Context is Enough: Empirical Validation of *Sequentiality* on Essays

Anonymous ACL submission

Abstract

Recent work has proposed using Large Language Models (LLMs) to quantify narrative flow through a measure called sequentiality, which combines topic and contextual terms. A recent critique argued that the original results were confounded by how topics were selected for the topic-based component, and noted that the metric had not been validated against ground-truth measures of flow. That work proposed using only the contextual term as a more conceptually valid and interpretable alternative. In this paper, we empirically validate that proposal. Using two essay datasets with human-annotated trait scores, ASAP++ and ELLIPSE, we show that the contextual version of sequentiality aligns more closely with human assessments of discourse-level traits such as Organization and Cohesion. While zero-shot prompted LLMs predict trait scores more accurately than the contextual measure alone, the contextual measure adds more predictive value than both the topic-only and original sequentiality formulations when combined with standard linguistic features. Notably, this combination also outperforms the zero-shot LLM predictions, highlighting the value of explicitly modeling sentence-to-sentence flow. Our findings support the use of context-based sequentiality as a validated, interpretable, and complementary feature for automated essay scoring and related NLP tasks.

1 Introduction

Large Language Models (LLMs) and measures derived from them have been used to investigate human behavior and cognition (Demszky et al., 2023; Mihalcea et al., 2024). However, despite their growing use in cognitive science and education research, LLM-derived measures require careful validation to avoid misinterpretation and bias. In the context of narrative understanding, Sap et al. (2022) introduced the concept of *sequentiality*, a measure combining a topic-based term and a contextual term to quantify narrative flow. A recent critique (Sunny et al., 2025) showed that the topic term introduced a confound: topics used in its computation were drawn only from autobiographical narratives, biasing comparisons across story types. When topics were sampled uniformly across story types, the reported differences diminished substantially. That work proposed a simplified, contextual version of the metric and called for empirical validation using data with human annotations. In this paper, we test and validate both components of the sequentiality formula using two essay datasets, ASAP++ (Mathias and Bhattacharyya, 2018) and ELLIPSE (Crossley et al., 2023), that include trait-level human scores for features aligned with narrative structure, such as Organization and Cohesion. Our results confirm that the contextual term captures narrative flow as previously hypothesized and that prior conclusions based on the full metric were likely affected by methodological artifacts.

To examine the validity of the contextual metric, Sunny et al. (2025) conducted a preliminary test using LLM-generated paragraphs with intentionally high and low narrative flow. The original sequentiality score failed to differentiate between them, while the contextual version captured the difference. These findings suggest that narrative flow is better described by how each sentence builds on its preceding context rather than by its topical similarity to a fixed set of story prompts.

To evaluate this hypothesis more rigorously, we use the ASAP++ and ELLIPSE datasets, which contain essays written in response to diverse prompts and annotated by human raters along traits such as Organization and Cohesion. These traits assess logical progression, sentence-level connectivity, and thematic development, which are elements that are conceptually linked to narrative flow. Both datasets have been widely used, with a few prior works focusing specifically on these trait-level scores (Doi

et al., 2024; Ding et al., 2024). However, no previous study has evaluated LLM-based sequentiality measures against these human-annotated traits.

In addition to sequentiality, we consider linguistic features commonly used in automated essay scoring, such as word/sentence count and number of lemmas (Hou et al., 2025). These features are known to correlate with general language abilities and content knowledge and have been effective in previous essay scoring models. However, they are only indirectly related to traits like Organization and Coherence, which depend on sentence-to-sentence flow. We evaluate whether sequentiality-based measures provide added predictive value beyond these linguistic baselines. Specifically, we compare the performance boost offered by the original sequentiality metric, its contextual variant, and its topic-only variant when combined with these features. This analysis addresses whether sequentiality captures distinct, conceptually relevant information about discourse flow that is not explained by surface-level linguistic features.

Our contributions in this study are threefold: 1) We validate the claim made by Sunny et al. (2025) that the formulation of Sap et al.'s (2022) sequentiality is flawed, and show that using only the contextual term achieves the best fit with human-annotated trait scores; 2) We assess the performance of the corrected sequentiality measure against direct zero-shot prompting of an LLM for flow-related trait scores, and show that while the metric captures aspects of flow, it is not optimized for the target trait in the way the prompted LLM response is; 3) We demonstrate that the inclusion of the contextual term provides a significant performance boost over established linguistic features, and does so more effectively than both the topic-only term and the original sequentiality formulation, suggesting that explicit modeling of sentence-to-sentence flow using only the contextual term improves predictive accuracy in automated essay scoring (AES) applications.

In the remainder of the paper, we describe the datasets and annotation schemes and formalize the sequentiality metrics. We then evaluate the original and reduced formulations against human ratings, compare them to established linguistic features, and assess whether the contextual term improves trait prediction in AES settings.

2 Methods

2.1 Datasets

We conduct experiments on two datasets, ELLIPSE and ASAP++. The ELLIPSE Corpus (Crossley et al., 2023) consists of 6,483 essays across 44 unique prompts. Each essay has been annotated on 6 analytical traits, of which we consider only "Cohesion" due to it being the closest trait to narrative flow. ASAP++ (Mathias and Bhattacharyya, 2018) is another popular dataset with trait-level scores across six distinct prompts. For our analyses, we focus on Prompts 1 and 2 (see Appendix A for details on prompts and traits), comprising 1,783 and 1,800 scored essays, respectively. We exclude the remaining prompts because they require examinees to incorporate external source materials, introducing direct in-text references that could confound our sequentiality metrics, and the topic part of the prompt becomes comparable in length to the essay for these prompts. We consider the trait "Organization" in this dataset as the closest match for narrative flow. ELLIPSE is a more challenging dataset for our analyses because the linguistic features used were tailor-made for ASAP++ (Hou et al., 2025). Furthermore, ELLIPSE is the bigger dataset with a greater number of prompts compared to ASAP++. Additional details and scoring rubrics are provided in Appendix A

2.2 Sequentiality

Sap et al. (2022) define sequentiality as the difference in negative log-likelihood (NLL) between two language model predictions: one conditioned on the topic alone ($\mathrm{NLL}_{\mathcal{T}}$), and one conditioned on both the topic and preceding context ($\mathrm{NLL}_{\mathcal{C}}$). For a sentence s_i, sequentiality is computed as

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

where $|s_i|$ is the number of tokens in the sentence, T is the topic, and $s_{0:i-1}$ are the preceding sentences. We define the sentence-level negative log-likelihood (NLL) under a language model (LM) as

$$\mathrm{NLL}(s|C) = -\frac{1}{|s|}\sum_{t=1}^{|s|}\log p_{\mathrm{LM}}(w_t|C, w_{0:t-1}), \tag{2}$$

where $|s|$ is the length of s, w_t is its t-th token, and C is contextual conditioning (i.e., topic T and preceding sentences). The sequentiality of a paragraph is the mean NLL across all sentences in that paragraph. Higher sequentiality values indicate that sentences are more predictable given the evolving context and fixed topic, whereas lower values signal greater divergence from the expectations set by the preceding text.

To isolate the influence of topic versus local context, we consider $\mathrm{NLL}_{\mathcal{T}}$ and $\mathrm{NLL}_{\mathcal{C}}$ separately to predict trait scores. We use the same 3 language models used in Sunny et al. (2025) to calculate these NLLs and sequentiality: LLaMa-3.1-8b-Instruct-AWQ (Llama; Grattafiori et al. (2024)), Falcon3-10b-Instruct-AWQ (Falcon; Team (2024)) and Qwen-2.5-7b-Instruct-AWQ (Qwen; Yang et al. (2024)). Please see Appendix B for details.

2.3 Validation Methodology

2.3.1 Ordinal regression and model selection

The sequentiality measure, a composite metric introduced by Sap et al. (2022), was proposed without empirical validation. To evaluate both its overall formulation and individual components, we compare them against trait-level human-annotated essay scores. Since the essay scores are ordinal and the sequentiality scores are continuous, we apply ordinal regression models where each component of the sequentiality term serves as a predictor and the human scores serve as the outcome variable.

We examine four models: a contextual model ($\mathrm{NLL}_{\mathcal{C}}$), a topic-based model ($\mathrm{NLL}_{\mathcal{T}}$), a combined model ($\mathrm{NLL}_{\mathcal{C}}$ and $\mathrm{NLL}_{\mathcal{T}}$ as separate predictors), and a sequentiality model (the composite metric in Eq. (1)). Model fit was evaluated using the Akaike Information Criterion (AIC), with lower values indicating better performance.

2.3.2 Zero-shot prompted LLM

We select the best-fitting measure based on AIC and assess its ability to predict essay trait scores compared to directly prompting an LLM. Prompts were prepared based on the rubric given to the human annotators for each dataset (see Appendix D for the prompts). We prompted the best model from Sunny et al. (2025), Llama, to generate trait scores.

We evaluate the results via a five-fold cross-validation on both datasets and report the mean Quadratic Weighted Kappa (QWK), in line with prior works (Taghipour and Ng, 2016; Ke and Ng, 2019).

Model	Features	ASAP_P1	ASAP_P2	ELLIPSE
		Organization		Cohesion
Llama	Seq	4833	5452	21871
	Topic	4823	5376	21477
	Context	4674	**5199**	**20790**
	Both	**4616**	**5179**	**20790**
Qwen	Seq	4839	5454	21870
	Topic	4821	5348	21430
	Context	4655	**5151**	**20690**
	Both	**4593**	**5138**	**20689**
Falcon	Seq	4820	5450	21874
	Topic	4828	5361	21337
	Context	4650	**5131**	**20674**
	Both	**4560**	**5099**	**20676**

Table 1: Model fit (AIC) of sequentiality variants for essay traits in ASAP++ and ELLIPSE. Lower is better. Bold indicates the lowest AIC (and second-lowest if within 50 units).

2.3.3 Incorporating linguistic features

Even if the contextual sequentiality term captures sentence-to-sentence flow, traits like cohesion and organization may not fully be captured by it. Further dimensions of the text may be required to fully explain these, for which we seek to incorporate some simple linguistic features with the goal of assessing the added benefit of sequentiality over and beyond these linguistic features. Based on Hou et al. (2025), we incorporate the 10 most impactful features in essay grading (see Appendix C for a detailed list). We compare the trait prediction performance of regression models with just the linguistic features and those with variants of sequentiality (described in Sec. 2.3.1) added to the linguistic features.

3 Results

3.1 Comparing contextual and topic-driven terms in sequentiality

Tab. 1 summarizes model fit (AIC) for sequentiality components against human-rated flow-related scores (see Appendix E.1 for additional traits). Across both datasets, $\mathrm{NLL}_{\mathcal{C}}$ consistently outperforms other terms, indicating better fit than either $\mathrm{NLL}_{\mathcal{T}}$ or full sequentiality. On ASAP, the combined model slightly outperforms $\mathrm{NLL}_{\mathcal{C}}$, but coefficient analysis reveals large differences from the original formulation: for ASAP, $(w_{\mathcal{T}} = -0.53, w_{\mathcal{C}} = 1.55)$; for ELLIPSE, $(w_{\mathcal{T}} = -0.004, w_{\mathcal{C}} = 1.23)$. These diverge from the original weights $(w_{\mathcal{T}} = -1.0, w_{\mathcal{C}} = 1.0)$ and support prior claims (Sunny et al., 2025) that the

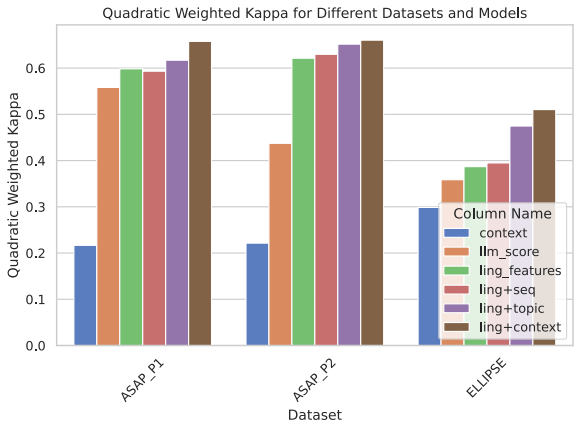

Figure 1: Comparison of QWK scores across approaches/features for each dataset's flow-relevant trait (Organization for ASAP prompts 1 and 2, Cohesion for ELLIPSE), using Llama.

context term captures flow and should drive any claimed differences in narrative flow (unlike in Sap et al. (2022), where effects were driven by the topic term). Notably, on the larger ELLIPSE dataset, $\mathrm{NLL}_{\mathcal{T}}$ adds almost no value (AIC unchanged; $w_{\mathcal{T}} = -0.004$). Similar patterns hold across LLM backbones.

3.2 Contextual term vs. zero-shot LLM

Having demonstrated the limitations of the original sequentiality formulation, we now evaluate how well the contextual term ($\mathrm{NLL}_{\mathcal{C}}$) performs relative to a zero-shot prompted LLM. To do this, we prompt the LLM directly with the essay and scoring rubric to predict trait scores. This approach is conceptually distinct from sequentiality, as the LLM is explicitly asked to optimize for the human-defined scoring criteria. On the ASAP++ dataset, the LLM achieves a QWK score of 0.49, while on ELLIPSE it scores 0.36. These results are broadly consistent with prior work (Hou et al., 2025), where prompted LLMs were shown to perform competitively on essay scoring tasks. The contextual sequentiality measure, while conceptually aligned with discourse flow, performs worse than the prompted LLM (llm_score in Fig. 1). This suggests that, as a standalone metric, $\mathrm{NLL}_{\mathcal{C}}$ may only partially capture the scoring rubric, potentially aligning with subcomponents like logical progression but not others such as completeness or relevance. Nevertheless, as we demonstrate next, this does not imply that sequentiality lacks utility as a feature in essay scoring.

3.3 Additive effect of context and linguistic features

To evaluate whether $\mathrm{NLL}_{\mathcal{C}}$ adds predictive value beyond existing baselines, we incorporate it into a standard feature-based AES model. As shown in prior work (Hou et al., 2025), simple linguistic features such as word count and lemma diversity are effective predictors of essay quality. We adopt a similar set of features (detailed in Appendix C) as our baseline model. Notably, these features alone (ling_features in Fig. 1) already outperform the zero-shot LLM on both datasets, though the performance gap narrows on the more challenging ELLIPSE dataset.

We then augment this baseline with the different sequentiality components: the topic-only term ($\mathrm{NLL}_{\mathcal{T}}$), the contextual term ($\mathrm{NLL}_{\mathcal{C}}$), and the original sequentiality (Eq. (1)). As shown in Fig. 1, all three offer improvements over the linguistic baseline, but the gains from $\mathrm{NLL}_{\mathcal{C}}$ are consistently the largest. This further supports our claims and those of Sunny et al. (2025), confirming that the contextual term is the most effective component of the sequentiality formulation for predicting human-annotated essay trait scores that are conceptually related to narrative flow. Additional results across other traits are provided in Appendix E.2, showing that the contextual term boosts performance on all traits and the overall score, consistent with prior suggestions that coherence measures are the most predictive of overall essay quality (Crossley and McNamara, 2010, 2011). $\mathrm{NLL}_{\mathcal{C}}$, therefore, is a quantitative measure that conceptually aligns more directly with human estimations of organization, coherence, and essay quality than surface-level linguistic features that are only indirectly (but strongly) indicative of essay quality.

4 Conclusion

We empirically validated the critique that the original sequentiality formulation is flawed due to its topic component (Sunny et al., 2025). Using trait-annotated essay datasets, we confirm that the context-only variant aligns better with human judgments of narrative flow. When used as a feature in automated essay scoring, this measure significantly improves predictive performance, highlighting the value of modeling sentence-to-sentence flow. These findings reinforce prior work (Crossley and McNamara, 2010, 2011) showing that local coherence is central to perceived essay quality.

Limitations

First, we rely on open-source mid-sized LLMs (LLaMa-3.1-7B, Falcon3-10B, and Qwen-2.5-7B) due to computational constraints. These models outperform GPT-3—the largest model evaluated in Sap et al. (2022) and represent a substantial improvement in reasoning and fluency. However, they may still lag behind the most capable proprietary models (e.g., GPT-4 or Claude 3), which could potentially offer even stronger contextual modeling. As such, our results may underestimate the full potential of sequentiality-based measures when used with the strongest available models.

Second, although we selected traits that are conceptually aligned with narrative flow (e.g., Organization, Cohesion), we observed performance gains across all traits and overall essay scores when including the context-based sequentiality term. While this is consistent with known correlations between trait scores that are often rated by the same human annotators within each dataset, it limits the specificity of our claims. That is, we cannot definitively conclude that context-sequentiality captures only sentence-to-sentence flow rather than broader indicators of essay quality. Future work should isolate this effect more clearly, for instance by comparing the impact of context-sequentiality against other discourse features or using targeted human judgments of local coherence and flow.

Ethics Statement

The datasets used in this study are publicly available and widely used for academic research in automated essay scoring. Both datasets are anonymized, with ASAP++ explicitly removing identifiable information. All model inference was performed using open-source LLMs run locally, eliminating the risk of data leakage and minimizing concerns around privacy.

An additional ethical consideration is that features derived from LLMs may encode biases present in their pretraining data. As a result, these features could favor certain writing styles or discourse patterns, potentially reinforcing normative preferences that disadvantage other valid forms of expression. Awareness of such biases is important when integrating LLM-derived measures into AES systems.

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Scott A. Crossley and Danielle S. McNamara. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson and R. Catrambone, editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 984–989. Cognitive Science Society.

Scott A Crossley and Danielle S McNamara. 2011. Text coherence and judgments of essay quality: Models of quality and coherence. In *33rd Annual Meeting of the Cognitive Science Society: Expanding the Space of Cognitive Science, CogSci 2011*, pages 1236–1241. The Cognitive Science Society.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

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.

Yuning Ding, Omid Kashefi, Swapna Somasundaran, and Andrea Horbach. 2024. When argumentation meets cohesion: Enhancing automatic feedback in student writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17513–17524, Torino, Italia. ELRA and ICCL.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Zhaoyi Joey Hou, Alejandro Ciuba, and Xiang Lorraine Li. 2025. Improve llm-based automatic essay scoring with linguistic features. *arXiv preprint arXiv:2502.09497*.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

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.

Amal Sunny, Advay Gupta, Yashashree Chandak, and Vishnu Sreekumar. 2025. From stories to statistics: Methodological biases in llm-based narrative flow quantification.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*.

A Dataset details

A.1 ASAP++ dataset

The dataset has 6 prompts; however, prompts 3-6 use texts from an external source and have in-text references which may potentially confound NLL calculations, thus we only use prompts 1 and 2 which can be found in Listing 1.

```
Prompt 1:
More and more people use computers, but not
everyone agrees that this benefits society. Those
who support advances in technology believe that
computers have a positive effect on people. They
teach hand-eye coordination, give people the
ability to learn about faraway places and people,
and even allow people to talk online with other
people. Others have different ideas. Some experts
are concerned that people are spending too much
time on their computers and less time exercising,
enjoying nature, and interacting with family and
friends.
Write a letter to your local newspaper in which you
state your opinion on the effects computers have on
people. Persuade the readers to agree with you.

Prompt 2:
"All of us can think of a book that we hope none of
our children or any other children have taken off
the shelf. But if I have the right to remove that
book from the shelf -- that work I abhor -- then
you also have exactly the same right and so does
everyone else. And then we have no books left on
the shelf for any of us." --Katherine Paterson,
Author
Write a persuasive essay to a newspaper reflecting
your views on censorship in libraries. Do you
believe that certain materials, such as books,
music, movies, magazines, etc., should be removed
from the shelves if they are found offensive?
Support your position with convincing arguments
from your own experience, observations, and/or
reading.
```

Listing 1: Prompts 1 and 2 from the ASAP++ dataset

For prompts 1 and 2, each essay is rated on the following five attributes:

1. **Ideas & Content**: The amount of content and ideas present in the essay.

2. **Organization**: How well structured the essay is.

3. **Word Choice**: Precision, variety, and impact of vocabulary.

4. **Sentence Fluency**: Rhythm, variation, and smoothness of sentences.

5. **Conventions**: Correctness of punctuation, spelling, grammar, and usage.

A.2 ELLIPSE dataset

Each essay receives a holistic score reflecting overall proficiency and communicative effectiveness. It is also scored across the following analytic dimensions:

1. **Cohesion**: Text organization and use of cohesive devices.

2. **Syntax**: Range and control of sentence structures.

3. **Vocabulary**: Precision, range, and appropriateness of word choice.

4. **Phraseology**: Use of idioms, collocations, and lexical bundles.

5. **Grammar**: Control of morphology and grammatical usage.

6. **Conventions**: Spelling, punctuation, and capitalization.

A.3 Scoring rubric

The rubric for the "Cohesion" trait in the ELLIPSE dataset is given in Listing 2 and for the "Organization" trait in the ASAP++ dataset can be found in Listing 3.

```
Cohesion
This property checks how well structured the essay
is.

Score 5: Text organization consistently well
controlled using a variety of effective  linguistic
features such  as reference and transitional words
and phrases to connect ideas across sentences and
paragraphs; appropriate  overlap of ideas.

Score 4: Organization generally well controlled; a
range of cohesive devices used appropriately such
as reference and transitional words and phrases to
connect ideas; generally appropriate overlap of
ideas.

Score 3: Organization generally controlled;
cohesive devices used but limited in type; Some
repetitive, mechanical, or faulty use of cohesion
use within and/or between sentences and paragraphs.

Score 2: Organization only partially developed with
a lack of logical sequencing of ideas; some basic
cohesive devices used but with inaccuracy or
repetition.

Score 1: No clear control of organization; cohesive
devices not present or unsuccessfully used;
presentation of ideas unclear.
```

Listing 2: Cohesion rubric

```
Organization
This property checks how well structured the essay
is. NOTE: Since the dataset has the essays
compressed into one line, please bear in mind that
the paragraph information is lost. Hence, give
writers the benefit of the doubt here.

Score 6: The essay is well-organized. There is a
clear flow of ideas with each idea self-contained
(this is where we assume that each idea is
contained in a paragraph). The essay has the
appropriate form as a letter to the editor.

Score 5: The essay shows good organization. There
is a flow of ideas. However, the ideas are mostly
self-contained. The essay has the appropriate form
as a letter to the editor.

Score 4: The essay shows satisfactory organization.
It contains a basic introduction, body and
conclusion.

Score 3: The essay shows some organization. Its
form may not be that of a letter to the editor. Its
ideas are not necessarily self-contained.

Score 2: Shows little or no evidence of
organization.

Score 1: The essay is awkward and fragmented. Ideas
are not self-contained.
```

Listing 3: Organization rubric

B LLM implementation

All models were implemented using `transformers` v4.46.3 (Wolf et al., 2020). Sequentiality was computed using default parameters following Sunny et al. (2025). For LLM scoring, we set the temperature to 0.0001 to ensure reproducibility, leaving other parameters at their defaults. Computation times for LLM scoring and sequentiality were approximately 24 and 40 hours, respectively, on a single RTX 2080Ti across all three datasets.

C Linguistic features

Tab. 2 summarizes the set of linguistic features used in this study. We follow Hou et al. (2025), who selected the top 10 non-prompt-specific features from state-of-the-art AES models (Ridley et al., 2021; Li and Ng, 2024). Feature extraction follows the same pipeline: total word and sentence counts are computed using `nltk` (Bird and Loper, 2004); lemma, noun, and stopword counts are obtained using `spaCy`'s `en_core_web_sm` model (Honnibal et al., 2020); and Dale–Chall difficult word count and long word count are extracted using the `readability` library[1] based on Dale and Chall (1948). See Tab. 2 for feature definitions.

[1] https://pypi.org/project/readability/

Feature	Definition
Total number of unique words	Count of word types that occur exactly once in the essay; a proxy for lexical diversity.
Total number of words	Total token count of the essay, indicating overall length and opportunity for argument development.
Total number of sentences	Number of sentence boundaries detected, used to gauge sentence complexity (e.g., average words per sentence).
Long word count	Count of long words implemeted in the Readabiliity package
Character count (non-space, non-punctuation)	Number of letters and digits only, excluding spaces and punctuation; aligns with some readability metrics.
Character count (all characters)	Total count of all characters including spaces and punctuation; reflects full essay length as stored.
Total number of lemmas	Count of base-form tokens after lemmatization, measuring conceptual variety beyond inflected forms.
Total number of nouns	Count of tokens tagged as nouns (POS = NN, NNS, etc.), indicating topical density and concreteness.
Total number of stopwords	Count of high-frequency function words (e.g., "the", "is", "and"), reflecting balance between function and content words.
Dale–Chall difficult word count	Number of words not in the Dale–Chall list of 3,000 common words (understood by 80% of fifth graders), measuring vocabulary challenge.

Table 2: Descriptions of the linguistic features.

D LLM prompts

Listing 4 and Listing 5 show the LLM prompt templates used to get "Cohesion" and "Organization" scores, respectively. Each of these has the `{rubric}`, `{topic}` and `{essay}` replaced with their corresponding rubrics (the same rubrics provided to the annotators in Appendix A.3), essay and topics from the essay being scored.

```
You are an annotator highly competent in grading
English essays. Your task is to grade the following
essay, given the topic the essay was written on and
a rubric to grade the essay.

Rubric: {rubric}
Topic: {topic}
Essay: {essay}
```

Listing 4: Prompt template for Cohesion annotation

```
You are an annotator highly competent in grading
English essays. Your task is to grade the following
essay, given the prompt used to write the essay and
a rubric to grade the essay. Additionally consider
that all the essays are anonymized. This means that
the named entities (people, places, dates, times,
organizations, etc.) are replaced by placeholders
(Eg. @NAME1, @LOCATION1, etc.). In addition to
this, capitalized phrases are anonymized as @CAP1,
@CAP2, etc. These anonymizations should not affect
your scoring. You are free to replace the
anonymizations with any placeholders.

Rubric: {rubric}
Prompt: {topic}
Essay: {essay}
```

Listing 5: Prompt template for Organization annotation

E Additional results

E.1 Model fit comparisons

Tab. 3 reports model fit comparisons across additional traits that may reflect narrative flow. The results mirror those observed for the primary traits: the contextual term alone generally provides the best fit, while the combined contextual+topic term yields marginal improvements on ASAP++, but shows negligible gains on the larger ELLIPSE dataset.

E.2 Additional trait-level performance comparisons

Fig. 2 shows performance comparisons across different input features and trait scores. Overall, the trends align with those observed in the main analysis: the contextual term underperforms direct LLM scoring, linguistic features alone outperform the LLM, and the inclusion of the contextual term yields the largest performance gain when added to linguistic features. Notably, these improvements extend even to traits not directly related to narrative flow. One possible explanation is that the sequentiality measure captures broader aspects of essay quality, which are correlated with multiple trait scores, thereby improving predictions across a wider range of dimensions.

Model	Features	ASAP P1			ASAP P2			ELLIPSE	
		Organization	Sentence fluency	Content	Organization	Sentence fluency	Content	Cohesion	Overall
Llama	Seq	4833	4919	4988	5452	5259	5600	21871	21277
	Topic	4823	4854	4981	5376	5166	5509	21477	20626
	Context	4674	**4718**	**4876**	**5199**	**4959**	**5345**	**20790**	**19714**
	Both	**4616**	**4698**	**4838**	**5179**	**4960**	**5334**	**20790**	**19714**
Qwen	Seq	4839	4920	4992	5454	5249	5599	21870	21276
	Topic	4821	4849	4979	5348	5077	5479	21430	20554
	Context	4655	**4693**	**4862**	**5151**	**4907**	**5297**	**20690**	**19563**
	Both	**4593**	**4670**	**4820**	**5138**	**4909**	**5291**	**20689**	**19563**
Falcon	Seq	4820	4908	4972	5450	5261	5600	21874	21278
	Topic	4828	4865	4985	5361	5094	5493	21337	20448
	Context	4650	4677	4849	**5131**	**4884**	**5283**	**20674**	**19550**
	Both	**4560**	**4622**	**4776**	**5099**	**4882**	**5264**	**20676**	**19549**

Table 3: Model fit (AIC) comparison of sequentiality variants—$\mathrm{NLL}_{\mathcal{T}}$, $\mathrm{NLL}_{\mathcal{C}}$, and their combination—across an extended list of essay traits in ASAP++ and ELLIPSE. Each cell reports the AIC for models using the specified features. Lower values indicate better fit. Bold indicates the lowest AIC (and second-lowest if within 50 units).

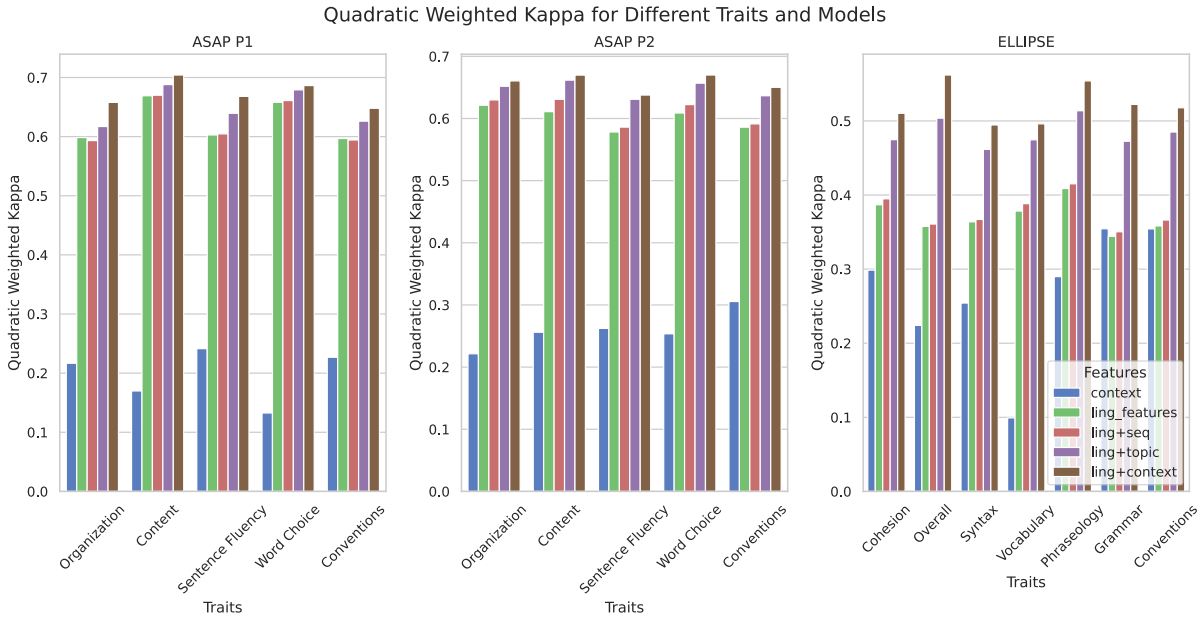

Figure 2: Comparison of QWK scores across approaches/features for all traits for both datasets, using LLaMa.