the number of endpoint pairs times the number of path types (random vs. maximum capacity). Spearman correlations showed similar patterns.

Evaluation Mode	News Data Set ($n = 400$)		VisPub Data Set ($n = 200$)	
	Min Coh.	Avg Coh.	Min Coh.	Avg Coh.
Simple	0.40	0.41	0.56	0.61
CoT	0.46	0.38	0.63	0.58
Standard	0.36	0.39	0.57	0.62
Complex	0.34	0.38	0.62	0.65

All values are statistically significant with $p < 0.001$. Min Coh. and Avg Coh. refer to minimum and average coherence along paths.

These differences suggest that LLM judges perform better in diverse, heterogeneous content where they can leverage broader knowledge to identify thematic connections. The focused nature of the news data set, covering a single topic over a short time span, may limit the discriminative signals available to LLM evaluators, as even a random sample might suffice to describe a coherent narrative when there is not much freedom of choice.

5.4. Inter-Agent Agreement

Our experimental design used $K = 5$ agents for the news data set and $K = 3$ for the VisPub data set. We note that this difference arose from our sequential experimental approach: initial results from the news data set demonstrated high inter-rater reliability (ICC(2,1) = 0.961), indicating that fewer agents would suffice for reliable evaluation. Therefore, we reduced the number of agents to $K = 3$ for the computationally more expensive VisPub data set. We note that this decision was validated by consistent high reliability metrics (ICC(2,1) > 0.92) across both data sets. These results suggest that even K = 3 provides more than adequate statistical power for the narrative evaluation based on LLMs.

In general, Table 3 demonstrates high inter-rater reliability in all evaluation modes. The values of ICC(2,K) exceed 0.97 for all conditions, indicating that the average of the K agents provides highly reliable scores. ICC(2,1) values all exceed 0.92 in all conditions, thus

showing that even if we followed a single-agent approach, the LLM-as-a-judge approach to evaluate narrative extraction has remarkably high consistency. Krippendorff's alpha values above 0.92 confirm substantial agreement even when accounting for chance.

Table 3. Inter-rater reliability metrics across evaluation modes.

Metric	News Data Set ($K = 5$)			VisPub Data Set ($K = 3$)		
	Simple	**Standard**	**Complex**	**Simple**	**Standard**	**Complex**
ICC(2,1)	0.961	0.924	0.952	0.963	0.924	0.963
ICC(2,*K*)	0.992	0.984	0.990	0.987	0.973	0.987
Krippendorff's α	0.961	0.924	0.952	0.962	0.924	0.963
Kendall's W	0.912	0.799	0.816	0.925	0.841	0.863
Mean abs. dev.	0.150	0.119	0.183	0.126	0.089	0.134
Exact agreement	81.2%	61.2%	25.6%	81.2%	58.2%	29.7%
Within-1 agreement	99.8%	100%	98.6%	99.8%	100%	99.5%

The simple evaluation mode achieves the highest exact agreement rates (>81%), while the complex modes show lower exact agreement but maintain near-perfect *within-1 point* agreement. This pattern suggests that, while agents may differ in precise scores in complex evaluations, they consistently identify the same quality gradients.

For practitioners, our analysis of results suggests that $K = 3$ agents represent an optimal balance between reliability and computational cost, as the marginal improvement from additional agents is negligible given the ICC(2,K) values already exceed 0.97. Indeed, the ICC(2,1) values above 0.92 indicate that even single-agent evaluation could be sufficient for many practical applications, offering a cost-effective alternative without substantial reliability loss.

5.5. Discrimination Between Path Types

Table 4 reveals that LLM evaluations successfully discriminate between high-quality and baseline narratives, although with varying effectiveness between data sets. For the VisPub data set, the complex LLM evaluation achieves discrimination ($d = 1.316$) that is close to that of mathematical coherence ($d = 1.353$). The news data set shows weaker discrimination for all LLM modes (d = 0.35–0.58), suggesting that focused, temporally constrained narratives pose greater challenges for LLM evaluation.

Table 4. Effect sizes (Cohen's *d*) for discriminating between max capacity and random paths.

Metric	News Data Set		VisPub Data Set	
	Cohen's *d*	**AUC-ROC**	**Cohen's *d***	**AUC-ROC**
Min. coherence	1.591	0.852	1.353	0.834
Avg. coherence	1.998	0.891	1.481	0.840
LLM simple	0.568	0.660	1.183	0.795
LLM CoT	0.345	0.594	1.083	0.777
LLM standard	0.566	0.650	1.188	0.798
LLM complex	0.582	0.663	1.316	0.823

The **AUC-ROC** values confirm these patterns, with VisPub achieving values near 0.80 for most LLM modes, comparable to mathematical metrics. The news data set shows moderate discrimination (AUC 0.59–0.66), still significantly better than chance but indicating room for improvement.

5.6. Cost–Performance Trade-Offs

Table 5 presents detailed computational requirements for each evaluation mode based on smaller-scale experiments with the VisPub data set using 60 narratives evaluated with a single agent ($K = 1$) to isolate the per-evaluation costs. The narratives were generated by taking 30 endpoint pairs and generating narratives using both the max capacity and the random path approaches. Token counts were measured using OpenAI's `tiktoken` library and include both the evaluation prompt and the narrative content (article titles formatted as a numbered list).

The simple evaluation approach uses approximately 123 input tokens per narrative, while the complex evaluation approach requires approximately 225 tokens. Token usage for Chain-of-Thought evaluation is highest at 279 tokens due to the structured reasoning steps. While timing measurements show low variance (0.016 to 0.114 s standard deviation), token usage exhibits higher variability ($\pm$53.9 tokens) due to the diverse narrative lengths in our sample (5.6 ± 3.4 documents).

The variability in our results reflects real-world usage patterns, where narratives naturally vary in length. The 60-narrative sample captures this distribution adequately, as larger samples would show similar variance patterns rather than converging to a tighter estimate. The consistent mean token differences between the evaluation modes (123 for simple vs. 279 for CoT) demonstrate that the sample size is sufficient to distinguish computational requirements between approaches. We note that these measurements reflect input tokens only and do not include the model's response tokens.

Table 5. Computational resource requirements for different evaluation methods on the VisPub data set (n = 60 narratives).

Method	Time per Narrative (s)	Tokens per Narrative
LLM simple	0.647 ± 0.082	123.0 ± 53.9
LLM CoT	6.001 ± 0.109	279.0 ± 53.9
LLM standard	0.867 ± 0.016	214.0 ± 53.9
LLM complex	1.034 ± 0.114	225.0 ± 53.9

The relationship between cost and performance is non-linear. For the VisPub data set, where our computational measurements were conducted, the progression from the simple evaluation approach ($r = 0.61$) to the complex evaluation approach ($r = 0.65$) represents a 7% improvement at 83% increased token usage. Chain-of-Thought evaluation shows particularly poor cost-effectiveness, requiring 127% more tokens than simple evaluation while achieving a lower correlation ($r = 0.58$). Although we did not measure computational costs on the news data set, the correlation patterns observed suggest even weaker cost–performance benefits from increased complexity. While Chain-of-Thought achieves the highest correlation for news ($r = 0.46$) compared to simple evaluation ($r = 0.41$), this 12% improvement would come at the cost of 127% more tokens if token usage patterns are similar across data sets, which is a reasonable assumption given comparable prompt structures. Complex evaluation performs worst on news ($r = 0.38$), suggesting that increased complexity does not guarantee better performance.

Beyond token usage, we note that the temporal costs vary substantially across evaluation modes. Chain-of-Thought evaluation requires 6.001 s per narrative, approximately 9.3 times longer than simple evaluation (0.647 s), while achieving mixed results across data sets. Standard and complex evaluations require 0.867 and 1.034 s, respectively, representing 34% and 60% increases over the simple evaluation. These timing measurements used a single agent ($K = 1$) to isolate the per-evaluation costs; our main experiments used $K = 5$ agents for news and $K = 3$ for VisPub, multiplying these costs proportionally.

The practical implications become apparent at scale. Evaluating 1K narratives with a single agent would require approximately 11 min with simple evaluation versus 1.7 h with Chain-of-Thought. Our main experiments, which evaluated 600 narratives across four evaluation modes with multiple agents, required approximately 12 h of execution time in total. However, we note that this includes network latency, API rate limiting, and data processing overhead.

In theory, using only the simple evaluation mode would reduce this to approximately 1.5 to 2 h. For studies processing larger narrative collections, the selection of the evaluation mode significantly impacts computational feasibility. These findings suggest that simple evaluation provides the best balance between computational efficiency and evaluation quality, achieving 85–90% of complex approach performance while requiring minimal resources.

5.7. Out-of-Distribution Robustness

To test whether our approach correctly identifies adversarial narratives, we created 60 out-of-distribution (OOD) narratives by mixing documents from news articles and scientific papers. We generated two types of adversarial sequences: (1) alternating narratives that switch between domains at every step, and (2) random mixed narratives that select documents randomly from the combined corpus. In particular, we generated 30 OOD narratives of each type for our evaluation.

Both OOD narrative types received significantly lower scores than typical narratives in our main experiments. Alternating narratives achieved mean LLM scores of 1.99 ± 0.35 and mathematical coherence of 0.303 ± 0.009, while random mixed narratives scored slightly higher at 2.41 ± 0.69 and 0.337 ± 0.108, respectively. These scores fall well below the coherent narrative baselines observed in our main experiments (typically 6–8 for LLM scores and 0.6–0.8 for mathematical coherence).

Notably, the correlation between mathematical coherence and LLM scores remains strong even for these adversarial examples ($r = 0.729$, $p < 0.0001$), demonstrating that the LLM-as-a-judge approach maintains its validity outside normal operating conditions. The significant difference between alternating and random mixed narratives ($t = -2.93$, $p = 0.0048$, Cohen's $d = -0.77$) suggests that LLMs can distinguish between different types of incoherence, with strictly alternating domain switches being recognized as particularly disruptive to narrative flow.

These results confirm that our approach exhibits appropriate behavior when faced with adversarial inputs, correctly identifying incoherent narratives regardless of how they were constructed. Thus, the ability of the proposed LLM-as-a-judge approach to maintain a strong correlation with mathematical metrics while providing interpretable assessments in an adversarial context provides further support for its use in practical applications.

5.8. Follow-Up Analyses

Figure 9 summarizes extensive follow-up analyses on the main results that provide a deeper understanding of the evaluation behavior of LLMs. In detail, Figure 9 shows the following results:

(a) A bar graph comparing the performance of predictive modeling (R^2 scores) obtained by different regression models to predict LLM evaluation scores from mathematical coherence metrics.

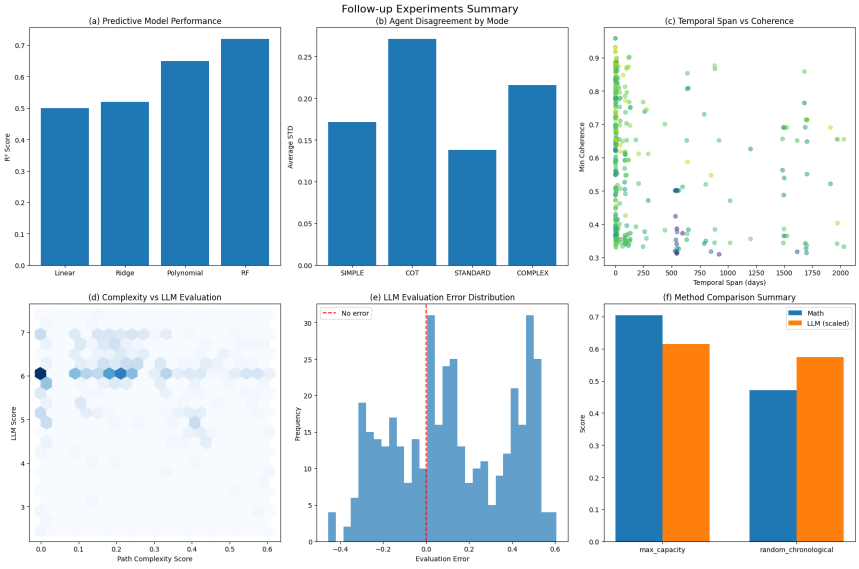
(b) A bar chart displaying the average standard deviation of agent evaluations across different evaluation modes, illustrating the level of inter-agent disagreement for each approach.

(c) A scatter plot examining the relationship between temporal span (in days/years) and minimum coherence, with points colored by LLM standard evaluation scores to show how temporal distance affects both mathematical and LLM-based coherence assessments.

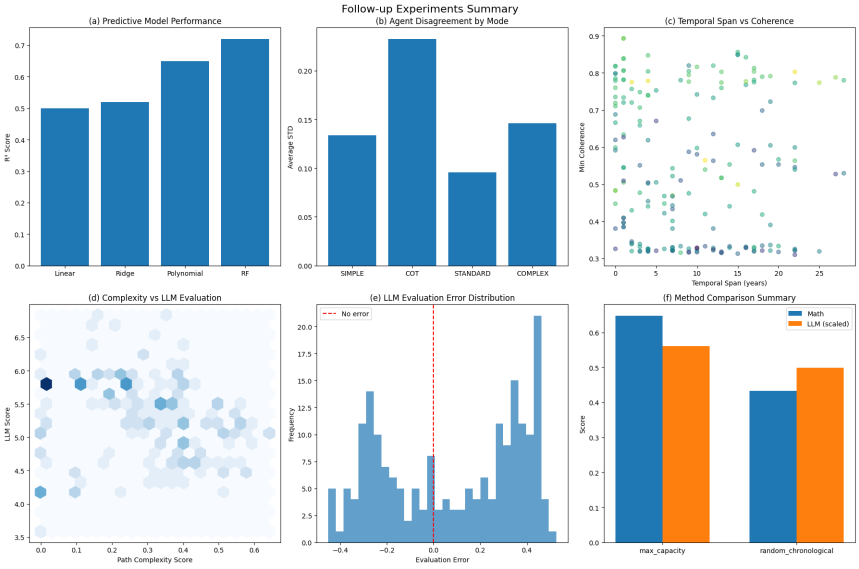
(d) A hexbin density plot showing the relationship between path complexity scores (derived from coherence variance, path efficiency, and weak link ratio) and LLM standard evaluation scores, revealing how narrative complexity impacts evaluation.

(e) A histogram showing the distribution of evaluation errors (the difference between normalized LLM scores and normalized mathematical coherence), with a vertical reference line at zero indicating perfect agreement.

(f) A grouped bar chart comparing mean mathematical coherence and mean LLM evaluation scores (scaled) between max capacity and random chronological path generation methods, providing a summary comparison of method performance.



A News data set follow-up experiments.



B VisPub data set follow-up experiments.

Figure 9. Summary of follow-up experiments including predictive modeling, agent disagreement analysis, temporal effects, and error patterns. The VisPub data set shows stronger predictive performance and clearer temporal patterns.

5.8.1. Predictive Modeling

Random forest models that predict LLM scores from mathematical coherence metrics achieve R^2 values of 0.40 to 0.55 for the VisPub data set, compared to 0.29 to 0.41 for news data. A feature importance analysis reveals that *average coherence* along the path contributes more to predictions (52–62% importance) than *minimum coherence* (27–36%) or *path length* (8–13%). This suggests that LLMs consider the overall narrative flow rather than focusing solely on the weakest links. Table 6 shows the performance of the evaluated models and Table 7 shows the analysis of the importance of the characteristics for the best model (random forests).

Table 6. Predictive modeling performance (R^2 scores) for predicting LLM scores from mathematical coherence metrics.

Model	News Data Set				VisPub Data Set			
	Simple	**CoT**	**Standard**	**Complex**	**Simple**	**CoT**	**Standard**	**Complex**
Linear	0.188 (±0.034)	0.188 (±0.071)	0.188 (±0.049)	0.208 (±0.062)	0.327 (±0.101)	0.356 (±0.085)	0.348 (±0.115)	0.421 (±0.100)
Ridge	0.186 (±0.032)	0.176 (±0.066)	0.187 (±0.051)	0.206 (±0.065)	0.328 (±0.089)	0.358 (±0.080)	0.350 (±0.101)	0.420 (±0.096)
Random forest	0.332 (±0.115)	0.315 (±0.075)	0.350 (±0.110)	0.290 (±0.121)	0.406 (±0.147)	0.410 (±0.149)	0.393 (±0.121)	0.554 (±0.078)
Polynomial	0.166 (±0.059)	0.168 (±0.050)	0.164 (±0.075)	0.184 (±0.089)	0.277 (±0.128)	0.328 (±0.083)	0.312 (±0.117)	0.398 (±0.087)

Table 7. Feature importance in random forest models for each evaluation mode.

Evaluation Mode	News Data Set			VisPub Data Set		
	Min. Coh.	**Avg. Coh.**	**Path Len.**	**Min. Coh.**	**Avg. Coh.**	**Path Len.**
Simple	0.602	0.315	0.083	0.312	0.562	0.127
CoT	0.636	0.282	0.082	0.517	0.373	0.110
Standard	0.542	0.340	0.117	0.268	0.621	0.111
Complex	0.469	0.402	0.129	0.366	0.526	0.108

5.8.2. Agent Disagreement Patterns

Agent disagreement varies systematically with score ranges and evaluation modes. The Chain-of-Thought evaluation approach shows the highest disagreement, with a mean standard deviation of 0.27 for news articles and 0.23 for scientific articles. In contrast, the simple evaluation approach maintains the lowest variations with standard deviations of 0.17 and 0.13, respectively. Disagreement peaks for narratives that receive medium scores (5–7 range) and decreases for very high or very low scores, suggesting greater consensus on clearly good or bad narratives. Table 8 summarizes the results of agent disagreement patterns.

Table 8. Agent disagreement analysis summary.

Evaluation Mode	News Data Set		VisPub Data Set	
	Avg. STD	**CoV**	**Avg. STD**	**CoV**
Simple	0.171	0.025	0.134	0.024
CoT	0.271	0.035	0.233	0.036
Standard	0.138	0.023	0.096	0.018
Complex	0.215	0.036	0.146	0.028

5.8.3. Temporal Analysis

Temporal span shows minimal correlation with coherence for both mathematical metrics and LLM evaluations. We note that news data spans are measured in days and

publication data spans are measured in years. The news data set does not show significant temporal effects. In contrast, the VisPub data set reveals slight negative correlations ($r = -0.08$ to -0.14) between the temporal span and the coherence scores, indicating that narratives that span many years face modest coherence challenges. However, optimal coherence occurs for moderate temporal spans (3–5 years), balancing topical evolution with maintaining connections.

5.8.4. Path Complexity Effects

An analysis of path complexity beyond simple length reveals interesting patterns. The *coherence variance* along paths correlates negatively with LLM scores ($r = -0.22$ to -0.30), indicating that consistent quality throughout the narrative matters more than achieving high peak coherence. The *weak link ratio* (edges below median coherence) shows the strongest negative correlation with LLM scores ($r = -0.30$), confirming that LLMs penalize narratives with notable weak points.

5.9. Cross-Validation Stability

Five-fold cross-validation confirms the stability of our findings. The correlation between mathematical coherence and LLM evaluations shows minimal differences between train and test sets in all folds (mean absolute difference < 0.05), indicating a stable generalization. The effect sizes for the discriminating path types remain stable across folds, with standard deviations below 0.2 for Cohen's d values. These results indicate that our findings generalize well within each data set and are not artifacts of particular endpoint selections.

6. Discussion

6.1. Data Set Characteristics and LLM Performance

The substantial performance difference between the data sets (maximum $r = 0.65$ for VisPub versus $r = 0.46$ for news) illuminates when the LLM-as-a-judge approach excels. The VisPub data set spans 30 years and 171 topics, providing a rich semantic space for LLMs to identify meaningful connections. Topics range from early volume rendering research to modern applications of machine learning in visualization, offering diverse conceptual bridges for narrative construction.

In contrast, the focus on the Cuban protests within a single year constrains the semantic space associated with the news data set. Although this provides a clear temporal narrative structure, it can limit the distinctiveness of features that LLMs use for quality assessment. The bimodal distribution of coherence scores in the news data set suggests that even random paths within this focused collection maintain reasonable coherence, making discrimination more challenging.

6.2. Evaluation Mode Effectiveness

The Chain-of-Thought evaluation shows interesting data-set-dependent behavior, achieving the best performance for both data sets in terms of *minimum coherence* (i.e., max capacity coherence). In particular, this method obtained a higher correlation in the scientific papers data set ($r = 0.63$) compared to the news data set ($r = 0.46$). However, in terms of *average coherence*, we see that the other approaches provide a better result in terms of correlations. The best approach for the news data set is the simple prompt and for the scientific papers data set is the complex prompt.

However, in general, we found that the simple evaluation achieves a performance of 85-90% of the more complex evaluation approaches—including CoT and complex versions of judge prompts—while requiring less computational resources, which has important practical implications. For large-scale narrative evaluation tasks, simple prompts provide

an efficient solution without substantial quality sacrifice. The marginal improvements from complex evaluation modes may not justify their computational costs for most applications.

6.3. Reliability and Consistency

The high inter-rater reliability (ICC > 0.96) across all conditions demonstrates that LLM evaluations are highly reproducible when using multiple agents with controlled variation. The near-perfect within-1 point agreement (>98%) indicates that while agents may differ in exact scores, they consistently identify quality gradients. This reliability is crucial for practical applications where consistent evaluation standards are required.

The lower exact agreement for complex evaluation modes (25–30%) compared to simple modes (>81%) suggests that increased evaluation complexity introduces more subjective interpretation. However, the maintained high correlation with mathematical metrics indicates that this variation occurs around consistent quality assessments rather than fundamental disagreements.

It should be noted that multi-agent systems show diminishing returns for most evaluation tasks [61]. In particular, as the number of agents increases, further coordination structures are required to obtain significant performance improvements [62].

6.4. Implications for Narrative Evaluation

Our results suggest different strategies for different content types. For diverse, long-spanning document collections, such as research papers, LLM-as-a-judge approaches provide effective proxies for mathematical coherence, potentially replacing complex calculations with interpretable assessments. For focused collections with limited topical variation, mathematical metrics may remain necessary to capture subtle coherence distinctions that LLMs struggle to identify.

The success of simple evaluation modes democratizes narrative assessment, making it accessible to practitioners without deep technical expertise. Rather than understanding embedding spaces and information theory, users can leverage LLM evaluations to assess narrative quality in natural language terms.

6.5. Limitations and Future Directions

Several limitations warrant consideration. Our evaluation uses a single LLM family (GPT-4), and performance may vary between different architectures. The English-only evaluation leaves unanswered questions about the multilingual narrative assessment. The two data sets, while contrasting and providing an appropriate baseline for news data and scientific papers data, do not represent all types of document collection.

Additionally, we do not propose new model architectures, training paradigms, or data collection strategies; instead, our contribution lies in demonstrating that existing LLMs can effectively evaluate narrative extraction without requiring specialized models or training data, which itself has practical value for immediate deployment.

Furthermore, although recent work by Chen et al. [18] has documented significant biases in LLM judges including position bias, verbosity bias, and self-enhancement bias, these limitations can be partially mitigated through multi-agent evaluation and careful prompt design.

Future work should explore performance across different LLM architectures, including open-source models that enable a deeper analysis of evaluation mechanisms. Multilingual evaluation presents unique challenges, as narrative coherence could manifest differently across languages and cultures. Domain-specific fine-tuning could potentially improve performance for specialized collections while maintaining general capabilities.

Hybrid approaches that combine mathematical and LLM-based metrics merit investigation. LLMs could provide interpretable explanations for mathematical coherence

scores, or mathematical metrics could guide LLM attention to critical narrative elements. Such combinations might achieve better performance than either approach alone while maintaining interpretability.

7. Conclusions

We demonstrate that LLM-as-a-judge approaches provide effective proxies for mathematical coherence metrics in narrative evaluation, with performance strongly influenced by data set characteristics. For diverse document collections spanning multiple topics and time periods, LLM evaluations achieve correlations up to 0.65 with mathematical coherence while successfully discriminating between algorithmically optimized and random narratives with effect sizes exceeding 1.0. For focused collections with limited topical variation, the correlations remain moderate (up to 0.46) but still significant.

Simple evaluation prompts emerge as the most practical choice, achieving 85–90% of complex mode performance with lower computational costs. Combined with high inter-rater reliability (ICC > 0.96) and near-perfect within-1 point agreement, this makes LLM evaluation accessible for large-scale narrative assessment tasks.

These findings have immediate practical applications for the digital humanities, investigative journalism, and reviews of the literature in scientific research, where understanding the narrative connections between document collections is crucial. By providing interpretable assessments without requiring technical expertise in embeddings or information theory, LLM-as-a-judge approaches democratize narrative evaluation while maintaining strong alignment with mathematical quality metrics. As LLM capabilities continue to advance, these evaluation approaches will likely become increasingly central to narrative understanding and knowledge discovery applications.

Funding: This work has been supported by ANID (National Research and Development Agency of Chile) FONDECYT de Iniciación en Investigación 2025 Grant 11250039 Project "Interactive Narrative Analytics: Developing scalable knowledge-based narrative extraction models and visual analytics systems for sensemaking in complex information landscapes." The author is also supported by Project 202311010033-VRIDT-UCN.

Data Availability Statement: All data is available at the Narrative Trails repository https://github.com/faustogerman/narrative-trails accessed on 1 June 2025.

Acknowledgments: During the preparation of this manuscript, the author used Grammarly and Writefull integrated with Overleaf for the purposes of paraphrasing and improving English writing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Keith Norambuena, B.F.; Mitra, T.; North, C. A Survey on Event-Based News Narrative Extraction. *ACM Comput. Surv.* **2023**, *55*, 1–39. [CrossRef]
2. Ranade, P.; Dey, S.; Joshi, A.; Finin, T. Computational Understanding of Narratives: A Survey. *IEEE Access* **2022**, *10*, 101575–101594. [CrossRef]
3. Santana, B.; Campos, R.; Amorim, E.; Jorge, A.; Silvano, P.; Nunes, S. A survey on narrative extraction from textual data. *Artif. Intell. Rev.* **2023**, *56*, 8393–8435. [CrossRef]
4. Keith Norambuena, B.F.; Mitra, T. Narrative Maps: An Algorithmic Approach to Represent and Extract Information Narratives. *Proc. ACM Hum.-Comput. Interact.* **2021**, *4*, 1–33. [CrossRef]
5. German, F.; Keith, B.; North, C. Narrative Trails: A Method for Coherent Storyline Extraction via Maximum Capacity Path Optimization. In Proceedings of the Text2Story 2025 Workshop@ECIR2025. CEUR-WS, Lucca, Italy, 10 April 2025; pp. 15–22.
6. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]

7. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
8. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
9. Hu, Q.; Moon, G.; Ng, H.T. From moments to milestones: Incremental timeline summarization leveraging large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 7232–7246.
10. Qorib, M.R.; Hu, Q.; Ng, H.T. Just What You Desire: Constrained Timeline Summarization with Self-Reflection for Enhanced Relevance. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 27 February–4 March 2025; Volume 39, pp. 25065–25073.
11. La Quatra, M.; Cagliero, L.; Baralis, E.; Messina, A.; Montagnuolo, M. Summarize dates first: A paradigm shift in timeline summarization. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 418–427.
12. Xu, G.; Isaza, P.T.; Li, M.; Oloko, A.; Yao, B.; Sanctos, C.; Adebiyi, A.; Hou, Y.; Peng, N.; Wang, D. Nece: Narrative event chain extraction toolkit. *arXiv* **2022**, arXiv:2208.08063.
13. 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, Washington, DC, USA, 25–28 July 2010; KDD '10; pp. 623–632. [CrossRef]
14. Gómez-Rodríguez, C.; Williams, P. A Confederacy of Models: A Comprehensive Evaluation of LLMs on Creative Writing. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 14504–14528.
15. Zhou, H.; Hobson, D.; Ruths, D.; Piper, A. Large Scale Narrative Messaging around Climate Change: A Cross-Cultural Comparison. In Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), Bangkok, Thailand, 16 August 2024; pp. 143–155.
16. Chan, C.M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; Liu, Z. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
17. Doostmohammadi, E.; Holmström, O.; Kuhlmann, M. How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs? In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 6321–6336.
18. Chen, G.; Chen, S.; Liu, Z.; Jiang, F.; Wang, B. Humans or LLMs as the Judge? A Study on Judgement Bias. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 8301–8327.
19. Shahaf, D.; Guestrin, C.; Horvitz, E. Trains of thought: Generating information maps. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; WWW '12; pp. 899–908. [CrossRef]
20. Shahaf, D.; Guestrin, C.; Horvitz, E. Metro maps of science. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; KDD '12; pp. 1122–1130. [CrossRef]
21. Cheng, J.; Liu, W.; Wang, Z.; Ren, Z.; Li, X. Joint event extraction model based on dynamic attention matching and graph attention networks. *Sci. Rep.* **2025**, *15*, 6900. [CrossRef]
22. Keith Norambuena, B.F.; Mitra, T.; North, C. Mixed Multi-Model Semantic Interaction for Graph-based Narrative Visualizations. In Proceedings of the 28th International Conference on Intelligent User Interfaces, Sydney, Australia, 27–31 March 2023; IUI '23; pp. 866–888. [CrossRef]
23. Castricato, L.; Frazier, S.; Balloch, J.; Riedl, M. Fabula Entropy Indexing: Objective Measures of Story Coherence. In Proceedings of the Third Workshop on Narrative Understanding, Online, 11 June 2021; pp. 84–94.
24. Bansal, N.; Akter, M.; Santu, S.K.K. SEM-f1: An automatic way for semantic evaluation of multi-narrative overlap summaries at scale. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 780–792.
25. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical neural story generation. *arXiv* **2018**, arXiv:1805.04833.
26. Goldfarb-Tarrant, S.; Chakrabarty, T.; Weischedel, R.; Peng, N. Content Planning for Neural Story Generation with Aristotelian Rescoring. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 4319–4338.
27. Yi, Q.; He, Y.; Wang, J.; Song, X.; Qian, S.; Yuan, X.; Zhang, M.; Sun, L.; Li, K.; Lu, K.; et al. Score: Story coherence and retrieval enhancement for ai narratives. *arXiv* **2025**, arXiv:2503.23512.
28. Zhu, D.; Wu, W.; Song, Y.; Zhu, F.; Cao, Z.; Li, S. CoUDA: Coherence Evaluation via Unified Data Augmentation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024; pp. 967–978.

29. Amorim, E.; Campos, R.; Jorge, A.; Mota, P.; Almeida, R. text2story: A python toolkit to extract and visualize story components of narrative text. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 15761–15772.
30. Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; Liu, Y. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv* **2024**, arXiv:2412.05579.
31. Sun, Y.; Zhu, D.; Chen, Y.; Xiao, E.; Chen, X.; Shen, X. Fine-Grained and Multi-Dimensional Metrics for Document-Level Machine Translation. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 1–17.
32. Jeong, Y.; Kim, M.; Hwang, S.W.; Kim, B.H. Agent-as-Judge for Factual Summarization of Long Narratives. *arXiv* **2025**, arXiv:2501.09993.
33. Kranti, C.; Hakimov, S.; Schlangen, D. clem: Todd: A Framework for the Systematic Benchmarking of LLM-Based Task-Oriented Dialogue System Realisations. *arXiv* **2025**, arXiv:2505.05445.
34. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 2511–2522.
35. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
36. Fu, J.; Ng, S.K.; Jiang, Z.; Liu, P. GPTScore: Evaluate as You Desire. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024; pp. 6556–6576.
37. Shen, J.; Mire, J.; Park, H.; Breazeal, C.; Sap, M. HEART-felt Narratives: Tracing Empathy and Narrative Style in Personal Stories with LLMs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 1026–1046.
38. Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. Training verifiers to solve math word problems. *arXiv* **2021**, arXiv:2110.14168.
39. Amini, A.; Gabriel, S.; Lin, S.; Koncel-Kedziorski, R.; Choi, Y.; Hajishirzi, H. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 2357–2367.
40. Zeng, Z.; Chen, P.; Liu, S.; Jiang, H.; Jia, J. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv* **2023**, arXiv:2312.17080.
41. Xia, S.; Li, X.; Liu, Y.; Wu, T.; Liu, P. Evaluating mathematical reasoning beyond accuracy. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 27 February–4 March 2025; Volume 39; pp. 27723–27730.
42. Ye, J.; Wang, Y.; Huang, Y.; Chen, D.; Zhang, Q.; Moniz, N.; Gao, T.; Geyer, W.; Huang, C.; Chen, P.Y.; et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv* **2024**, arXiv:2410.02736.
43. Verga, P.; Hofstatter, S.; Althammer, S.; Su, Y.; Piktus, A.; Arkhangorodsky, A.; Xu, M.; White, N.; Lewis, P. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv* **2024**, arXiv:2404.18796.
44. Li, Z.; Wang, C.; Ma, P.; Wu, D.; Wang, S.; Gao, C.; Liu, Y. Split and Merge: Aligning Position Biases in LLM-based Evaluators. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 11084–11108.
45. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
46. Tonmoy, S.; Zaman, S.; Jain, V.; Rani, A.; Rawte, V.; Chadha, A.; Das, A. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv* **2024**, arXiv:2401.01313.
47. Bechard, P.; Ayala, O. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), Mexico City, Mexico, 16–21 June 2024; pp. 228–238.
48. Kollias, G.; Das, P.; Chaudhury, S. Generation constraint scaling can mitigate hallucination. *arXiv* **2024**, arXiv:2407.16908.
49. Long, D.X.; Nguyen, N.H.; Sim, T.; Dao, H.; Joty, S.; Kawaguchi, K.; Chen, N.; Kan, M.Y. LLMs Are Biased Towards Output Formats! Systematically Evaluating and Mitigating Output Format Bias of LLMs. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2025; pp. 299–330.
50. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv* **2022**, arXiv:2203.11171.

51. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-Verification Reduces Hallucination in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 3563–3578.
52. Li, J.; Zhang, Q.; Yu, Y.; Fu, Q.; Ye, D. More agents is all you need. *arXiv* **2024**, arXiv:2402.05120.
53. Taubenfeld, A.; Sheffer, T.; Ofek, E.; Feder, A.; Goldstein, A.; Gekhman, Z.; Yona, G. Confidence Improves Self-Consistency in LLMs. *arXiv* **2025**, arXiv:2502.06233.
54. Naismith, B.; Mulcaire, P.; Burstein, J. Automated evaluation of written discourse coherence using GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Toronto, ON, Canada, 13–14 July 2023; pp. 394–403.
55. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
56. Hackl, V.; Müller, A.E.; Granitzer, M.; Sailer, M. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Front. Educ.* **2023**, *8*, 1272229. [CrossRef]
57. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258.
58. Chhun, C.; Suchanek, F.M.; Clavel, C. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 1122–1142. [CrossRef]
59. Lynch, C.J.; Jensen, E.; Munro, M.H.; Zamponi, V.; Martinez, J.; O'Brien, K.; Feldhaus, B.; Smith, K.; Reinhold, A.M.; Gore, R. GPT-4 Generated Narratives of Life Events using a Structured Narrative Prompt: A Validation Study. *arXiv* **2024**, arXiv:2402.05435.
60. Isenberg, P.; Heimerl, F.; Koch, S.; Isenberg, T.; Xu, P.; Stolper, C.; Sedlmair, M.; Chen, J.; Möller, T.; Stasko, J. vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2199–2206. [CrossRef]
61. Chen, Z.; Wang, S.; Tan, Z.; Fu, X.; Lei, Z.; Wang, P.; Liu, H.; Shen, C.; Li, J. A survey of scaling in large language model reasoning. *arXiv* **2025**, arXiv:2504.02181.
62. Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; Tu, Z. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 17889–17904.