Fiction Flows: A Replication and Reinterpretation of Narrative Sequentiality

Andrew Piper
McGill University

Sil Hamilton
Cornell University

Haiqi Zhou
McGill University

Federico Pianzola
University of Groningen

Abstract

Narrative flow emerges from the interplay between memory and expectation, shaping how stories are both produced and understood. To operationalize this construct, Sap et al. (2022) propose *sequentiality*, a language-model–based measure of sentence-level predictability, and report that imagined stories flow better than recalled ones. We conduct a large-scale replication across multiple language models, examine how modeling choices shape the original findings, and test generalization beyond crowdworker data using passages from published fiction and narrative non-fiction. Although the original contrast replicates under their initial formulation, it diminishes substantially under alternative specifications, suggesting that it reflects properties of the measurement setup rather than a stable feature of narrative flow. By contrast, fiction does appear to exhibit a robust sequentiality advantage over reality-bound genres under a minimal context-only formulation. However, mixed-effects analyses indicate that this advantage is not reducible to standard coherence measures, underscoring the need for further theoretical and empirical grounding of narrative flow.

1 Introduction

Sap et al. (2022, 2020) sought to quantify "narrative flow"—the degree to which one event leads naturally to the next in a story—by introducing a computational measure called *sequentiality*. Using GPT-3 to estimate sentence-level likelihoods, they find higher sequentiality in imagined than autobiographical stories, arguing that imagined narratives follow shared schemas more closely, while recollections include episodic details that disrupt flow.

Narrative flow is an important construct because it marks the meeting point of cognitive processes such as memory, prediction, schema activation, and their realization in text. As a longstanding concept in narrative theory (Pianzola et al., 2021), studying

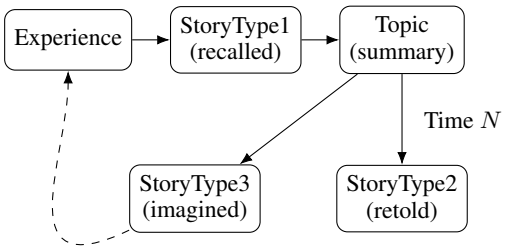

Figure 1: Schematic of the generative framework in Sap et al. (2022), modeling the cognitive relationship between lived experience, recalled narration, summary formation, and subsequent retold and imagined stories.

flow can give insight into how minds produce and make sense of narrative sequences. Understanding how language models can operationalize flow to predict sequences substantially contributes to computational approaches to narrative understanding.

Flow is a psychological construct that traditionally refers to human experiences (Csikszentmihalyi, 1990), including readers' engagement with a narrative (Thissen et al., 2018), emerging from the interplay of different cognitive processes (Pier, 2016; Pianzola et al., 2021). Because it arises from subtle interactions between syntax, semantics, and world knowledge, narrative flow has proven difficult to study at scale, despite its centrality to how minds produce and make sense of unfolding events. The concept of "narrative sequentiality" (Grabes, 2014) refers to the specific textual dimension of this phenomenon, describing either features inherent in a text (Graesser et al., 2004) or the ordered cognitive processes through which readers construe coherence and progression (Graesser et al., 1994).

Large language models (LLMs) offer a novel lens onto these dynamics. Trained to predict next tokens on vast corpora, they internalize many of the distributional cues that support human expectations about what typically follows in a narrative. Prior research suggests LLMs can serve both as method-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34158–34174

ological tools for quantifying aspects of narrative flow and as computational hypotheses about how authors and readers generate and anticipate narrative continuations (Naismith et al., 2023; Zhang and Long, 2025; Pianzola, 2024).

Despite its contributions, the original study by Sap et al. (2022) leaves several questions unresolved. First, sequentiality provides only a narrow operationalization of flow, equating it with a model's predictive ease under the assumption this quantity is correlated with readers' cognitive and textual processes. Second, because Sap et al. (2022)'s results depend on a single proprietary model (GPT-3 Davinci), it remains unclear how robust their findings are to model architecture, scale, or training data. Third, the implementation of the sequentiality formula itself introduces several idiosyncrasies, most notably around the inclusion of plot summaries (see Fig. 1), that make it difficult to isolate which components of the measure drive the observed differences.

We are not the first to raise concerns about the findings and methods of the original paper. Sunny et al. (2025) use two alternative datasets and three open-source models to highlight the strong likelihood of measurement bias in Sap et al. (2022)'s formula. Given these initial concerns, along with the value of measuring narrative flow computationally, we undertake an even more expansive replication effort here. Our core contributions are as follows:

Model robustness. We show that the imagined > recalled sequentiality effect reported by Sap et al. (2022) replicates under their original formulation across seven language models of differing size and architecture, but becomes unstable and model-dependent once the topic term is removed.

Formula sensitivity. We demonstrate that the original effect hinges almost entirely on the inclusion of the topic (summary) term: under a context-only formulation intended to isolate local continuity, the imagined > recalled contrast largely collapses, indicating sensitivity to story–summary mismatch rather than contextual flow.

Data dependence. We find that this failure to replicate is specific to the HIPPOCORPUS regime. When moving beyond crowd-sourced diary narratives to published genres, fictional narratives—here a proxy for imagined storytelling—exhibit a consistent sequentiality advantage over matched nonfiction (recalled) genres. This suggests sequentiality is better understood as a property of fictional narration than of imagination *per se*.

Measurement validity. Finally, we show that LLM-based sequentiality is not reliably predicted by standard indicators of discourse coherence or continuity. This suggests that sequentiality reflects a distinct contextual expectation that may not align strongly with traditional notions of narrative flow.

Taken together, these findings move beyond direct replication and point toward a need to reconsider how narrative flow is operationalized in computational settings. Rather than providing a direct measure of flow, LLM-based sequentiality appears to capture a more limited form of contextual expectation that varies with modeling assumptions and data context. These findings can provide the foundation for future investigations into this important narrative mechanism.[1]

2 Defining Flow and Sequentiality

Narrative flow and *narrative sequentiality* are foundational concepts in understanding how narratives function both cognitively and textually. For NLP, these concepts are critical in modelling narrative comprehension, generation, and evaluation — but these two concepts are not synonymous.

Flow is a broad concept transversal to various human experiences, (mainly) conceptualised in relation to the balance between a person's skills and the complexity of a perceptual stimulus (Csikszentmihalyi, 1990). It is specifically characterized by the perception of a sense of challenge. With respect to narrative, the comprehension of the content of the story is an aspect which can be connected to this sense of challenge, since the right match between the complexity of a story and cognitive skills is relevant for the reader to have an engaging experience (Thissen et al., 2018; Pianzola et al., 2021).

Neuroscientific and psychological studies demonstrate that narrative flow involves the dynamic integration of incoming narrative information into a causally coherent structure, which is essential for comprehension and memory (Song et al., 2021; Chang et al., 2022). Computational analogs in NLP attempt to simulate these processes via neural architectures that track dependencies over time, such as recurrent neural networks (RNNs), transformers, and multi-agent AI, which model sequential data and hidden states reflecting narrative dynamics (Rashkin et al., 2020; Wilner et al., 2021; Guan et al., 2023; Balestri and

[1]We release our code and data: https://github.com/dot-txtlab/fiction-flow.

Pescatore, 2025). However, NLP systems often conflate perceived narrative flow (a subjective psychological state) with narrative sequentiality (a property directly related to the text itself). While computational measures like sequentiality scores quantify temporal dependencies, they do not fully capture the whole subjective experience that define narrative flow. Addressing this gap requires uncovering how to best model sequentiality such that this construct can be later integrated into a broader computational model of narrative flow.

Narrative sequentiality concerns the organization of narrative events into a structured, ordered sequence, incorporating temporal and causal dependencies. These sequences can be both hierarchical and non-linear (e.g. nested narratives, flashbacks, and subplots) which complicates the temporal ordering and require sophisticated discourse processing. Textual definitions emphasize the role of linguistic cues—such as tense, discourse markers, and anaphora—that signal event order and relationships (Grabes, 2014). The concept of "scripts" highlights that narratives often follow structured sequences of actions, which facilitate comprehension by providing predictable patterns (Schank and Abelson, 1977; Herman, 2002).

3 Summary of Original Study

Sap et al. (2022) investigate how the narrative flow of stories differs when people recall autobiographical experiences versus imagined events (Figure 1). They define narrative flow as *narrative sequentiality*, the extent to which each sentence in a story follows from the evolving context of preceding sentences and from a shared schematic (script-like) understanding of how events typically unfold.

To operationalize narrative sequentiality computationally, Sap et al. (2022) frame it as a problem of conditional predictability under a language model. Their sequentiality measure estimates how much more predictable a sentence becomes when conditioned on prior narrative context, relative to a global topic baseline provided by a brief story summary.

Formally, for each sentence s_i in a story associated with a topic T and for a given history size h (the h preceding sentences $s_{i-h}, \ldots, s_{i-1}$), sentence-level predictability is estimated under two conditioning regimes using a language model.

Let $s_i = (w_{i,1}, \ldots, w_{i,|s_i|})$ denote the sequence of tokens in sentence s_i. The topic-only negative log-likelihood is defined as

$$\mathrm{NLL}_T(s_i) = -\frac{1}{|s_i|} \sum_{t=1}^{|s_i|} \log p_\theta\big(w_{i,t} \mid T\big), \quad (1)$$

and the contextual negative log-likelihood, conditioning on both the topic and the h preceding sentences, as

$$\mathrm{NLL}_C(s_i, h) = -\frac{1}{|s_i|} \sum_{t=1}^{|s_i|} \log p_\theta\big(w_{i,t} \mid T, s_{i-h}, \ldots, s_{i-1}\big). \quad (2)$$

Sequentiality for sentence s_i at history size h is then defined as the length-normalized difference

$$\mathrm{SEQ}(s_i, h) = \mathrm{NLL}_T(s_i) - \mathrm{NLL}_C(s_i, h), \quad (3)$$

which captures the gain in predictability attributable to local narrative context beyond the topic baseline. Story-level sequentiality for a story d is obtained by averaging over its n_d sentences:

$$\mathrm{SEQ}(d, h) = \frac{1}{n_d} \sum_{i=1}^{n_d} \mathrm{SEQ}(s_i, h). \quad (4)$$

Results are reported for history sizes $h = 1, \ldots, 9$ as well as for the full-story history.

This formulation is instantiated using the HIPPOCORPUS dataset,[2] which operationalizes both narrative context and topic information in a specific way. The dataset contains 6,854 short "diary-like" narratives written by crowd workers as *recalled*, *retold* (3–6 months later), or *imagined* stories about the same underlying topics. Crucially, authors of recalled experiences first write a brief summary (the "topic"), which is then reused to condition both subsequent retellings of the same experience and the generation of imagined stories (Figure 1). As a result, the topic term T functions as a shared global conditioning signal across story types, while local narrative context varies within each story.

Using GPT-3, Sap et al. (2022) report that imagined stories exhibit higher sequentiality than recalled autobiographical stories, with retold narratives occupying an intermediate position. These differences persist across context lengths, leading the authors to argue that imagined narratives more closely follow shared schematic structures, whereas autobiographical recollection introduces idiosyncratic details that disrupt narrative flow.

[2] https://www.microsoft.com/en-us/download/details.aspx?id=105291

4 Methods

We evaluate the robustness and interpretation of sequentiality by varying the language model, likelihood formulation, data source, and comparison measures for assessing narrative structure.

4.1 Model Effects

Prior work shows that many NLP findings are sensitive to model architecture, scale, and training regime (Bommasani et al., 2021; Kaplan et al., 2020; Longpre et al., 2024). Because narrative sequentiality is itself a model-derived quantity, we test whether the effects reported by Sap et al. (2022) are robust across contemporary language models with different design assumptions.

We evaluate seven causal language models spanning multiple generations of Transformer architectures, parameter scales, and training regimes. These include a GPT-3 baseline (`davinci-002`), used to approximate the original model in Sap et al. (2022), as well as a set of open-weight models that vary in size and pretraining data (Table 1). To support faithful replication, we focus primarily on base models, matching the original next-token prediction setting used by Sap et al. (2022) and enabling comparisons across model generations. This choice is also theoretically motivated: because sequentiality is defined in terms of token-level log-likelihoods, its interpretation depends on the underlying probability distribution learned during pretraining. Instruction-tuned models alter this distribution through alignment objectives that prioritize coherence and helpfulness, potentially distorting likelihood-based estimates of contextual expectation. Nevertheless, to assess the impact of such tuning, we additionally evaluate instruction-tuned variants of all open-weight models considered.

Together, our models range from early decoder-only Transformers trained on relatively small web corpora to more recent architectures with longer context windows and more diverse pretraining mixtures. All open-weight models are quantized to 4 bits and run with `llama.cpp` modified to echo NLL probabilities in each response. To manage computational cost, we evaluate each model on a random 50% subsample of the original corpus.

4.2 Formula Effects

A central design choice—and potential confound—in Sap et al. (2022)'s formulation is the use of a topic (summary) term as the baseline against which contextual predictability is measured. Because sequentiality is defined as the difference between a topic-only and a topic-plus-context negative log-likelihood, variation in the resulting score can reflect not only gains from local narrative context but also divergence between a sentence and the topic summary. Sentences that are equally well supported by their preceding context may receive different sequentiality scores solely due to differences in how closely they match the summary.

Model	Parameters	Tokens	Year
GPT-2	1.5B	40B	2019
GPT-3	175B	300B	2020
LLaMA 3.1	8B	15T	2024
Qwen 3	8B	36T	2025
Qwen 3	0.6B	36T	2025
SmolLM 3	3B	11.2T	2025
Mistral 3	8B	-	2025

Table 1: Models considered in our experiment.

This concern is empirically substantiated by Sunny et al. (2025), who show that sequentiality differences are often driven primarily by the topic-driven likelihood rather than by contextual predictability. By varying how topics are generated, they demonstrate that the direction and magnitude of sequentiality effects can change substantially, and in some cases reverse entirely. Their analysis further shows that the contextual term in isolation behaves in line with intuitive notions of narrative flow, whereas the inclusion of an explicit topic baseline can obscure this signal.

Building on this diagnosis, we implement two complementary variants of the sequentiality formulation designed to separate contextual effects from topic-driven artifacts:

- **(A) Exact replication ("With Topic").** We first reproduce the original formulation in Eq. (4), reconstructing the procedure described by Sap et al. (2022) as closely as possible. Because neither code nor intermediate outputs were released with the original paper, this replication is necessarily approximate. Our implementation and outputs are publicly available.
- **(B) Context-only sequentiality ("No Topic").** We then construct a modified formulation that removes the topic from *both* the baseline and contextual terms in order

to isolate the contribution of local narrative context alone. Specifically, we replace the topic prompt with an empty string, yielding an (approximately) unconditional negative log-likelihood $\mathrm{NLL}_{\emptyset}(s_i)$ for sentence s_i.

This unconditional baseline captures how intrinsically familiar or difficult a sentence is for the language model, independent of any story-specific information. Subtracting the contextual likelihood from this baseline therefore measures whether a sentence becomes more predictable *because of* the preceding narrative, rather than because it is itself surprising or rare.

For the contextual probability, we condition only on the h preceding sentences $s_{i-h:i-1}$, yielding a length-normalized contextual likelihood $\mathrm{NLL}_{C\emptyset}(s_i \mid s_{i-h:i-1})$. Our modified sequentiality measure is defined as:

$$\mathrm{SEQ}'(s_i, h) = \mathrm{NLL}_{\emptyset}(s_i) - \mathrm{NLL}_{C\emptyset}(s_i \mid s_{i-h:i-1}). \quad (5)$$

Finally, to compare narrative conditions, we conduct pairwise comparisons between story types using Welch's two-sample t-tests and report standardized effect sizes using Hedges' g. Analyses are performed separately for each model, likelihood formulation, and context size. We also examine the unconditional term $\mathrm{NLL}_{\emptyset}(s_i)$ directly to verify what effect sentence-level predictability has on our formula. Unlike the linear regression framework used by Sap et al. (2022), our replication indicates passage length does not contribute meaningful explanatory power, motivating its exclusion and favoring standardized mean differences that do not rely on assumptions of linearity.

4.3 Data Effects

Another potential source of variation concerns the data generation regime used to instantiate autobiographical and imaginative narratives. While HIPPOCORPUS enables controlled contrasts between recalled, retold, and imagined stories, several aspects of its construction may shape observed sequentiality differences and limit generalization. The instruction to write "diary-like" stories, for example, imposes a narrow genre constraint that may not generalize to other forms of autobiographical narration. Similarly, imagined stories are generated in response to short summaries written by other participants, ensuring topical alignment but constraining imaginative variation. These design choices raise questions about whether the reported effects reflect properties of narrative flow more generally or of this specific crowd-sourced, summary-conditioned setting.

In their replication, Sunny et al. (2025) address some of these concerns by adopting a different data generation strategy, aligning passages from published autobiographies with biographies of the same historical figures to control for event content. This design offers tighter topical control and avoids crowdworker-generated diary prose, but introduces new constraints. In particular, it narrows the "imagined" condition to third-person biographical narration of real lives rather than genuinely imaginative storytelling. It also confounds memory condition with point of view, as autobiographical passages are first-person while biographical passages are third-person. As a result, the study probes a distinct contrast—authorial perspective on shared events—rather than a more broadly understood concept of imagination.

To examine the generalizability of sequentiality effects beyond crowd-sourced data, we introduce a third data generation process using passages from published fiction and non-fiction while controlling for point of view and narrativity. We sample two twenty-sentence passages from each book in the CONLIT dataset (Piper, 2022) and compare fiction to matched non-fiction genres, including autobiographical, biographical, and non-narrative texts. Non-fiction works are drawn from Amazon bestseller lists, while fiction is sourced from novels reviewed in the *New York Times*. Although this design forgoes explicit topical alignment, it provides a stricter operationalization of autobiographical writing and a broader conception of imaginative narration. For comparability with prior work, we also include an autobiography–biography contrast following Sunny et al. (2025).

4.4 Measurement Validity

We next evaluate the construct validity of NLL-based sequentiality as a measure of narrative flow. Sap et al. (2022) relate sequentiality to auxiliary covariates including realis event counts, LIWC categories, lexical concreteness, and human judgments of event salience. While informative, these measures are not designed to capture narrative coherence directly and often reflect adjacent phenomena such as narrative absorption or episodic detail.

To more directly assess narrative organization, we compare sequentiality to sentence-level lin-

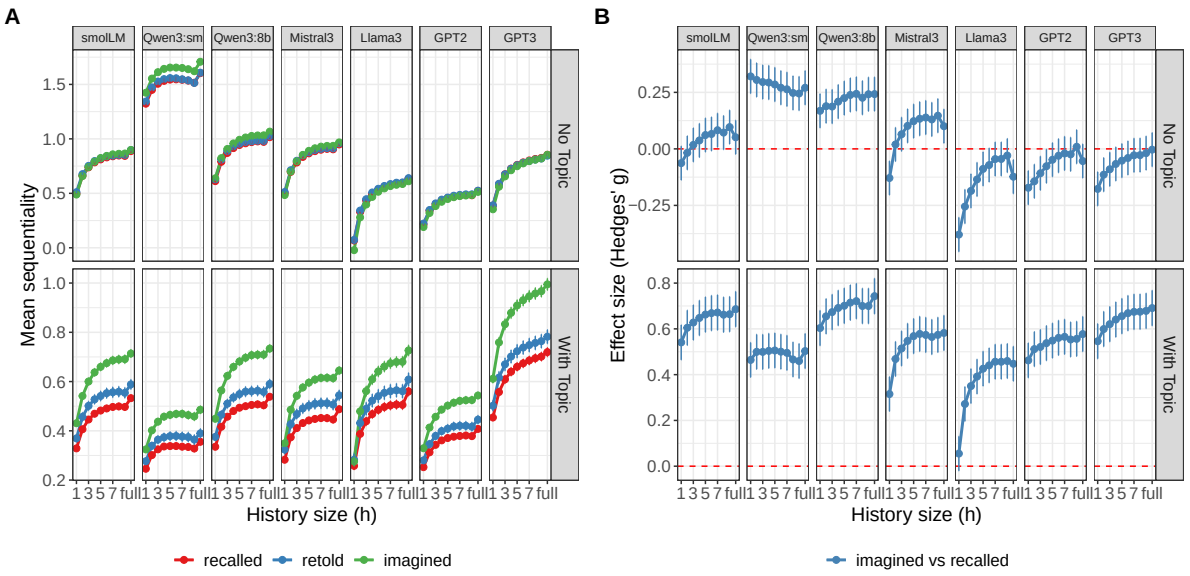

Figure 2: **Model- and formulation-level effects on narrative sequentiality (original dataset).** (**A**) Mean sequentiality as a function of context length (h) under topic-conditioned and context-only formulations across seven language models for recalled, retold, and imagined narratives. (**B**) Effect sizes (Hedges' g) for the imagined vs. recalled contrast across models, context lengths, and formulations. The dashed line at $g = 0$ indicates no difference. Error bars show 95% confidence intervals.

guistic and structural indicators grounded in prior work on discourse coherence and text comprehension—particularly models of local coherence and situation model construction that emphasize lexical overlap, referential continuity, and discourse relations (Graesser et al., 1994, 2004; Kintsch, 1988). We group these measures into three complementary dimensions capturing local coherence, situation model continuity, and processing difficulty. All measures are computed at the sentence level and aligned to adjacent sentence pairs and more fully described in Table C1.

Lexical and semantic overlap captures whether adjacent sentences share surface or latent semantic content, a core component of local coherence. We include (i) lexical overlap, measured via the Jaccard index over content-word lemmas, and (ii) semantic overlap, measured as cosine similarity between adjacent sentences' embedding vectors.

Discourse and situation continuity captures whether adjacent sentences are linked by logical and rhetorical structure and maintain a stable situation model. We operationalize this using indicators for discourse connectives, RST-derived relations, tense continuity, and agent continuity via subject coreference—classic dimensions of discourse coherence shown to support narrative integration.

Sentence-level difficulty controls for the possibility that sequentiality primarily reflects processing difficulty rather than narrative structure. These include sentence length, word rarity, clause count, and clause tree depth.

Using the context-only formulation (subsection 4.2), we then compute sentence-pair NLL-based sequentiality for immediate (H1) and medium-range (H3) windows and fit hierarchical regressions that predict sequentiality from between-sentence relational continuity and sentence-level difficulty with story-level random effects.[3]

5 Results

5.1 Direct Replication is Successful

As shown in Figure 2 (bottom row), across all models the original formulation by Sap et al. (2022) yields medium to large effect sizes between recalled and imagined stories as initially reported, with a minimum median g of 0.43 (Llama 3.1 8B), a maximum of 0.70 (Qwen 3 8B), and a mean across all models of 0.58. For retold versus recalled, we see the smaller effect sizes also reported in the original paper (min=0.16 (Llama 3.1), max=0.24 (smolLM), mean=0.20).

[3]For more, see Appendix B.

5.2 But Everything Hinges on the Topic Term

However, when we remove the topic term and condition only on context (Figure 2, top row) we see effect sizes collapsing effectively to 0 for both the retold-recalled conditions (mean g across all models = 0.04) and imagined-recalled (mean g = 0.07). Older models (GPT-2/3, Llama 3.1) actually show a reversed effect at small context windows that attenuates with larger contexts. Qwen 3 is the only model family to show a consistent positive effect across windows, though effect sizes remain small (mean $g = 0.28$ for 0.6B; $g = 0.22$ for 8B). See Appendix A for full results.

One possible conclusion is that these results underscore the importance of the topic term: only when topical alignment is controlled do imagined stories appear to flow more consistently. Yet this leads to a counterintuitive implication. Under Sap et al. (2022)'s formulation, flow largely reflects divergence from a story's summary rather than narrative continuity. Imagined stories score higher not because they are more sequential, but because they are less similar to the provided summaries, whereas autobiographical stories are unsurprisingly closer to summaries written by their own authors. Consistent with this interpretation, Sunny et al. (2025) show that the topic-based formulation has no predictive power on synthetic data designed to exhibit high and low flow.

5.3 Fictional Stories Do Exhibit More Flow

The failure of the modified formulation to replicate the original results does not rule out the fact that imagined stories may still exhibit more sequentiality than autobiographical stories even if we use the context-only approach. While the crowdworker-produced stories of HIPPOCORPUS do not exhibit these differences using the context-only measure beyond potential very small effects shown by a single model, it is still possible that other definitions of imagined and recalled writing might show more significant and consistent effects.

Using our genre-based operationalization, we find clear and consistent differences in sequentiality across all fiction conditions that contrast sharply with the null results observed for crowdworker-generated narratives. As shown in Figure 3 and Table 2, fiction exhibits substantially higher sequentiality than recalled genres with medium to large mean effect sizes. Similar to Sunny et al. (2025), we do not find meaningful differences between the

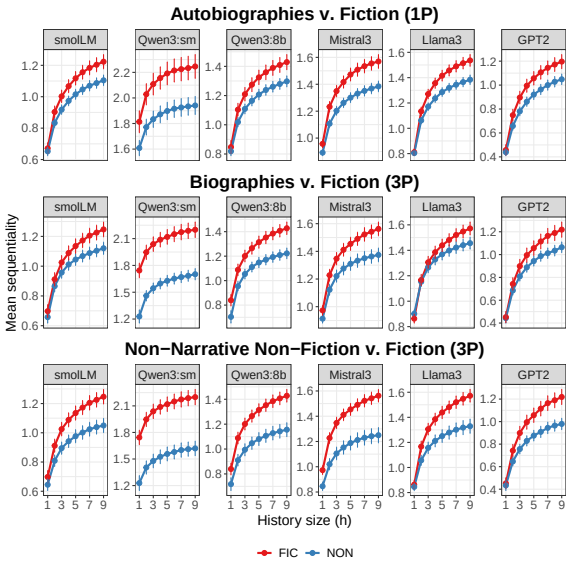

Figure 3: Mean sequentiality scores are shown across nine context window sizes (h = 1-9) for six language models, comparing fiction (FIC) against three non-fiction genres (NON) aligning for point-of-view. Error bars represent 95% confidence intervals.

two forms of narrative non-fiction (autobiography and biography) suggesting that the observed effect depends on modelling "imagination" as the construction of non-real events rather than real ones.

Recalled	Imagined	g	N
Autobiography	FIC (1P)	.40	223/221
Biography	FIC (3P)	.46	181/196
Non-Narrative	FIC (3P)	.69	115/196
Autobiography	BIO	-.06	223/181

Table 2: Mean effect sizes (Hedges' g) across all models and contexts using Sap et al. (2022)'s recalled/imagined framing. N refers to the number of books per category.

To verify that fiction's sequentiality advantage reflects narrative structure rather than baseline linguistic rarity, we examined the unconditional term $\mathrm{NLL}_{\emptyset}(s_i)$ directly. Fiction exhibits substantially higher baseline NLL across all conditions, indicating that fictional sentences are more surprising unconditionally. This pattern validates the context-only formulation: fiction operates in a rarer linguistic space yet exhibits larger gains from narrative context, demonstrating that sequentiality successfully isolates narrative dependencies from lexical frequency effects.

5.4 Standard Coherence Indicators Do Not Predict NLL-Based Sequentiality

We do not find associations between standard coherence measures and NLL-based sequentiality. To conclude this, we model NLL-based sequentiality using mixed-effects regressions that jointly estimate the contribution of model identity, between-sentence continuity measures, discourse relations, and sentence-level difficulty. Across both immediate (H1) and medium-range (H3) context windows, we find that contextual gain in predictability is not reliably associated with any of our discourse coherence measures when controlling for sentence-level difficulty (all $p > .20$). The one exception was change in sentence length at the 1-sentence context window, where sentences that were longer relative to their immediate predecessor exhibited reduced flow ($p = 0.0016$). These results suggest that NLL-based sequentiality reflects a form of contextual expectation that is not well captured by standard surface coherence indicators and that warrants further investigation.

5.5 Instruction-Tuned Models Reveal Instability

To test whether our results are specific to base LLMs, we evaluate the post-trained variants of our open-weight models on the core HIPPOCORPUS replication task (Appendix A).[4]

We find that instruction-tuned models reproduce the original pattern under the topic formulation (Eq. 4), with imagined narratives more sequential than recalled ones ($g = 0.465$), though the effect is attenuated relative to base models. Under the no-topic formulation (Eq. 5), a small positive effect remains ($g = 0.221$), suggesting—on the surface—that imagined narratives do exhibit greater sequentiality under instruct-tuning, partially supporting, although with much weaker effects, the initial finding of (Sap et al., 2022). However, decomposition of this measure reveals a reversal: both the unconditional baseline ($\text{NLL}_{\emptyset}$, $g = -0.196$) and the contextual likelihood ($\text{NLL}_{C\emptyset}$, $g = -0.213$) independently favor *recalled* narratives.

This apparent contradiction arises because sequentiality is defined as the difference between baseline and contextual predictability, so the result depends on their relative change rather than their independent direction. Because both components independently favor recalled narratives, the positive combined effect for imagined stories should not be interpreted as indicating greater contextual predictability in imagined stories. Instead, the decomposition reveals that the full formulation is reflecting the interaction of distortions introduced by instruction tuning. This contrasts with the CONLIT results, where decomposition aligns with the combined effect, indicating that the formulation behaves coherently under base models but becomes unstable under instruction tuning.

6 Conclusion

In this study we set out to evaluate Sap et al. (2022)'s central claim that imagined narratives exhibit higher "flow" than recalled narratives, a pattern they interpret as evidence that imagination draws more heavily on shared, schema-like event structures whereas memory introduces more idiosyncratic details that disrupts predictability.

Our first core finding is that the `imagined > recalled` effect replicates only under Sap et al. (2022)'s original With-Topic formulation (Eq. 4). When the topic (summary) term is removed—so that models condition solely on prior context—the contrast largely collapses and becomes unstable across models and datasets, posing a substantive challenge to the original interpretation. Because the Topic formulation measures whether prior context improves predictability relative to a summary written by participants, it effectively rewards divergence from that summary: continuations that depart from it can appear to "flow better" because context contributes more new information. As a result, the metric risks capturing story–summary mismatch rather than narrative sequentiality. When the summary term is removed, this confound disappears, and so does the effect, indicating that the original measure does not validly index narrative flow but instead conflates contextual predictability with summary misalignment.

A complementary pattern emerges under instruction tuning. While post-trained models still reproduce the topic-based effect, the no-topic formulation yields only weak and unstable differences. Crucially, decomposition reveals that both the baseline and contextual components independently favor recalled narratives, even as their difference produces a small positive effect for imagined stories. This divergence indicates that the residual signal arises from the interaction of distorted likelihood

[4] All models are evaluated with their respective prompt template to ensure fairness.

estimates rather than a coherent gain in contextual predictability.

Second, we find that the above failure to replicate the contrast between `imagined` > `recalled` stories is only true under their specific data-generation regime: short, diary-like narratives produced by crowd workers and derived from summary prompts. When we move beyond this setup to longer, formally published genres—comparing published fiction to matched non-fiction genres—the imagined condition (here fiction) shows consistently higher sequentiality across models and conditions using the more appropriate base-model approach. This indicates narrative sequentiality appears to be a meaningful property of fiction rather than a valid marker of reality-based imagination (biography) or memory-based cognition as revealed in prompted continuation tasks.

Third, our covariate analysis suggests that NLL-based contextual gain is not reliably aligned with standard measures of discourse coherence, one of the primary ways "flow" has been understood. When controlling for sentence-level difficulty, mixed-effects models show no consistent relationship between contextual gain and lexical overlap, semantic similarity, entity continuity, tense alignment, or discourse relations.

6.1 Future Work

Our results raise a core interpretive question for future work: what is NLL measuring when it comes to narrative sequentiality? One possibility is that our covariate coherence measures are too limited and better measurements would align with sequentiality. A second possibility is that NLL-based contextual gain captures event-level predictability—how much a sentence aligns with distributional expectations about what typically happens next—rather than surface cohesion. Alternatively, it may reflect model-internal priors about narrative progression or style that differ by training data and fine-tuning regime. A fourth possibility is that it responds to local surprisal management: how well a sentence resolves expectations generated by the prior discourse, in ways not well captured by classical coherence metrics. These interpretations remain open, but they shift how the original result should be understood. Imagined stories may exhibit greater contextual gain than recalled ones, yet this advantage does not align with narrative coherence as it is typically operationalized. Instead, NLL-based sequentiality appears to capture a distinct dimension of narrative expectation, offering a promising direction for refining computational theories of narrative flow.

Limitations

We acknowledge the following limitations in our work:

Absence of a prototypical narrative baseline. Our context-only formulation replaces the topic-conditioned baseline with an unconditional language-model likelihood in order to isolate the contribution of local narrative context. While this choice offers a transparent and easily interpretable comparator, it does not model expectations about what constitutes a *prototypical* narrative sequence. In other words, it captures whether a sentence becomes more predictable given preceding text, but not whether that sentence conforms to culturally or generically typical event progressions.

An ideal baseline for narrative flow would reflect schematic knowledge about what usually happens next in a story—independent of any specific narrative realization. Constructing such a baseline is challenging, however, because narrative prototypes vary systematically across cultures, genres, historical periods, and intended audiences. Even within a single language, expectations about event order differ markedly between, for example, children's fiction, realist novels, genre fiction, and personal narratives. Developing baselines that capture these higher-level narrative priors, while remaining comparable across domains, remains an open methodological problem. Our results therefore speak to local contextual predictability rather than to alignment with an abstract narrative schema.

Limited coverage of narrative genres. Our interpretation of fiction's higher sequentiality relative to non-fiction emphasizes the role of non-real, hypothetical, or counterfactual events in shaping narrative expectation. However, our fictional and non-fictional genres are limited. As a result, the observed fiction–non-fiction contrast may partially reflect genre-specific conventions rather than a general distinction between imagined and factual narratives. Other forms of narrative non-fiction—such as long-form journalism, historical writing, or creative non-fiction—may exhibit different sequentiality profiles, just as other genres of fiction may alter this relationship as well.

Distance from the original psychological construct. Sap et al. (2022) sought to capture a cognitive distinction between recalling and imagining under conditions of spontaneous story generation. Our findings, together with those of Sunny et al. (2025), suggest that under the constraints of the original corpus, these cognitive conditions yield little systematic difference in narrative sequentiality. It remains possible that alternative data collection strategies designed to elicit more spontaneous or less summary-mediated narratives could yet produce effects closer to those originally hypothesized. We therefore emphasize that our alternative data generation procedures are not intended to approximate the original psychological task, but rather to clarify how the categories of "autobiographical" and "imagined" operate under different textual and modeling conditions.

Limited cultural diversity. All analyses in this study (and the original) were conducted on English-language corpora, which necessarily reflect a narrow set of cultural contexts and narrative traditions. Narrative flow, as well as expectations about event ordering, pacing, and emphasis, may be shaped by culturally specific storytelling conventions. As a result, the patterns we observe may reflect specific narrative cultural norms rather than more general properties of storytelling.

Moreover, language models trained predominantly on English-language data may internalize culture-specific assumptions about what constitutes a coherent or well-formed narrative. These assumptions can influence measures of contextual predictability independently of any universal cognitive or narrative principles. Extending this work to culturally diverse corpora and narrative traditions along with more culturally-sensitive language models will be essential for assessing the broader applicability of NLL-based measures of narrative flow and for distinguishing culturally situated narrative expectations from more generalizable features of narrative organization.

AI Usage Disclosure

We used AI tools to assist with editing and generate draft code for data analysis. All code and text were reviewed by the authors, and all analyses, interpretations, and final wording are the authors' own.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Social Sciences and Humanities Research Council of Canada (SSHRC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) ainsi que le Conseil de recherches en sciences humaines du Canada (CRSH) de leur soutien.

References

R. Balestri and G. Pescatore. 2025. Multi-agent system for ai-assisted extraction of narrative arcs in tv series. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, volume 1, pages 663–670.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv e-prints*, pages arXiv–2108.

C. H. C. Chang, Samuel A. Nastase, and Uri Hasson. 2022. Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences of the United States of America*, 119(51).

Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. Harper and Row, New York.

Herbert Grabes. 2014. Sequentiality. In Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert, editors, *The Living Handbook of Narratology*. Hamburg University, Hamburg.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

Jian Guan, Zhenyu Yang, Rongsheng Zhang, Zhipeng Hu, and Minlie Huang. 2023. Generating coherent narratives by learning dynamic and discrete entity states with a contrastive framework. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

David Herman. 2002. *Story Logic: Problems and Possibilities of Narrative*. University of Nebraska Press, Lincoln, NE.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and 1 others. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. 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)*, pages 394–403.

Federico Pianzola. 2024. Dynamical systems, literary theory, and the computational modelling of narrative. *Interdisciplinary Science Reviews*, 49(2):222–236.

Federico Pianzola, Giuseppe Riva, Karin Kukkonen, and Fabrizio Mantovani. 2021. Presence, flow, and narrative absorption: an interdisciplinary theoretical exploration with a new spatiotemporal integrated model based on predictive processing. *Open Research Europe*, 1.

John Pier. 2016. The configuration of narrative sequences. In Raphaël Baroni and Françoise Revaz, editors, *Narrative Sequence in Contemporary Narratologies*, pages 20–36. Ohio State University Press.

Andrew Piper. 2022. The conlit dataset of contemporary literature. *Journal of Open Humanities Data*, 8.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A Smith, and James Pennebaker. 2020. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1970–1978.

Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A Smith, James W Pennebaker, and Eric Horvitz. 2022. Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45):e2211715119.

Roger Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

Hyunmi Song, Byung Yoon Park, Hyun Park, and Woo Min Shim. 2021. Cognitive and neural state dynamics of narrative comprehension. *The Journal of Neuroscience*, 41(43):8972–8990.

Amal Sunny, Advay Gupta, Yashashree Chandak, and Vishnu Sreekumar. 2025. From stories to statistics: Methodological biases in llm-based narrative flow quantification. In *The SIGNLL Conference on Computational Natural Language Learning*.

B. A. K. Thissen, Winfried Menninghaus, and Wolff Schlotz. 2018. Measuring optimal reading experiences: The reading flow short scale. *Frontiers in Psychology*, 9.

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative embedding: Re-Contextualization through attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinming Zhang and Yunfei Long. 2025. MLD-EA: Check and complete narrative coherence by introducing emotions and actions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1892–1907, Abu Dhabi, UAE. Association for Computational Linguistics.

A Full Effect Size Reporting

In this section we report all effect size calculations as Hedges' g across datasets, models, context windows and comparison conditions. For each dataset we also decompose our context-only formula (Eq. 5) into its component parts.

HIPPOCORPUS Results

Condition	Comparison	GPT2	GPT3	Llama3	Mistral3	Qwen3:8b	Qwen3:small	smolLM	Mean
With Topic	Imag vs Rec	0.551	0.662	0.433	0.566	0.700	0.500	0.662	0.582
	Ret vs Rec	0.180	0.185	0.160	0.227	0.231	0.175	0.237	0.199
No Topic	Imag vs Rec	-0.0510	-0.0462	-0.106	0.112	0.225	0.278	0.0563	0.0669
	Ret vs Rec	0.0443	-0.0158	0.0382	0.0602	0.0582	0.0319	0.0522	0.0384

Table A1: HIPPOCORPUS median Hedges' g across context windows, models, and conditions.

Comparison	GPT2	Llama3	Mistral3	Qwen3:8b	Qwen3:small	smolLM	Mean
Imag vs Rec	0.105	-0.063	-0.005	0.016	0.024	0.026	0.017
retold vs recalled	0.078	0.035	0.043	0.029	0.045	0.042	0.045

Table A2: HIPPOCORPUS median Hedges' g for Eq. (5) $\mathrm{NLL}_{C\emptyset}(s_i \mid s_{i-h:i-1})$ only formulation and all context windows.

Comparison	GPT2	Llama3	Mistral3	Qwen3:8b	Qwen3:small	smolLM	Mean
Imag vs Rec	0.084	-0.130	-0.152	-0.179	-0.268	-0.060	-0.117
retold vs recalled	0.054	0.001	-0.006	-0.021	-0.001	-0.003	0.004

Table A3: HIPPOCORPUS Hedges' g for Eq. (5) $\mathrm{NLL}_{\emptyset}(s_i)$ only formulation.

CONLIT Results

Imagined	Recalled	GPT2	Llama3	Mistral3	Qwen3:8b	Qwen3:small	smolLM	Mean
FIC1P	MEM	0.377	0.450	0.539	0.367	0.493	0.362	0.431
FIC3P	BIO	0.307	0.224	0.451	0.538	1.00	0.303	0.471
FIC3P	MIX	0.483	0.607	0.858	0.743	1.10	0.532	0.721
BIO	MEM	0.0735	0.301	0.0322	-0.245	-0.583	0.0846	-0.056

Table A4: CONLIT median Hedges' g for Eq. (5) full formulation and all context windows.

Imagined	Recalled	GPT2	Llama3	Mistral3	Qwen3:8b	Qwen3:small	smolLM	Mean
FIC1P	MEM	0.408	0.743	0.627	0.511	0.270	0.541	0.517
FIC3P	BIO	0.125	0.476	0.368	0.327	0.0177	0.281	0.266
FIC3P	MIX	0.214	0.503	0.430	0.357	0.117	0.306	0.321
BIO	MEM	0.425	0.196	0.274	0.202	0.299	0.251	0.275

Table A5: CONLIT median Hedges' g for Eq. (5) $\mathrm{NLL}_{C\emptyset}(s_i \mid s_{i-h:i-1})$ only formulation and all context windows.

Imagined	Recalled	GPT2	Llama3	Mistral3	Qwen3:8b	Qwen3:small	smolLM	Mean
FIC1P	MEM	0.744	1.04	1.00	0.735	0.638	0.746	0.817
FIC3P	BIO	0.407	0.636	0.736	0.707	0.862	0.483	0.639
FIC3P	MIX	0.675	0.982	1.12	0.931	1.06	0.699	0.911
BIO	MEM	0.487	0.465	0.272	0.00579	-0.306	0.317	0.207

Table A6: CONLIT Hedges' g for Eq. (5) $\mathrm{NLL}_{\emptyset}(s_i)$ only formulation.

Instruction-Tuned Results. We report effect sizes for instruction-tuned model variants evaluated on HIPPOCORPUS at $h \in \{0, 1, 3, 9\}$.

Condition	Comparison	Llama3-I	Mistral3-I	Qwen3:8b-I	Qwen3:sm-I	smolLM-I	Inst Mean
With Topic	Imag. vs Rec.	0.337	0.607	0.484	0.443	0.455	0.465
With Topic	Ret. vs Rec.	0.097	0.190	0.143	0.265	0.172	0.174
No Topic	Imag. vs Rec.	0.152	0.279	0.241	0.243	0.200	0.221
No Topic	Ret. vs Rec.	-0.021	0.014	—	—	—	-0.025

Table A7: HIPPOCORPUS median Hedges' g for instruction-tuned models, Eq. (4) and Eq. (5), $h \in \{1, 3, 9\}$.

Comparison	Llama3-I	Mistral3-I	Qwen3:8b-I	Qwen3:sm-I	smolLM-I	Inst Mean
Imag. vs Rec.	-0.137	-0.277	-0.233	-0.229	-0.190	-0.213
Ret. vs Rec.	0.029	0.013	0.015	0.064	0.008	0.026

Table A8: HIPPOCORPUS mean Hedges' g for $\mathrm{NLL}_{C\emptyset}$ (context only, no topic), instruction-tuned models, $h \in \{1, 3, 9\}$.

Comparison	Llama3-I	Mistral3-I	Qwen3:8b-I	Qwen3:sm-I	smolLM-I	Inst Mean
Imag. vs Rec.	-0.147	-0.274	-0.240	-0.200	-0.121	-0.196
Ret. vs Rec.	0.006	-0.007	-0.017	0.039	0.023	0.009

Table A9: HIPPOCORPUS Hedges' g for $\mathrm{NLL}_{\emptyset}$ (unconditional, $h = 0$, no topic), instruction-tuned models.

B Covariate Regression Analysis

To evaluate the construct validity of NLL-based sequentiality as a measure of narrative flow, we regressed sequentiality scores on sentence-level linguistic and structural indicators that operationalize narrative organization. All measures were computed at the sentence level and aligned to adjacent sentence pairs.

B.1 Covariate Groups

We organized covariates into three theoretically motivated groups:

Lexical and semantic overlap captures the extent to which adjacent sentences share surface or latent semantic content, a core component of local coherence:

- *Lexical overlap*: Jaccard index over content-word lemmas between adjacent sentences (H1 and H3 windows)
- *Semantic overlap*: Cosine similarity between adjacent sentences' embedding vectors (H1 and H3 windows)

Discourse and situation continuity captures whether adjacent sentences are linked by logical and rhetorical structure and maintain a stable situation model:

- *Discourse connectives*: Binary indicators for temporal, contingency, comparison, expansion, and onset markers
- *RST relations*: Categorical indicators for Rhetorical Structure Theory relations (H1 and H3 windows)
- *Entity continuity*: Subject coreference between adjacent sentences (H1 and H3 windows)
- *Tense continuity*: Tense agreement between adjacent sentences (H1 and H3 windows)

Sentence-level difficulty controls for the possibility that sequentiality primarily reflects processing difficulty rather than narrative structure:

- *Length*: Sentence length in tokens
- *Length delta*: Absolute difference in length between adjacent sentences (H1 and H3 windows)
- *Word rarity*: Rate of rare words per sentence
- *Clause count*: Number of clauses per sentence
- *Tree depth*: Maximum depth of syntactic dependency tree

B.2 Model Specifications

We fit hierarchical linear mixed-effects models predicting NLL-based sequentiality from the covariates described above, with random intercepts for story (indexed by `Row`). All continuous predictors were z-scored prior to modeling. We fit separate models for two context window sizes (H1 and H3) and two outcome specifications:

Gain models predict the reduction in NLL from adding context:

$$\text{Gain}_{ij} = \text{NLL}_{\text{solo},ij} - \text{NLL}_{\text{context},ij} \tag{6}$$

where i indexes sentence pairs and j indexes stories. The gain score represents how much predictability improves when adjacent sentences are available as context.

The regression equation for gain models is:

$$\begin{aligned}\text{Gain}_{ij} = \beta_0 + \beta_{\text{model}} + \beta_{\text{lex}}\text{Lexical}_{ij} + \beta_{\text{sem}}\text{Semantic}_{ij} \\ + \beta_{\text{ent}}\text{Entity}_{ij} + \beta_{\text{tense}}\text{Tense}_{ij} + \beta_{\text{len}\Delta}\text{LengthDelta}_{ij} \\ + \sum_k \beta_k \text{Connective}_{k,ij} + \beta_{\text{RST}}\text{RST}_{ij} \\ + \beta_{\text{len}}\text{Length}_{ij} + \beta_{\text{rare}}\text{RareRate}_{ij} \\ + \beta_{\text{clause}}\text{ClauseCount}_{ij} + \beta_{\text{depth}}\text{TreeDepth}_{ij} + u_j + \epsilon_{ij}\end{aligned} \tag{7}$$

where $u_j \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for story j.

Raw context models predict the absolute NLL with context, controlling for the solo NLL:

$$\begin{aligned}\text{NLL}_{\text{context},ij} = \beta_0 + \beta_{\text{solo}}\text{NLL}_{\text{solo},ij} + \beta_{\text{model}} \\ + \beta_{\text{lex}}\text{Lexical}_{ij} + \beta_{\text{sem}}\text{Semantic}_{ij} + \beta_{\text{ent}}\text{Entity}_{ij} \\ + \beta_{\text{tense}}\text{Tense}_{ij} + \beta_{\text{len}\Delta}\text{LengthDelta}_{ij} \\ + \sum_k \beta_k \text{Connective}_{k,ij} + \beta_{\text{RST}}\text{RST}_{ij} \\ + \beta_{\text{len}}\text{Length}_{ij} + \beta_{\text{rare}}\text{RareRate}_{ij} \\ + \beta_{\text{clause}}\text{ClauseCount}_{ij} + \beta_{\text{depth}}\text{TreeDepth}_{ij} + u_j + \epsilon_{ij}\end{aligned} \tag{8}$$

Models were fit using maximum likelihood estimation via the `lmer` function in the `lme4` R package. Statistical significance was assessed using Satterthwaite approximations for degrees of freedom via the `lmerTest` package, with p-values adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate procedure.

C Full Description of Covariate Coherence Measures

We describe our implementation of narrative coherence measures in Table C1.

Table C1: Cohesion and Complexity Measures

Category	Measure	Description	Implementation Details
Lexical and semantic overlap	Content-word overlap (lexical repetition)	Measures the overlap of content words (nouns, verbs, adjectives, adverbs) between adjacent sentences using lemmatized forms	For each pair of adjacent sentences s_i and s_{i+1}: (1) Tokenize sentences; (2) POS-tag and keep only content words; (3) Lemmatize words; (4) Compute overlap using Jaccard Similarity: $Overlap(i) = \frac{\lvert L(s_i) \cap L(s_{i+1}) \rvert}{\lvert L(s_i) \cup L(s_{i+1}) \rvert}$ where $L(s)$ = set of lemmas in sentence s.
	Sentence embedding similarity (contextual semantic cohesion)	Measures latent semantic coherence between adjacent sentences using contextual embeddings.	Encode each sentence as a sentence-level vector using BERT-base CLS embedding (mean-pooled contextual embeddings). Compute cosine similarity: $SemSim(i) = \cos(\vec{s_i}, \vec{s_{i+1}})$
Discourse and situation continuity	Lexical discourse connectives	Measures the presence and category of explicit discourse connectives that structure narrative progression.	For each sentence s_i, extract binary indicators using curated lexicons from discourse parsing literature (PDTB-style inventories): $TemporalConnective(s_i) \in \{0, 1\}$ (*then, next, later, suddenly, meanwhile*); $CausalConnective(s_i) \in \{0, 1\}$ (*because, so, therefore, thus*); $ContrastiveConnective(s_i) \in \{0, 1\}$ (*but, however, although*); $AdditiveConnective(s_i) \in \{0, 1\}$ (*also, in addition, furthermore*). We also compute sentence-onset connectives—which often signal event sequencing: $OnsetConnective(s_i) = 1$ if s_i begins with any connective (signals event sequencing).

Continued on next page

Table C1 – *Continued from previous page*

Category	Measure	Description	Implementation Details
	Rhetorical Structure Theory (RST) relations	Provides a principled representation of how adjacent text units relate (e.g., *Elaboration, Sequence, Cause, Contrast*).	(1) Run a neural RST parser on each story; (2) For each sentence pair, extract the predicted relation $RSTRelation(i)$ consisting of Elaboration, Sequence, Cause, Contrast; (3) Collapse into flow-promoting vs. flow-disrupting relations: Flow-promoting (*Sequence, Elaboration, Cause, Joint*); Flow-disrupting (*Contrast, Background, Explanation/Boundary*).
	Tense distance (temporal continuity)	Measures temporal continuity by quantifying dissimilarity in verb tense usage between adjacent sentences. Higher distance indicates tense shifts that may mark new narrative segments or temporal jumps.	Using spaCy POS tagging: (1) Extract all verb and auxiliary verb fine-grained tags from each sentence; (2) Compute tense distance using Jaccard distance: $TenseDistance(i) = 1 - \frac{\lvert T(s_i) \cap T(s_{i+1}) \rvert}{\lvert T(s_i) \cup T(s_{i+1}) \rvert}$ where $T(s)$ = set of verb tense tags in sentence s. Distance ranges from 0 (identical tenses) to 1 (completely different tenses).
	Agent continuity (coreference-based)	Measures protagonist/agent continuity between adjacent sentences by tracking how mentions of the main agent (grammatical subject) from the previous sentence appear in the current sentence through coreference chains. Higher values indicate stronger agent tracking across sentence boundaries.	Using spaCy dependency parsing and Maverick coreference resolution model: (1) Extract grammatical subjects from previous sentence(s) using dependency parsing (tokens with 'subj' relation); (2) Identify main agent as first subject; (3) Run coreference resolution on combined text to obtain mention clusters; (4) Find clusters containing the main agent; (5) Collect all tokens from agent clusters and current sentence; (6) Compute Jaccard similarity: $AgentContinuity(i) = \frac{\lvert A(s_{i-k}) \cap W(s_i) \rvert}{\lvert A(s_{i-k}) \cup W(s_i) \rvert}$ where $A(s_{i-k})$ = set of tokens in agent coreference clusters from previous sentence(s), $W(s_i)$ = set of tokens in current sentence. Can be computed with different context windows (e.g., $k = 1$ or $k = 3$).

Continued on next page

Table C1 – *Continued from previous page*

Category	Measure	Description	Implementation Details
Sentence-level difficulty controls	Sentence length and length delta	Measures sentence length in tokens and the difference in length between current and previous sentence(s). Length variation can indicate pacing changes, with shorter sentences potentially creating faster narrative rhythm and longer sentences providing more detailed description.	Using spaCy tokenization (excluding whitespace): $Length(s_i)$ = token count; $LengthDelta_{p1}(i) = Length(s_i) - Length(s_{i-1})$; $LengthDelta_{p3}(i) = Length(s_i) - \frac{1}{3}\sum_{k=1}^{3} Length(s_{i-k})$. Positive values indicate current sentence is longer than context.
	Word rarity (rare-word rate)	Measures lexical sophistication by computing the proportion of rare content words (nouns, verbs, adjectives, adverbs) in a sentence. Uses Zipf frequency scores to identify rare words, where lower scores indicate less common words.	Using spaCy POS tagging and wordfreq library: (1) Extract content words (NOUN, VERB, ADJ, ADV, PROPN) excluding stop words and punctuation; (2) For each content word, compute Zipf frequency score (scale 0-8, where higher = more common); (3) Count words with Zipf score < 3 as rare; (4) Compute rare-word rate: $RareRate(s_i) = \frac{\text{number of rare content words}}{\text{total content words}}$.
	Clause count (syntactic complexity)	Measures syntactic complexity by counting the number of finite clauses in a sentence. Higher clause counts indicate more complex sentence structures with multiple propositions.	Using spaCy dependency parsing: Count tokens with POS tags VERB or AUX that serve as clause markers. Include: (1) ROOT verbs (main clause); (2) Clausal complements and modifiers with dependencies: ccomp, xcomp, advcl, acl, relcl. $ClauseCount(s_i)$ = total number of identified clause markers.
	Clause tree depth (syntactic complexity)	Measures hierarchical syntactic complexity by computing the maximum depth of the dependency parse tree. Deeper trees indicate more levels of syntactic embedding and subordination.	Using spaCy dependency parsing: (1) Identify ROOT token(s); (2) Recursively traverse dependency tree from root, tracking depth at each level; (3) Compute maximum depth across all branches: $TreeDepth(s_i)$ = maximum path length from root to any leaf node. For multi-sentence input, returns maximum depth across all sentence trees.