From Stories to Statistics: Methodological Biases in LLM-Based Narrative Flow Quantification

Amal Sunny
IIIT-Hyderabad
amal.sunny@research.iiit.ac.in

Advay Gupta
IIIT-Hyderabad
advay.gupta@research.iiit.ac.in

Yashashree Chandak
Stony Brook University
ychandak@cs.stonybrook.edu

Vishnu Sreekumar
IIIT-Hyderabad
vishnu.sreekumar@iiit.ac.in

Abstract

Large Language Models (LLMs) have made significant contributions to cognitive science research. One area of application is narrative understanding. Sap et al. (2022) introduced *sequentiality*, an LLM-derived measure that assesses the coherence of a story based on word probability distributions. They reported that recalled stories flowed less sequentially than imagined stories. However, the robustness and generalizability of this narrative flow measure remain unverified. To assess generalizability, we apply *sequentiality* derived from three different LLMs to a new dataset of matched autobiographical and biographical paragraphs. Contrary to previous results, we fail to find a significant difference in narrative flow between autobiographies and biographies. Further investigation reveals biases in the original data collection process, where topic selection systematically influences sequentiality scores. Adjusting for these biases substantially reduces the originally reported effect size. A validation exercise using LLM-generated stories with "good" and "poor" flow further highlights the flaws in the original formulation of sequentiality. Our findings suggest that LLM-based narrative flow quantification is susceptible to methodological artifacts. Finally, we provide some suggestions for modifying the *sequentiality* formula to accurately capture narrative flow.

1 Introduction

The advent of Large Language Models (LLMs) over the last few years has fundamentally changed the landscape of language and cognitive research as we know it (Zhao et al., 2023). These models have gotten sophisticated to the point where the debate now is whether they display emergent properties (Wei et al., 2022). LLMs can solve real-world problems ranging from code synthesis (Nijkamp et al., 2022) to lie detection (Loconte et al., 2023). Such advances have further motivated research using these models to gain insight into human behavior and cognition (Demszky et al., 2023; Mihalcea et al., 2024). While LLMs are impressive, rigorous use of statistics and better metrics have previously demonstrated that initial claims about their capabilities were overblown (Schaeffer et al., 2023). Narrative understanding is one important area of human cognition that finds a natural application for LLMs. In this study, we rigorously assess a recently proposed LLM-derived measure of narrative flow (Sap et al., 2022) and find that methodological biases drive a large part of the original results. We discuss conceptual issues with the original formulation and propose some ways to address these biases. We expect that our analytical approach comprising both conceptual and direct replications, with appropriate randomization procedures in the evaluation pipeline, will serve as a template for future assessments of LLM-derived measures of human cognition.

Narrative understanding has historically been hindered due to difficulty in quantifying narrative elements in texts at scale (Toubia et al., 2021). Sap et al. (2022) introduced *sequentiality*, a measure of how well an LLM predicts sentences in a narrative based on the preceding sentences and the topic (i.e., *context*), compared to predictions based on the *topic* alone. This relative measure, averaged over all sentences in a story, characterizes the narrative flow, that is, the organization of the sequence of events and how well they progress from one to the next. This formulation was a breakthrough as the measure considered the entire story in its whole context, whereas previous methods relied on either detecting certain words from sentences (Sims et al., 2019; Li et al., 2013) or tracking particular elements over time in stories (e.g., sentiment, emotion, categories of words, or sentence embeddings (Reagan et al., 2016; Boyd et al., 2020; Toubia et al., 2021)). Specifically, Sap et al. (2022) captured signatures of episodic memory retrieval in narra-

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 201–215

tives by contrasting the sequentiality of a "recalled" narrative (a story that happened to person A) versus an "imagined" narrative (a story that person B made up given the topic of person A's recalled narrative). Recalled stories were less sequential compared to imagined ones, suggesting that spontaneously recalled stories deviate more from the schema of the event due to recalled events possibly triggering memories of other events. Imagined stories, on the other hand, stick to themes that are predictable from the event schema. The work of Sap et al. (2022) was seminal, inspiring multiple studies that directly incorporated their sequentiality metric (Cohen et al., 2025; Cornell et al., 2023) and numerous others (Demszky et al., 2023; Mihalcea et al., 2024; Piper et al., 2023) that built upon its findings as a foundational result.

However, LLM-based research is prone to biases (Gallegos et al., 2024; Zhou et al., 2024). A large body of recent NLP research is dedicated to evaluating biases already inherent in LLMs in different domains (Yeh et al., 2023; Huang et al., 2024) and to evaluating biases in methods and metrics that utilize LLMs (Lin et al., 2025; Hu et al., 2024; Zheng et al., 2023; Ni et al., 2024). Broadly, we can classify the biases in the field as i) methodological ones - where the bias exists in how the methods were framed, overlooking some aspects of the problem or favoring certain assumptions, and ii) data-driven biases, which arise from the dataset used and manifest in ways specific to the task at hand, influencing model predictions based on patterns, imbalances, or artifacts present in the underlying datasets (Yeh et al., 2023). We propose to evaluate Sap et al. (2022) for methodological biases, by evaluating their *sequentiality* metric as is on an entirely different curated dataset of matched autobiographies and biographies (analogous to "recalled" and "imagined" stories from Sap et al. (2022)) to see if we can capture the same difference across the two groups demonstrating properties of episodic memory retrieval. A successful replication of the results on a dataset from a conceptually similar domain would ameliorate concerns of bias by demonstrating generalizability.

To further assess methodological biases in Sap et al. (2022), we examine the generalizability of their findings across different LLMs. Since these prior experiments were conducted, the LLM landscape has evolved rapidly, with the development of significantly more advanced models (Zhao et al., 2023). These newer models have been trained on larger and more diverse datasets spanning multiple domains and incorporate key advancements such as alignment techniques (Rafailov et al., 2023; Ouyang et al., 2022). As a result, they may offer a more human-representative estimation of *sequentiality*. We experiment with several open-source and cutting-edge models such as LLaMa-3.1 (Grattafiori et al., 2024), Qwen-2.5 (Yang et al., 2024) and Falcon3 (Team, 2024) by first replicating the analysis on the original Sap et al. (2022) data to ensure parity.

Then, we consider the possibility that data-driven biases drive the original *sequentiality* results. Models applied to biased data will produce biased outcomes. We need to ensure that there are no confounds in the data that can explain the results before we make claims about complex measures such as narrative flow being a useful metric for various downstream applications. We discover a possible data-driven bias in how topics are collected and find that this bias directly influences the results. On correcting this data-driven bias using randomization techniques, we find a significantly decreased effect size from the original finding.

Apart from data-driven biases, there are concerns about the formulation itself that contains both a topic-driven term and a contextual term (that incorporates the topic in it as well). Sap et al. (2022) did not attempt to validate the formula on stories that are known to have good/poor narrative flow. Therefore, to further assess the *sequentiality* formulation, we conduct a small-scale experiment of generating stories that exhibit visibly good and poor narrative flow and apply the sequentiality measure to these stories. We do not find the expected sequentiality difference between the two group of stories. This further reinforces our concerns with the *sequentiality* formulation.

Our contributions in this study are fourfold - 1) we curate a dataset of matched autobiographical and biographical accounts and attempt to *conceptually* replicate Sap et al.'s (2022) sequentiality measure; 2) we *directly* replicate their results on the original dataset using more modern LLMs to ensure generalizability; 3) we demonstrate a methodological bias that when corrected leads to a much smaller effect size than originally claimed; 4) we demonstrate that even after removing the methodological bias emanating from the topic, the formulation does not work for stories generated with explicit good and poor narrative flow. We conclude that *sequentiality*, as originally formulated, is not

an appropriate measure to compute narrative flow of a story.

2 Methods

2.1 Datasets

2.1.1 Hippocorpus

Hippocorpus[1], used in Sap et al.'s (2022) original study, consists of 6854 stories collected by crowd sourcing from Amazon Mechanical Turk (MTurk) human workers. The dataset consists of three different categories of stories (recalled, retold, and imagined), but we only use two - the recalled and imagined stories. The recalled group consists of stories written by the workers ranging from 15-25 sentences about a memorable or salient event that they experienced in the past 6 months. The same workers also provided 2-3 sentence summaries that served as the topics of these stories. The imagined group consists of stories written by another set of workers who are given the summaries from the recalled group and told to write an imagined story about the same topics.

Hippocorpus contained more than one imagined story for some topics. We restrict our analysis to topics that had exactly one recalled and one imagined story. We obtained 2395 such matched recalled-imagined story pairs. The recalled stories had an average of 18.5 sentences and 277.5 words and the imagined stories had an average of 17.7 sentences and 240.1 words.

2.1.2 Autobiography-Biography Dataset

We collected autobiography-biography book pairs on the lives of 63 distinct individuals (126 books in total). Paragraphs in the books were embedded using *gte-base-en-v1.5* (Zhang et al., 2024). We used these embeddings to match paragraphs for thematic content across the autobiography and biography of any given personality to obtain autobiographical and biographical narratives of the same events. We retained auto-bio paragraph pairs with cosine similarity > 0.7 ensuring alignment of thematic content, while avoiding verbatim overlap using a ROUGE-L threshold < 0.4 (Lin, 2004). An auto-bio book pair was retained only if it contained > 25 pairs of paragraphs meeting the above criteria for semantic similarity. 4175 story pairs from 38 pairs (76 books) matched all the criteria and were retained for further analysis. The autobiographical paragraphs contained an average of 7.0 sentences and 116.8 words and the biographical paragraphs contained an average of 8.2 sentences and 136.0 words. Examples of matched auto-bio paragraph pairs and information about the books retained in the dataset can be found Appendix A.

2.1.3 Synthetic "Good" and "Poor" Flow Stories for Validating Sequentiality

Given that the sequentiality formulation was proposed but not validated using "ground-truth" stories by Sap et al. (2022), we attempt to validate the measure by prompting an LLM to generate stories with "good flow" and "poor flow"(see Appendix E for examples of generated stories and Appendix B.3 for the prompt used). We used *Mistral-7B-Instruct-v0.2-AWQ* (Jiang et al., 2023) (henceforth *Mistral*) to generate these stories, given randomly sampled topics from the Hippocorpus dataset. We manually verified and filtered the generated stories to ensure agreement with their respective flow labels ("good" or "poor") before computing their sequentiality scores.

As illustrated in Fig. 1, our analytical approach begins by prompting a high-performing LLM to generate a topic for each paragraph before computing the *sequentiality* of each paragraph. The details of topic generation and the sequentiality formulation are provided below.

2.2 Topic Generation

Topics were generated using *Mistral* for a given paragraph/story. The model was given a structured prompt containing a definition of topic as the "main idea of a paragraph," an example paragraph and topic (see Appendix B.1 for the example and prompt), and an instruction to return the topic in one to two sentences. We refer to this strategy as "one-shot prompting" and use this as the default topic generation strategy throughout this study unless specified otherwise. A small minority of the responses ($< 5\%$) did not strictly adhere to the instructions to return just the topic. The additional paragraphs generated, providing additional context or justification for the choice of the topic, were discarded, retaining only the generated topic.

As we will see in the next subsection, the sequentiality formula has a topic-driven term and a contextual term. In Sap et al.'s (2022) dataset, the topics are written only by the recalled group and they report that sentences in the recalled stories are better predicted by the topic compared to the sentences in

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

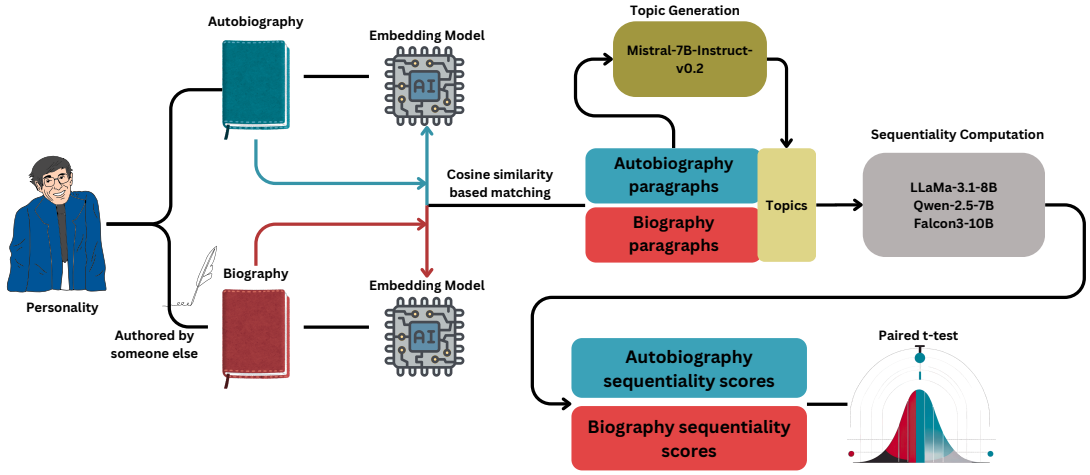

Figure 1: Overview of the methodology and analysis pipeline.

the imagined stories. Critically, they report no significant differences in the contextual term between recalled and imagined stories. To evaluate potential topic-driven bias in these results, we select topics in three different ways in our direct replication: i) exclusively generated from the autobiographical/recalled stories, ii) exclusively generated from the biographical/imagined story, and iii) generated from a story that is randomly selected from the autobiography/recalled or biographical/imagined group. The third approach guards against potential bias due to selecting topics from only one group.

To ensure that our results are not specific to the topic generation strategy described above, we repeat the analysis using an alternative method. In this approach, we generate topics using a zero-shot strategy, where the model receives only a prompt without examples, based on the instructions in Sap et al. (2022) (see Appendix B.2 for the prompt). We evaluate the effect of this strategy on sequentiality difference across groups in the most unbiased condition (i.e., randomly selecting topics from either group) and compare it against the results we obtained with the default one-shot strategy.

2.3 Sequentiality

Following Sap et al. (2022), we use *sequentiality* to quantify the extent to which the ideas/sentences in a story proceed in a well-organized sequence. Sequentiality of a sentence s_i compares the likelihood of the sentence under two probabilistic models: i) a **topic-driven model**, which assumes that the sentence is generated conditioned only on the overarching theme (topic) of the story, denoted by $\mathcal{T}$, and ii) a **context-driven model**, which assumes that the sentence depends on both the theme $\mathcal{T}$ and the preceding sentences $s_{0:i-1}$.

Sequentiality of s_i is computed as the difference in negative log-likelihoods between the topic ($\text{NLL}_{\mathcal{T}}$) and contextual ($\text{NLL}_{\mathcal{C}}$) models:

$$\Delta\ell(s_i) = -\frac{1}{|s_i|}\Big[\underbrace{\log p_{LM}(s_i \mid T)}_{\text{topic-driven}} - \underbrace{\log p_{LM}(s_i \mid T, s_{0:i-1})}_{\text{contextual}}\Big], \quad (1)$$

where the log-probability of a sentence s given some context C (e.g., topic T and preceding sentences $s_{0:i-1}$) is the sum of log-probabilities of its tokens w_t given the same context: $\log p_{LM}(s|C) = \sum_t \log p_{LM}(w_t|C, w_{0:t-1})$; and the likelihoods are normalized by sentence length $|s_i|$. Finally, the sequentiality of a paragraph is computed by averaging the sequentiality of all the sentences in the paragraph. Higher values of sequentiality are taken to indicate that sentences are highly predictable from the topic and context of the unfolding story whereas lower values indicate greater deviation from the ideas predicted by the preceding sentences. However, we note here that true sequentiality differences between stories should be driven primarily by $\text{NLL}_{\mathcal{C}}$ with $\text{NLL}_{\mathcal{T}}$ providing a "baseline" topic-based likelihood. If results are primarily driven by significant differences in the topic-based likelihood with no differences in the contextual likelihood, as in Sap et al. (2022), the measure would be incongruent with the intuitive concept of "sequentiality" as a measure of flow from the preceding context. Sap et al. (2022)'s main argument seems to rest on the difference in effect sizes between the overall sequentiality measure and the topic-term. Therefore, we compute the effect sizes associated with overall sequentiality,

$\text{NLL}_{\mathcal{T}}$, and $\text{NLL}_{\mathcal{C}}$ to make our arguments.

We estimated the likelihoods using three different models, (i) *Meta-Llama-3.1-8B-Instruct-AWQ-INT4* (henceforth Llama-3.1), (ii) *Qwen2.5-7B-Instruct-AWQ* (henceforth Qwen-2.5), and (iii) *Falcon3-10B-Instruct-AWQ* (henceforth Falcon3), all of which are trained on extensive high-quality text corpora featuring web, code, STEM, and curated high-quality and multilingual data. We used these different models to test whether results are model-dependent. We conduct our main analysis with the model producing the closest direct replication of Sap et al.'s (2022) inferences on their own dataset.

2.4 Statistical Analysis

For comparing sequentiality and its constituent terms across the two groups, we use a paired t-test and report the t-statistic, p-value, degrees of freedom (df) and Cohen's d (effect size). We emphasize Cohen's d over p-values because given a dataset of this size, statistical significance could be trivial if not weighted properly with the corresponding effect size (Sullivan and Feinn, 2012).

3 Conceptual Replication

We computed the sequentiality of matched pairs of autobiographical and biographical paragraphs in our curated dataset and compared them using a paired t-test. We randomly pick one of the two paragraphs from a given pair of paragraphs to pass to an LLM to generate a topic for the sequentiality computation. We find no significant sequentiality differences between topic-matched biographical and autobiographical paragraphs ($t = 0.11, p = 0.90, d = 0.001, df = 4174$) using LLaMa-3.1 (Fig. 2). Even after aggregating the scores for each personality, we do not find a significant difference across biographies and autobiographies ($t = -1.22, p = 0.23, d = 0.19, df = 37$).

To evaluate whether this result was specific to the LLM we used, we experimented with two other high-performing LLMs, *Qwen-2.5* and *Falcon3* and report our results in Tab. 1. All the models display either no statistically significant differences between the groups or a statistically significant difference but with a negligible effect size. Clearly, none of the models replicate the large differences reported in Sap et al. (2022). Our conceptual replication could have failed due to one or more of three reasons - 1) the autobiographies in our dataset have potentially undergone heavy editing (or ghost writing) and multiple passes by the author resulting in a more narrativized version that doesn't contain any trace of autobiographical memory retrieval, 2) our implementation of sequentiality calculation is flawed, 3) the original analysis was biased in some way. While it is difficult to evaluate the effect of editing in the dataset, our implementation can be verified by directly applying it to the Hippocorpus dataset to replicate Sap et al.'s (2022) results. Furthermore, as part of this direct replication, we can also examine the impact of slightly different but conceptually valid methodological choices on the original results.

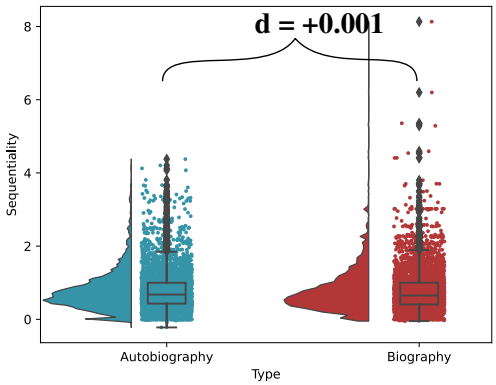

Figure 2: Distribution of sequentiality scores for topic-matched autobiographical and biographical paragraphs. d represents Cohen's d (effect size)

Model	t-stat	p-val	Cohen's d
LLaMa-3.1	0.11	0.90	0.001
Falcon3	3.05	*0.001*	0.047
Qwen-2.5	0.76	0.44	0.014

Table 1: Results of a paired t-test comparing sequentiality of biographical and autobiographical stories across three different LLMs.

4 Direct Replication

4.1 Original Topics

To verify that the failure to replicate the original result is not due to an implementation difference/error or due to specificity of the LLM used for the task, we compute sequentiality of the stories in the original Hippocorpus dataset. Using the same story pairs and topics they collected, we replicate their finding that imagined stories flow more sequentially than recalled stories

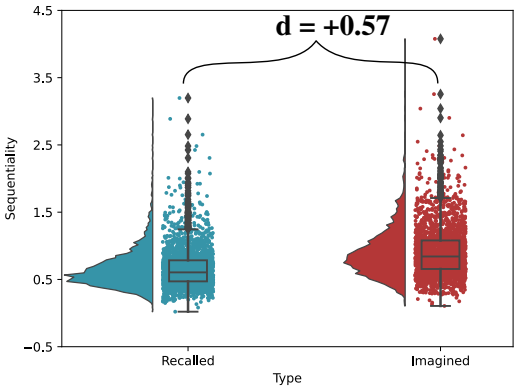

Figure 3: Distribution of sequentiality scores for recalled and imagined stories, given human-generated topics from the Hippocorpus dataset. d represents Cohen's d (effect size)

($t = 28.29, p < 0.001, d = 0.57, df = 2394$) using LLaMa-3.1 (Fig. 3). We also examine the difference between the topic-driven ($t = -19.48, p < 0.001, d = 0.39, df = 2394$) and contextual terms ($t = -2.24, p = 0.02, d = 0.04, df = 2394$) from Eq. (1) and, similarly to the original study, find a sizeable difference in the topic-driven term but not in the context-driven term. We repeat the analysis using different LLMs (Qwen-2.5 and Falcon3) and report similar results in Tab. 2.

These results indicate that the sequentiality measure generalizes to different LLMs and verifies that our implementation of sequentiality works. However, this "replication" does not fully alleviate concerns about the formulation or the potential for bias driving the original results. Sap et al. (2022) reported that the difference between the two groups was predominantly driven by the topic-driven term in Eq (1) and not the contextual term. While we replicated the same patterns, we also notice that the contextual term shows a slight difference in the opposite direction, i.e., recalled sentences flow better from the context of the unfolding story compared to the sentences in the imagined stories. This observation, combined with the fact the topic for these stories are provided only by the recalled group and not the imagined group, points towards a potential source of bias. To further assess this potential bias, we replaced the human-generated topics/summaries with LLM-generated topics but using i) only the recalled stories, ii) only the imagined stories, and iii) stories randomly sampled from the recalled and imagined conditions as input for topic generation. We expect these three conceptually equivalent ways of picking topics to yield similar results if there is no bias from the topic term.

4.2 Different Topic Generation Strategies

We find drastically different sequentiality patterns depending on how the topics are generated. We report results for LLaMa-3.1 since it displayed the largest difference (i.e. replicated Sap et al.'s (2022) results the best) between the groups in the previous section. We report results from the other models in Appendix C.

On generating topics using the imagined stories, which is a valid way of choosing topics since both recalled and imagined stories are about the same themes, we find that recalled stories are more sequential than imagined ones ($t = -4.91, p < 0.001, d = 0.10, df = 2394$; Fig. 4a), completely flipping the direction of the original results. The topic-driven differences ($t = 5.63, p < 0.001, d = 0.11, df = 2394$) and context-driven differences ($t = 3.11, p = 0.001, d = 0.06, df = 2394$) have also changed directions, compared to the results in Tab. 2.

On the other hand, when generating topics using the recalled stories as in the original work, we find that the imagined stories flow more sequentially than the recalled ones ($t = 18.27, p < 0.001, d = 0.37, df = 2394$, Fig. 4b), albeit the effect size is smaller than when using the original human-generated topics ($d = 0.37$ vs $d = 0.57$). We also observe that the directions of the topic-driven ($t = -15.44, p < 0.001, d = 0.31, df = 2339$) and context-driven ($t = -3.59, p < 0.001, d = 0.07, df = 2360$) differences replicate what were originally reported. Critically, we note here that the effect size differences between the overall sequentiality measure ($d = 0.37$) and topic-driven NLL ($d = 0.31$) are not as stark as the original findings, likely due to the topics being generated by an LLM rather than the same humans who recalled the stories. This result already calls into question the validity of the overall sequentiality measure if the observed differences are driven almost entirely by the topic term.

A conceptually equivalent way of generating topics completely flipping the results reported originally strongly indicates bias from the topic-driven term in the sequentiality measure. When the same people who recalled events from their lives also generate summaries that are used as the topics in the analysis, it should be expected that the sentences in the recalled stories would be significantly

Model	Metric	t-stat	p-val	Cohen's d	NLL_I	NLL_R
LLaMa-3.1-7b	Topic-driven $(-)$	-19.48	10^{-78}	0.39	-3.39	-3.12
	Context-driven $(+)$	-2.24	0.025	0.04	-2.48	-2.46
	Sequentiality $(+)$	**28.29**	$\mathbf{10^{-152}}$	**0.57**	0.90	0.66
Falcon3-10b	Topic-driven $(-)$	-17.41	10^{-64}	0.35	-3.81	-3.51
	Context-driven $(+)$	-2.62	0.008	0.05	-2.69	-2.66
	Sequentiality $(+)$	23.89	10^{-113}	0.48	1.12	0.85
Qwen-2.5-7b	Topic-driven $(-)$	-18.73	10^{-73}	0.38	-3.52	-3.24
	Context-driven $(+)$	-2.62	0.008	0.05	-2.59	-2.56
	Sequentiality $(+)$	27.04	10^{-141}	0.55	0.93	0.67

Table 2: Sequentiality comparison of imagined and recalled stories using different LLMs and human-generated topics taken from the Hippocorpus dataset. NLL_I & NLL_R are the mean values for the negative log likelihood of imagined and recalled stories, respectively. $(+)$ and $(-)$ indicate the expected direction of difference for the metric to replicate Sap et al.'s (2022) result. For example, Sequentiality $(+)$ indicates that a positive value in that row replicates the finding that imagined stories flow more sequentially than recalled ones. LLaMa is chosen for further analyses based on the strongest replication effects identified in **bold**. Additionally, our requirement that the contextual term drives the effect is also indicated by a $(+)$. Extremely small p-values are approximated to the closest power of 10.

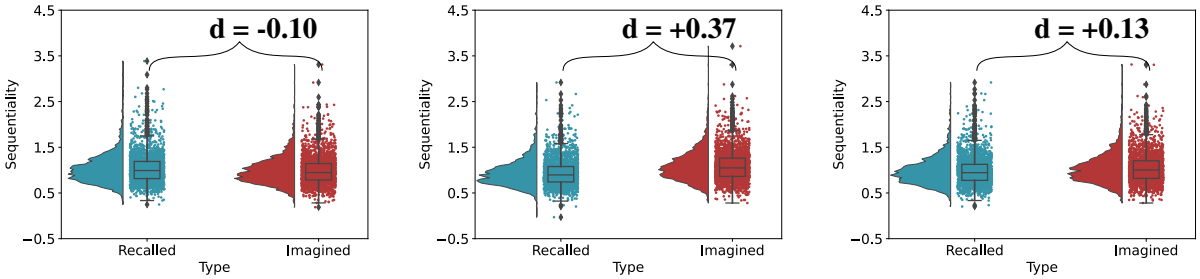

Figure 4: Distribution of sequentiality scores for recalled and imagined stories, given topics generated from the imagined (a), recalled (b), and random stories (c), respectively. A positive and large effect size (**d**) replicates Sap et al. (2022).

better predicted from the topics than those in the matched imagined stories. Since the topic also goes into the contextual term in Eq (1), the bias not only affects the topic-driven term but also the contextual term, explaining why the components also change directions when generating topics differently. However, the topic-driven term (which contributes negatively to the sequentiality formula) exerts a stronger influence and drives a decrease in sequentiality for the group from which topics were generated and trivially explains the original result of Sap et al. (2022).

We address the topic-driven bias by randomly sampling a topic either from the recalled or the imagined story. Now, we find that while imagined stories are still more sequential than recalled ones ($t = 6.62, p < 0.001, \mathbf{d = 0.13}, df = 2394$; Fig. 4c), the difference has a much lower effect size compared to the difference obtained by using the original human-generated topics ($d = 0.13$ vs $d = 0.57$, respectively). Furthermore, there is no context-driven difference ($t = -0.23, p = 0.81, d = 0.004, df = 2394$) and sequentiality difference is almost entirely topic-driven ($t = -4.56, p < 0.001, d = 0.09, df = 2394$). Repeating this analysis with a different topic generation strategy yields similar results (see Appendix D for details). Yet again, when not using human-generated biased topics, the effect sizes of the topic-driven term and the overall sequentiality measure are comparable, unlike those reported in Tab. 2, challenging Sap et al.'s (2022) claim that the overall measure explains narrative flow differences to a much greater extent than the topic-only term.

5 Sequentiality in LLM-generated Narratives

All experiments conducted thus far indicate a bias in the formulation primarily from the topic term. Here, we validate the sequentiality measure by applying it on a synthetic dataset of stories with "good" and "poor" flow (see Sec. 2.1.3).

Surprisingly, sequentiality of the stories generated to have good flow is lower than those with poor flow ($t = -12.59, p < 0.001, d = 0.51, df = 597$). Critically, we find that the contextual term behaves as we would expect intuitively: stories with good flow have a higher contextual likelihood than those with poor flow ($t = 10.39, p < 0.001, d = 0.42, df = 597$). However, the topic-driven likelihood ($t = 14.93, p < 0.001, d = 0.61, df = 597$) cancels out the contextual effect and drives the overall measure in the opposite direction. Therefore, in this validation exercise as well, the topic-driven term is the driving force masking an actual observable and meaningful effect from the contextual term. These results suggest that the topic-driven sentence likelihoods are meaningful since even LLMs generate stories that are more predictable from the topics under a "good flow" instruction. However, a standalone topic term in the formula may be masking effects that are adequately captured by the contextual term, which already incorporates the topic as part of its context. Therefore, a modified sequentiality formula with only the contextual term may be a more appropriate measure of narrative flow, especially when the comparison is between stories about the same topics.

6 Conclusion

To summarize, in the current study, we curated a novel dataset of matched autobiographies and biographies and applied Sap et al.'s (2022) sequentiality measure but using more modern open-source LLMs and failed to find a significant difference in narrative flow between the two. We then directly replicated Sap et al.'s (2022) results using the same LLMs applied to the original dataset. To understand why the narrative flow differences were so stark in Sap et al.'s (2022) data but not in our carefully curated dataset of matched autobiographical and biographical paragraphs, we analyzed both the data and the formula in (Sap et al., 2022) and report a bias in how the topics were collected which directly influenced the results through the narrative flow formula. We corrected for this bias and found a significantly reduced effect size for Sap et al.'s (2022) original finding. We further curated an LLM-generated dataset of stories with good and poor narrative flow and demonstrated the same topic-driven bias that pushed sequentiality in the opposite direction from the expected pattern. However, the contextual term captured the expected difference in narrative flow, suggesting that the sequentiality formula modified to have only the contextual term may be adequate. Importantly, this modified formulation would indicate that there are no real narrative flow differences between recalled and imagined stories, contrary to Sap et al.'s (2022)'s claims. We confirmed that our findings were not influenced by the specific LLM we used and conclude that LLM-based sequentiality, as originally formulated, is not a suitable metric for analyzing narrative flow.

These results have direct implications for studies that use sequentiality as a measure. Cohen et al. (2025) evaluated the relationship between the readability of medical texts and sequentiality. Rather than considering sequentiality as a whole, they analyzed the topic term and contextual term separately. Their results showed that the contextual term (referred to as the chain model) performed best and was primarily used in their analysis. This approach avoids topic bias and provides further evidence in favor of a modified formulation that includes only the contextual term. Cornell et al. (2023) conducted a similar study to Sap et al. (2022), comparing sequentiality across groups, with the addition of a new group consisting of generated stories based on topics from Sap et al.'s (2022) dataset. Their goal was to evaluate LLM-generated storytelling in comparison to human storytelling to better understand underlying memory processes. They reported multiple significant differences across groups (imagined, recalled, generated zero-shot, generated few-shot), but only in terms of overall sequentiality, not its components (topic and contextual terms). These differences may stem from how closely the stories align with the topic rather than from the overall cohesion of the story and its ideas, indicating the need for further investigation.

More generally, we recommend that future research on LLM-derived measures adopt proper randomization of aspects (such as topic-generation) that enter the formula and can potentially bias results. It is also important to provide an independent validation of such measures before they are adopted widely, to prevent accumulating biased methods of

assessing/measuring human cognition from LLMs that have been argued to be cognitively implausible (Connell and Lynott, 2024).

Limitations

Due to computational limitations, we were unable to evaluate the generalizability of the sequentiality metric on larger models. However, given that the disparity in computational capabilities across model sizes has been narrowing and that all the models we utilized surpass GPT-3—the largest model examined by (Sap et al., 2022), this limitation may be less consequential. Furthermore, we were able to reproduce the results across all the models we considered, ameliorating concerns about generalizability of our findings. We were also unable to extensively evaluate different LLMs for topic generation. However, we found that our topic generation methods replicated the original results, when applied to the recalled stories, and also helped demonstrate the issue of topic-driven bias. While we speculated that the contextual term in the formula by itself should work as a measure of narrative flow, a finer-grained investigation of how the sequentiality formula can be modified, possibly by incorporating weights for the topic and contextual terms, may be warranted but is beyond the scope of the current work. A more carefully curated synthetic or actual dataset with ground-truth flow is necessary to accurately assess measures of narrative flow. We attempted to validate the sequentiality measure using stories generated by an LLM to have good and poor flow and showed that the contextual term can indeed capture narrative flow differences. However, further validation of the idea that the contextual term is sufficient by itself through more elaborate experiments is required.

Code and Data

The code and data extraction methods are available at `https://github.com/mandalab/narrative-flow-autobio`. To recreate the autobiography-biography dataset, we provide a pipeline to extract and match paragraphs from the books, provided a text file version of all these books. Other datasets can be accessed from their referenced locations.

Acknowledgments

We would like to thank Aruneek Biswas, Pooja R, Pritha Ghosh, and Gargi Shukla for providing useful feedback on earlier versions of the manuscript. We also thank IIIT Hyderabad for the Faculty Seed Fund (VS, IIIT/R&D Office/Seed-Grant/2021-22/018).

References

Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science advances*, 6(32):eaba2196.

Trevor Cohen, Weizhe Xu, Yue Guo, Serguei Pakhomov, and Gondy Leroy. 2025. Coherence and comprehensibility: Large language models predict lay understanding of health-related content. *Journal of Biomedical Informatics*, 161:104758.

Louise Connell and Dermot Lynott. 2024. What can language models tell us about human cognition? *Current Directions in Psychological Science*, 33(3):181–189.

Charlotte Cornell, Shuning Jin, and Qiong Zhang. 2023. The role of episodic memory in storytelling: Comparing large language models with humans. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis,

Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Position: TrustLLM: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 598–604.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025. Investigating bias in llm-based bias detection: Disparities between llms and human perception. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649.

Riccardo Loconte, Roberto Russo, Pasquale Capuozzo, Pietro Pietrini, and Giuseppe Sartori. 2023. Verbal lie detection using large language models. *Scientific reports*, 13(1):22849.

Rada Mihalcea, Laura Biester, Ryan L Boyd, Zhijing Jin, Veronica Perez-Rosas, Steven Wilson, and James W Pennebaker. 2024. How developments in natural language processing help us in understanding human behaviour. *Nature Human Behaviour*, 8(10):1877–1889.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Haiquan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. In *International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Andrew Piper, Hao Xu, and Eric D Kolaczyk. 2023. Modeling narrative revelation. In *CHR*, pages 500–511.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, 5(1):1–12.

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.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3623–3634.

Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Falcon-LLM Team. 2024. The falcon 3 family of open models.

Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. 2021. How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, 118(26):e2011695118.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating interfaced llm bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 292–299.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jianlong Zhou, Heimo Müller, Andreas Holzinger, and Fang Chen. 2024. Ethical chatgpt: Concerns, challenges, and commandments. *Electronics*, 13(17).

A Dataset: Autobiography and Biography

A.1 Selected Books

Tab. 3 lists all the books with their respective ISBN codes, retained in the autobiography-biography dataset after matching and filtering as described in Sec. 2.1.2.

A.2 Examples of Auto-Bio Paragraph Pairs Describing the Same Events

Listing 1 and Listing 2 are examples of matching autobiographical and biographical paragraphs narrating the same events from Nelson Mandela's and Luis Suarez's lives, respectively.

```
Personality: Nelson Mandela
AUTOBIOGRAPHY:
Mr. de Klerk decided to gamble. He announced that as a result
of the by-election in Potchefstroom he would call a
nationwide referendum for March 17 so that the white people
of South Africa could vote on his reform policy and on
negotiations with the ANC. He stated that if the referendum
was defeated, he would resign from office. The referendum
asked a plain and direct question of all white voters over
the age of eighteen: "Do you support the continuation of the
reform process which the state president began on 2 February
1990 which is aimed at a new constitution through
negotiation?"

BIOGRAPHY:
Rather than delay the matter any further, within days of the
Potchefstroom defeat de Klerk announced the holding of a
white referendum on the reform process on 17 March. A clear
majority in favour, he said, would obviate the need for the
government to return to the white electorate. The risks he
took were considerable. A defeat for the government would
almost certainly have led to civil war. All the resources of
the white establishment were thrown into securing a 'Yes'
vote. De Klerk explicitly promised the electorate that he
would prevent majority rule. The National Party's advertising
campaign was based on the slogan, 'Vote Yes, if you're scared
of majority rule'. The Conservative Party campaigned for a
'No' vote, claiming that the choice for Afrikaners was
between 'the survival of the Afrikaner volk and two cars in
the garage'. The ANC did its best to support the 'Yes'
campaign, reassuring whites about its good intentions in the
negotiating process and promising job security for civil
servants.
```

Listing 1: Sample paragraphs from Nelson Mandela's autobiography and biography pertaining to the same theme.

```
Personality: Luis Suarez

AUTOBIOGRAPHY:
What came next was Ghana in the quarter-final and a penalty
shoot-out to send Uruguay through to the semi-final for the
first time in forty years. After the sending off for having
'saved' a goal on the line, I watched Asamoah Gyan miss his
penalty from the entrance to the tunnel. I was in tears,
heartbroken, but when the ball flew over the bar, I ran off
celebrating. Maybe it had been worth it.

BIOGRAPHY:
It was an instinctive move and it did the job. Ghana hadn't
scored; Uruguay were still in the tournament but now down to
ten men, as Luis was sent off. Of course, Ghana still had the
chance to seal that semi-final spot from the consequent
penalty but Asamoah Gyan could only watch in despair as his
spot-kick hit the bar and bounced away to safety.
```

Listing 2: Sample paragraphs from Luis Suarez's autobiography and biography pertaining to the same theme.

B Prompts

B.1 Prompt for Topic Generation - One Shot

Listing 3 shows the prompt we used to generate topics given a story/paragraph. The prompt consists of an example paragraph and its corresponding topic (topic manually annotated, refer Listing 4 for the example paragraph and topic) to condition the LLM (*Mistral*) on the expected form of topic generation - which is commonly known as one-shot prompting. Subsequently, the prompt accepts the given story/paragraph as input and generates a topic based on it.

```
<s>[INST] A topic is the main idea discussed in a
text.
Identify the topic for the paragraph. Return only
the topic in 1-2 sentences.
Example:
Paragraph: The man who explained the difference ...
[/INST]
Topic: Bruno Silva's vital support in Groningen.
</s>
[INST] Identify the topic for the given paragraph.
Return a single topic which is most relevent for
the paragraph. Return only the topic in 1-2
sentences with no additional text or information.
```

```
Paragraph: {paragraph_text} [/INST]
```

Listing 3: Structured prompt for topic generation for stories/paragraphs using the Mistral-7B model

```
Example
Topic: Bruno Silva's vital support in Groningen.

Paragraph: The man who explained the difference between net
and gross to me, and much more besides, was Bruno Silva.
Bruno would be our salvation at Groningen in those first few
months. I remembered him as a Uruguayan international and as
a player for Danubio - one of the third teams in Uruguay
along with Defensor, behind Nacional and Penarol. We used to
get together to watch games from the Uruguayan league or we
would meet up for family barbecues. We couldn't find any
Uruguayan steak so we managed to get hold of some Brazilian
meat instead from a Brazilian who had played for many years
in Groningen called Hugo Alves Velame. He was coaching in the
academy at that point and he was someone else who was great
with Sofi and me, becoming our translator whenever we had to
deal with the club.
```

Listing 4: Example of the topic and paragraph pair used in the prompt for topic generation

B.2 Prompt for Topic Generation - Zero Shot

Listing 5 shows the prompt we used to generate topics given a story/paragraph, in a zero shot setting. Unlike the previous prompt, there is no example given and the instructions correspond to those provided to the participants of Sap et al.'s (2022) study to collect the summary. The prompt accepts the given story/paragraph as input and generates a topic based on it.

```
<s>[INST] You are given a story or event that has
happened in the given paragraph. Come up with a
short summary of the event (2-3 sentences), written
with enough details that you will remember what you
wrote about in the future. Return only the summary
in the format of "Topic: <summary>" with no
additional text.
"Paragraph: {paragraph_text}" </s>[/INST]
```

Listing 5: Zero shot structured prompt for topic generation for stories/paragraphs using the Mistral-7B model

B.3 Prompt for Story Generation

Listing 6 details the prompt template utilized for generating stories with "Good" narrative flow. The prompt for "Poor" flow replaces the "Good" for "Poor" in the instructions on generating the story, but is identical otherwise.

```
<s>[INST] Narrative flow refers to the logical
progression of a story with smooth transitions
between events. A story with good narrative flow
implies it is well structured, predictable and
transitions smoothly around the events detailed,
while a story with poor narrative flow implies it
would be poorly structured, unpredictable and with
abrupt transitions across events making it harder
to understand. Generate a single distinct short
story based on the given topic - one with good
narrative flow
Create one story, written in around 200-300 words
and 15-20 sentences on the given topic and
structure it as follows:

Topic: {p1}
Story (Good Flow): <insert story here>

Do not write anything but the given template
above before or after the story. [/INST]
```

Listing 6: Structured prompt given to the Mistral-7B for generating a story with "good" narrative flow

C Sequentiality Computation of Recalled and Imagined Stories, Given Randomly Sampled Topics: Comparison Across LLMs

We report the results of sequentiality computation from Sec. 4.2 applied to all the models in Tab. 4. The other models produce similar results as LLaMa-3.1, with much smaller effect sizes for sequentiality differences between recalled and imagined stories when using randomly sampled topics compared to human-generated topics from the recalled stories (Qwen-2.5: $d = 0.12$ vs $d = 0.55$; Falcon3: $d = 0.05$ vs $d = 0.48$; compared against the direct replication reported in Tab. 2). We again note that this difference is primarily driven by the topic term, with the contextual term having a negligible effect. Additionally, we observe that Falcon reveals negligible sequentiality differences between recalled and imagined stories when using unbiased topics even though it yields a moderate effect size for the original replication in Tab. 2.

D Sequentiality Computation of Recalled and Imagined Stories, Given Randomly Sampled Topics: Comparison Across Topic Generation Strategies

To ensure the robustness of our results, we re-evaluate the sequentiality computation and comparison across the groups for randomly sampled topics using a different strategy for topic generation. We utilize the zero-shot prompting strategy detailed in Sec. 2.2. The results follow the same trends as using one-shot prompting in Sec. 4.2, with imagined stories still being more sequential than recalled stories ($t = 7.78, p < 0.001, \mathbf{d = 0.15}, df = 2394$) but with a much lower effect size compared to the difference obtained by using the original human-generated topics ($d = 0.15$ vs. $d = 0.57$) and comparable to one-shot prompting ($d = 0.15$ vs. $d = 0.13$). The sequentiality difference is almost entirely topic-driven ($t = -6.65, p < 0.001, d = 0.13, df = 2394$), but there is now a small difference between the context terms ($t = -2.27, p = 0.02, d = 0.04, df = 2394$).

E Generated Stories

Listing 7 shows examples of a few generated stories, demonstrating clear narrative flow differences between “good” and “poor” flow stories.

```
TOPIC: I had to have a thyroidectomy for thyroid
cancer. I ended up fracturing my wrist on the way
to surgery and then having a complication of a
seroma on my neck after surgery.

Good Flow: I had to have a thyroidectomy, a routine
surgery to remove my thyroid due to cancer. The day
of the procedure arrived, and I felt a strange
sense of calm. I had prepared myself mentally for
this moment. The sun was just starting to rise as I
stepped out of my house, ready to embark on this
new journey. But little did I know, my day was
about to take an unexpected turn. As I walked to
the car, I tripped over a loose stone and fell
hard, fracturing my right wrist. The pain was sharp
and intense, but I pushed it aside, knowing I
couldn't let my fear of surgery be outweighed by a
simple injury. At the hospital, I was reassured by
the medical staff. They assured me that my wrist
would heal, and the surgery would go on as planned.
I was reassured, but deep down, I couldn't shake
off the feeling of unease. The surgery went
smoothly, and I was relieved to wake up in
recovery, feeling groggy but otherwise fine. But
then, a few days later, I noticed a strange lump
forming on my neck. It was a seroma, a pocket of
fluid that had collected under the incision site. I
was prescribed antibiotics and given instructions
to keep the area clean. But as the days passed, the
seroma grew larger, and I began to feel a sense of
panic. I called my doctor, who scheduled an
emergency appointment. At the hospital, they
drained the seroma, and I was given more
antibiotics. This complication added to my recovery
time, but I remained determined. I knew that, with
time and patience, I would get through this.
Despite the setbacks, I remained hopeful. I knew
that the road to recovery would be long, but I was
ready to face it head-on. With each day that
passed, I grew stronger, both physically and
mentally. And though the memories of that fateful
day still lingered, I knew that I would emerge from
this experience stronger and more resilient than
ever before.


Poor Flow: I had to have a thyroidectomy for
thyroid cancer. The day of the surgery, I woke up
early and prepared myself. I packed my bag, ate a
light breakfast, and headed out the door. As I was
walking to the hospital, I tripped over a crack in
the pavement and fractured my wrist. I was shocked
and in pain, but I knew I couldn't let this stop me
from getting the surgery I needed. I made it to the
hospital and checked in. The surgery went well, but
a few days later, I developed a serious
complication. A seroma formed on my neck, a pocket
of fluid that caused discomfort and made it
difficult to move. I went back to the hospital, and
they drained the seroma several times. But it kept
coming back. The doctors were baffled, and I was
frustrated. I couldn't understand why this was
happening. The fractured wrist was a setback, but I
had expected to recover smoothly from the surgery.
But now, with this new complication, I was unsure
of what to expect. The days turned into weeks, and
the seroma refused to go away. I went through
numerous treatments and procedures, but nothing
seemed to work. I was losing hope, feeling like I
was stuck in a never-ending cycle of pain and
discomfort. Just when I thought things couldn't get
any worse, I received some unexpected news. The
pathology report from the thyroidectomy showed that
the cancer had spread. My world came crashing down
around me. I was devastated, feeling like I had
been dealt a terrible hand. The fractured wrist and
seroma seemed insignificant compared to this new
revelation. I was left with so many questions and
so few answers. The journey ahead was uncertain,
and I was unsure if I had the strength to face it.

TOPIC: I went to the fair with some of my friends
and my partner and we had a picture-perfect day.
Everything from the food, to the weather, to the
crazy fair art was amazing.

Good Flow: I went to the fair with my partners hand
in mine and my friends by my side, the sun casting
a golden glow over the vibrant scene before us. The
air was filled with the sweet scent of funnel cakes
and popcorn, and the sounds of laughter and
excitement echoed through the park. We began our
day with a leisurely stroll along the midway,
taking in the sights and sounds. My friends and I
marveled at the carnival games, our partners
teasing us as we tried our hand at ring toss and
balloon darts. The sun continued to shine, the
weather perfect, as we made our way to the food
stands. We shared a picnic blanket beneath the
shade of a large tree, enjoying the flavors of corn
dogs, cotton candy, and funnel cakes. My friends
and I reminisced about our childhood memories of
fair visits, while my partner shared stories of
their own. The day went on in this blissful manner,
with each event seamlessly leading to the next. We
explored the fair art exhibit, where our jaws
dropped in awe of the creative masterpieces. The
sun began to set, and as we watched the sky turn
pink and orange, we knew it was time to leave. As
we gathered our things and said our goodbyes, we
couldn't help but feel grateful for the
picture-perfect day we had shared together. The
memories would stay with us long after the fair, a
testament to the beauty of friendship and love.

Poor Flow: I went to the fair with some of my
friends and my partner. We had an amazing roller
coaster ride to start the day. Suddenly, we were at
the cotton candy stand. My partner wanted to buy
some, but we had lost track of my friends. I went
to look for them, but they were nowhere to be
found. Suddenly, it started raining. We took
shelter under a tree, but my partner was insistent
that we should go back home. I disagreed, and we
decided to continue our fair day. But as we were
walking, we came across a petting zoo. My partner
wanted to see the animals, but we ran out of
tickets. Suddenly, we heard a loud announcement
that the Ferris wheel was malfunctioning. We
decided to check it out and found a long line, but
my partner was impatient and insisted on cutting in
line. The crowd protested, but we managed to get
on. As we were enjoying the view, suddenly, the
rain stopped, and the sun came out. We left the
Ferris wheel and went to the art stand. My partner
wanted to buy a painting, but we couldn't decide
which one to choose. Suddenly, my friends appeared
out of nowhere and helped us make a decision. We
all had fun at the carnival games, but then, my
partner got sick and we had to leave. The day ended
abruptly, and we didn't even get to try the funnel
cakes. The fair day was full of unexpected twists
and turns, and the day ended as suddenly as it had
begun.
```

Listing 7: Two example story pairs, consisting of "good" and "bad" flow stories generated by Mistral-7B.

Name	Autobiography	Biography
Isaac Asimov	9780553569971	9780810831292
Diane von Furstenberg	9781451651546	9780062041234
Lucille Ball	9781101667088	9781504018920
Muhammad Ali	9781631680496	9780791081563
Anne Frank	9780553577129	9781408842119
Ansel Adams	9780316437011	9781620405550
Bruce Springsteen	9781501141515	9781101606247
Andrew Carnegie	9789354203503	9781101201794
Fidel Castro	9781416562504	9780745630069
Julia Child	9780307277695	9780307762856
Winston Churchill	9781587315367	9780805023961
Jacques Cousteau	9780792267966	9780307378279
Alex Ferguson	9780340919408	9780224083072
Richard Feynman	9780393355628	9781453210468
Benjamin Franklin	9781508475095	9780684807614
John Kenneth Galbraith	9780345303233	9781466893757
Mahatma Gandhi	9780486245935	9780307269584
Billy Graham	9780061171062	9780849917028
Che Guevara	9781644210963	9780802197252
Buster Keaton	9780306801785	9781497602311
Henry Kissinger	9781451636468	9780698195691
Langston Hughes	9781466883499	9780195146431
Niki Lauda	9781473577954	9781471192036
Malcolm X	9780345350688	9781101445273
Nelson Mandela	9780316548182	9781586489519
Michelle Obama	9781524763138	9780307958822
Paul Robeson	9780807096932	9781497635364
Theodore Roosevelt	9781438295343	9780307777829
Elizabeth Cady Stanton	9781505923551	9780195037296
Luis Suarez	9781472224255	9781784181949
Sachin Tendulkar	9781473605190	9788174363602
Nikola Tesla	9781684222063	9781585093083
Margaret Thatcher	9780062049452	9780713992823
Mark Twain	9780520267190	9780307874597
Mike Tyson	9780007502516	9781476618029
Edith Wharton	9780684847559	9780307555854
Virginia Woolf	9781448181889	9781407066240
Paramhansa Yogananda	9781565892125	9780190668051

Table 3: List of personalities and the ISBN codes of their autobiographies and biographies used in the analysis.

Model	Metric	t-statistic	p-val	Cohen's d	NLL_I	NLL_R
LLaMa-3.1-7b	Topic-driven$(-)$	-4.56	10^{-6}	0.09	-3.58	-3.52
	Context-driven$(+)$	-0.23	0.81	0.004	-2.54	-2.54
	Sequentiality$(+)$	6.62	10^{-11}	0.12	1.03	0.98
Falcon3-10b	Topic-driven$(-)$	-2.30	0.021	0.04	-3.98	-3.94
	Context-driven$(+)$	-0.008	0.24	0.01	-2.75	-2.74
	Sequentiality$(+)$	2.65	0.008	0.05	1.23	1.20
Qwen-2.5-7b	Topic-driven$(-)$	-4.51	10^{-6}	0.09	-3.82	-3.75
	Context-driven$(+)$	-0.35	0.72	0.007	-2.62	-2.61
	Sequentiality$(+)$	6.29	10^{-10}	0.12	1.20	1.14

Table 4: Sequentiality comparison of imagined and recalled stories using different LLMs, given LLM-generated topics sampled randomly from imagined and recalled stories. NLL_I & NLL_R are the mean values for negative log likelihood of imagined and recalled stories, respectively. $(+)$ and $(-)$ indicate the expected direction of difference for the metric to replicate Sap et al.'s (2022) result. For example, Sequentiality $(+)$ indicates that a positive value in that row replicates the finding that imagined stories flow more sequentially than recalled ones. Additionally, our requirement that the contextual term drives the effect is also indicated by a $(+)$. Extremely small p-values are approximated to the closest power of 10.