Literary Evidence Retrieval via Long-Context Language Models

Katherine Thai and Mohit Iyyer
UMass Amherst University of Maryland, College Park,
kbthai@umass.edu, miyyer@umd.edu

https://github.com/katherinethai/long_context_relic

Abstract

How well do modern long-context language models understand literary fiction? We explore this question via the task of *literary evidence retrieval*, repurposing the RELiC dataset of Thai et al. (2022) to construct a benchmark where the entire text of a primary source (e.g., *The Great Gatsby*) is provided to an LLM alongside literary criticism with a missing quotation from that work. This setting, in which the model must generate the missing quotation, mirrors the human process of literary analysis by requiring models to perform both global narrative reasoning and close textual examination. We curate a high-quality subset of 292 examples through extensive filtering and human verification. Our experiments show that recent reasoning models, such as GEMINI PRO 2.5 can exceed human expert performance (62.5% vs. 50% accuracy). In contrast, the best open-weight model achieves only 29.1% accuracy, highlighting a gap in interpretive reasoning. Despite their speed and apparent accuracy, even the strongest models struggle with nuanced literary signals and overgeneration, signaling open challenges for applying LLMs to literary analysis. We release our dataset and evaluation code to encourage future work in this direction.

1 Introduction

The emergence of long-context language models, which can process millions of tokens (Gemini Team, 2024), has unlocked new AI applications for literary analysis. In this paper, we focus on the task of *literary evidence retrieval*, in which a model must retrieve a supporting quotation from a primary source (e.g., a novel) to substantiate an excerpt of literary criticism.

Thai et al. (2022) frame literary evidence retrieval as a computational task by introducing RELiC, a dataset that contains short excerpts from published literary criticism that include quotations from famous novels. While RELiC was developed to benchmark and improve retriever models, which compute embeddings of claims and short chunks of the book text, we repurpose it as a testbed for long-context[1] language models: A model is given the entire text of the book and an excerpt of literary criticism with a missing quotation from that book, then asked to generate the missing quote (see Figure 1, top). This task does not exactly mirror the human process of literary analysis, in which a scholar typically develops claims and selects supporting quotes iteratively. However, the task requires the same understanding of and complex reasoning over plot, subtext, and other literary devices with the advantage of being easily verifiable.

To enable robust evaluation, we curated a high-quality subset of 292 examples from RELiC, verified through a combination of automated filtering and expert human review. These 292 examples contain claims that require both global reasoning over events and "close reading" over singular passages to solve. Our experiments reveal that state-of-the-art reasoning models significantly outperform previous LLMs and even surpass a human expert baseline: our best model, GEMINI PRO 2.5 obtains an accuracy of **62.5%** compared to **55.0%** for a human expert on a subset of the data. However, we also find that these models tend to overgenerate and struggle with subtle literary cues. Open-weight models perform substantially worse, suggesting that interpretive reasoning, not just long-context capacity, is essential to success in this domain, forming important directions for future research.

2 Dataset Curation

The RELiC dataset (Thai et al., 2022) consists of 78k excerpts of English-language literary analysis collected from scholarly journals where each excerpt includes a direct quotation from one of 79

[1] In this paper, we consider "long-context" to mean allowing a minimum of 128k tokens as input.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 369–380

Model + Simple Prompt	ALL (n=292)	(n=40)	(n=39)
Gemini Pro 2.5(Google, 2025)	63.7	57.5	**82.1**
o3(OpenAI, 2025b)	49.7	47.5	71.8
Gemini Pro 1.5 (Gemini Team, 2024)	36.1	22.5	43.6
o1 (OpenAI, 2024)	32.2	25.0	41.0
GPT-4o (OpenAI et al., 2024)	27.4	17.5	35.9
Qwen 2.5 Instruct (72B) (Qwen et al., 2025)	11.3	7.5	20.5
Llama 3.1 Instruct (8B) (Grattafiori et al., 2024)	5.1	5.0	2.6
Qwen 2.5 Instruct (7B)	2.7	5.0	2.6
Llama 3.3 Instruct (70B)	2.1	0.0	2.6
Model + Explanation Prompt			
Gemini Pro 2.5	**64.7**	**62.5**	79.5
GPT-4.1(OpenAI, 2025a)	51.0	47.5	69.2
o3	50.7	50.0	66.7
Gemini Pro 1.5	38.5	40.0	50.0
Claude Sonnet 3.7(Anthropic, 2025)	37.0	32.5	48.7
DeepSeek-R1(DeepSeek-AI et al., 2025)	29.1	15.0	38.5
GPT-4o	24.3	22.5	31.8
Qwen 3 (32B)(Qwen et al., 2025)	19.2	20.0	33.3
Qwen 3 (8B)	8.9	5.0	10.3
o3-mini (OpenAI, 2025c)	8.3	10.0	13.6
Baselines			
gte-Qwen2-7B-instruct	4.5	2.5	6.8
Human	-	55.0	-

Table 1: Percentage of test set examples where the model generated the correct ground truth quotation for different folds of the test set.

	Primary Sources (n=7)		Dataset Examples (n=292)		
	Tokens	Words	Tokens	Words	# Ex/Book
Mean	85,526	69,456	254.9	203.6	36.0
St. Dev.	26,6304	21,167	66.9	52.9	19.6
Max	124,544	102,549	492.0	385.0	52.0
Min	45,038	37,209	116.0	91.0	7.0

Table 2: Summary statistics for long-context RELiC. Token counts were computed with the `o200k_base` encoding via `tiktoken` (https://github.com/openai/tiktoken) and word counts were computed by splitting on whitespace.

primary source texts in the public domain. Each example in the dataset contains up to four sentences preceding the quotation and up to four sentences following the quotation. Together, these sentences of make up the "context" of the ground truth quotation. The primary source quotations may be up to five consecutive sentences in length.

2.1 Adapting RELiC for long-context reasoning

While large, RELiC is also noisy, which necessitates several filtering steps before we can evaluate models on it. We implemented an extensive data preprocessing pipeline[2] that involved several cleaning and filtering passes with GPT-4o-mini and GPT-4o supplemented by some programmatic heuristics, all geared towards mitigating the below issues:

Low data quality: Some RELiC examples contain **OCR artifacts** present in the primary source texts that render the prefixes and suffixes ungrammatical. There are also some examples that are **misclassified** as literary analysis.

Model exploits: Several aspects of RELiC examples can provide unintended cues to models, allowing them to bypass the reasoning challenge. One was the **disclosure of the location** of the quote in the prefix or suffix. Another was **quote leakage**, or the appearance of part or all of the ground truth quotation in the context. Finally, **data contamination** was a concern–the literary analysis excerpts could appear in the training data of the LLMs we benchmarked and may have been memorized, or the ground truth quotation is so prevalent in the training data (because it belongs to a public domain novel) that the model is able to retrieve it without needing the primary source text at all.

Human verification: After filtering, we create a high-quality human-verified subset of the dataset by having one of the authors, who has a degree in English literature and has read all of the primary source novels manually review 400 filtered examples. This author marked 292 examples that were well-formed instances of literary analysis with a ground truth quotation that could be identified given only the book and literary analysis context. See Table 2 for dataset statistics.

Dataset folds: Our new dataset contains two labeled folds of special data, (1) the Human Eval Set: 40 examples attempted by our author, and (2) the Close Reading Set: 39 examples labeled by our author as examples of **close reading**, a literary analysis technique in which the reader considers "linguistic elements, semantic aspects, syntax, rhetoric, structural elements, thematic, and generic references in the text" (Ohrvik, 2024). A natural consequence of this interpretive technique is the repeated citation of parts of the ground truth quote in the context. Since these examples contain lexical overlap with the ground truth quotation, models can exploit this during evidence retrieval.

3 Experimental Setup

We evaluated four closed-source and four open-weight models on the long-context RELiC dataset, prompting each in a zero-shot setting to retrieve the most fitting quotation for a literary analysis excerpt given the full primary source text. We tested two prompt types: (1) Simple, which requested only the quotation, and (2) Explanation, which first asked the model to justify its choice before

[2]All prompts, details of filtering heuristics, and descriptions of our manual validations are in Appendix A.

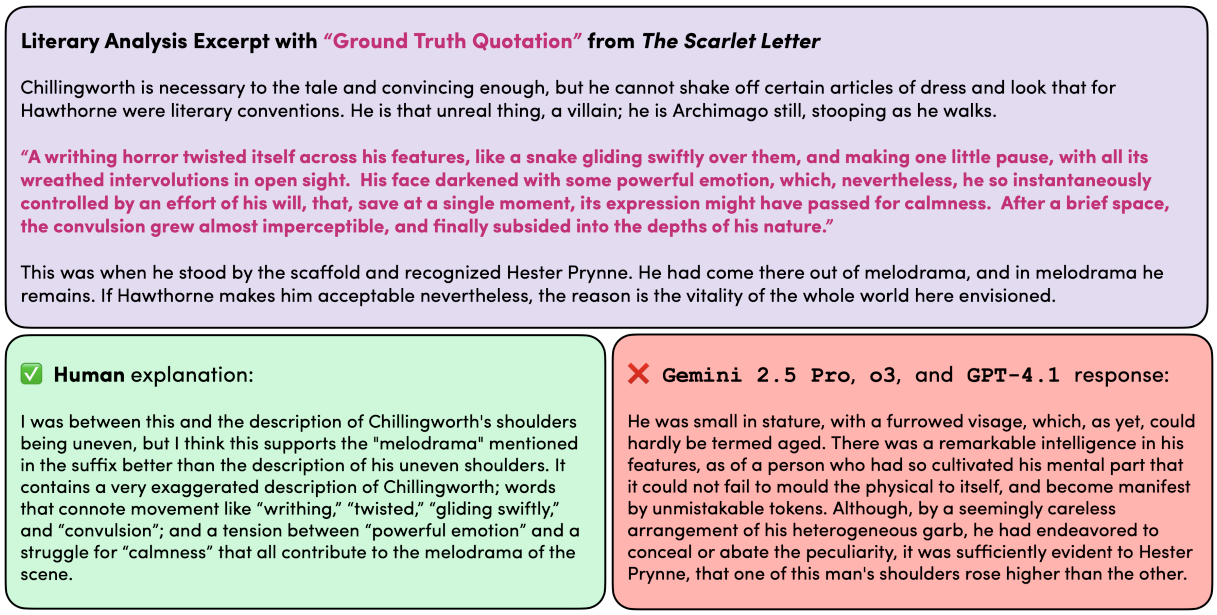

Figure 1: An example from long context RELiC where the HUMAN annotator successfully retrieves the quotation while the top-performing LLMs fail.

selecting a quotation[3]. Our prompts are adapted from Karpinska et al. (2024) in the NOCHA long-context benchmark. Additionally, we implemented an embedding-based retrieval baseline using GTE-QWEN2-7B-INSTRUCT, the top-performing text embedding model on MTEB[4] at the time of writing.

3.1 Human Evaluation

The author who manually validated high quality examples also attempted 40 examples from four previously-read primary source novels. The author had access to a digital copy of the primary source text with a string-matach search function and the corresponding Wikipedia summaries. In addition to selecting missing quotations, the author also wrote a short justification for each choice. The task took approximately 8 hours.

3.2 Evaluation Scheme

Our automatic evaluation of model responses applied partial ratio fuzzy matching (to account for minor typographical differences) to check that the model response was from the primary source text and to measure overlap between the response and the ground truth quotation[5]. For the embedding baseline, we calculate recall@1.

[3]Prompts & inference details in Appendix C.

[4]https://huggingface.co/spaces/mteb/leaderboard accessed February 2025

[5]Further details can be found in Appendix D.

4 Results & Analysis

Table 1 reports results for all evaluated models on the full set of 292 claims and the two folds.

LLMs outperform HUMAN and SOTA embedding baseline Google's GEMINI PRO 2.5 reasoning model achieves better accuracy than our HUMAN baseline in a fraction of the time—each API call took an average of 45 seconds vs. an average of 12 minutes for the HUMAN expert. Both Google and OpenAI's latest long-context models show substantial performance gains over their immediate predecessors on the literary evidence retrieval task. GEMINI PRO 2.5 outperforms GEMINI PRO 1.5 by a wide margin (64.7% vs. 38.5%), and OpenAI's GPT-4.1 improves significantly over GPT-4O (51.0% vs. 24.3%). Similarly, O3 surpasses O1 (48.7% vs. 32.2%). The embedding-based method achieved 4.5% accuracy, only 1.6% higher than the best recall@1 reported in the original RELiC paper from three years ago. LLMs' success over embeddings suggests the importance of contextualized representations that can enable reasoning over entire primary source texts.

Closed-source outperform open-weight LLMs Despite recent progress in open-weight LLM development, we observe a striking performance gap between closed-source and open-weight models on the literary evidence retrieval task. The best-performing open-weight model, DEEPSEEK-R1,

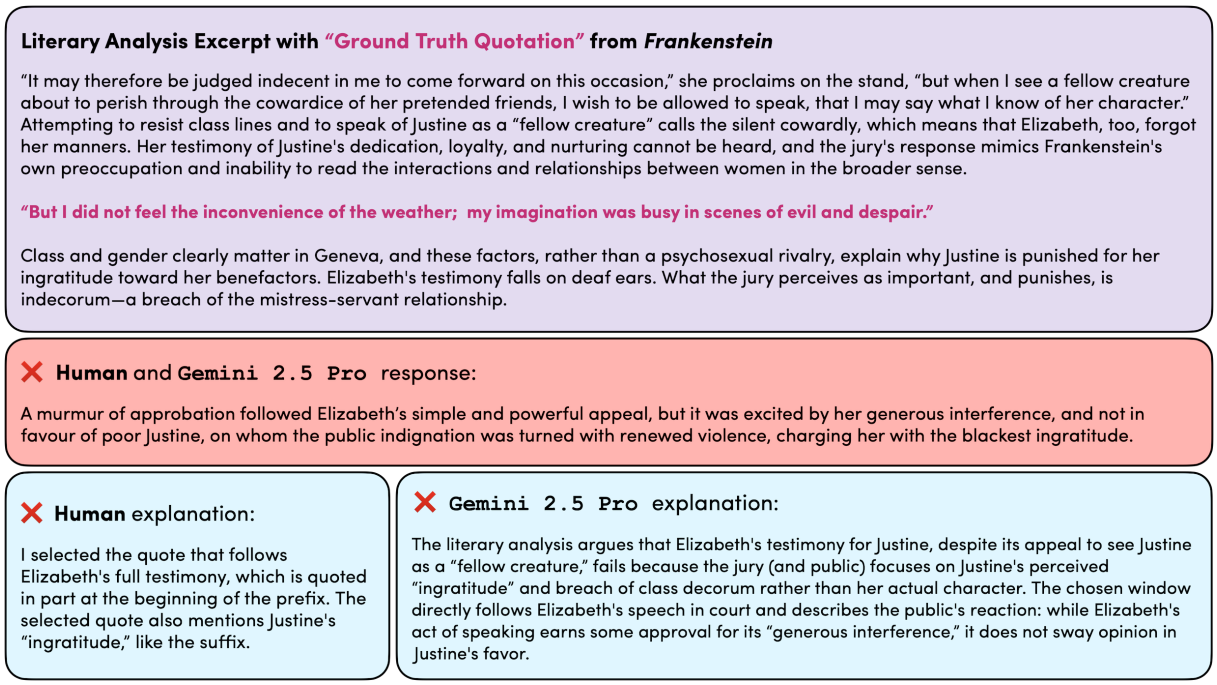

Figure 2: An example of long context RELiC where the HUMAN annotator and GEMINI PRO 2.5 both select the same quotation, but it is not the ground truth quotation.

achieves only 29.1% accuracy—less than half the performance of the top closed-source model, GEMINI PRO 2.5 (64.7%).

Small LLMs can't capitalize on close reading examples Nearly all models larger than 8B parameters show modest to significant improvements on the close reading fold, with some models improving by over 20%. This effect is expected due to ground truth leakage in the the provided excerpt of literary analysis, but the 7B and 8B parameter open-weight models experience nearly no performance boost on this fold—the performance of LLAMA 3.1 INSTRUCT (8B) actually suffers, suggesting that the smaller models lack the capacity to even exploit lexical overlap.

EXPLANATION prompt provides a glimpse into model reasoning We observed mixed results when querying non-reasoning LLMs with the EXPLANATION prompt. While GEMINI PRO 1.5 improved with explanation-based reasoning, GPT-4O saw a decline. Recent reasoning LLMs such as GEMINI PRO 2.5 and O3 use internal reasoning tokens that are not exposed via the API but influence the model's final output. These tokens are critical to the model's performance on complex tasks, but their inaccessibility complicates model evaluation and comparison. Applying the EXPLANATION prompt to these reasoning models preserved accuracy, despite the redundancy, while providing insight into the models' reasoning processes. See a comparison of human and model justifications in Figure 2.

All models tend to overgenerate Despite prompts explicitly instructing models to select and generate no more than five consecutive sentences from the primary source, we observe a consistent tendency across all evaluated models to produce significantly longer outputs. To quantify this overgeneration behavior, we compute the length ratio, defined as the ratio of the model's generated output length to the ground truth quotation length (measured in characters); values closer to 1.0 indicate greater adherence to the prompt constraints and minimal overgeneration. The length ratios are reported in Table 3.

We find that all models overgenerate, including the human annotator (average ratio of 2.1), likely due to natural variance in interpreting sentence boundaries. However, language models consistently surpass this baseline: for instance, state-of-the-art models such as GEMINI PRO 2.5 and GPT-4.1 have average ratios of 3.0 and 4.8, respectively. Smaller open models like LLAMA 3.1 INSTRUCT (8B) and LLAMA 3.3 INSTRUCT (70B) exhibit even more extreme overgeneration (ratios > 5.7), suggesting that weaker models may compensate for uncertainty by producing longer outputs.

LLMs struggle with literary nuance In Figure 1, we present a challenging example where all LLMs fail but our HUMAN expert correctly identified the ground truth quotation from *The Scarlet Letter*, a >80k-word novel. The context alludes to a description of the character Roger Chillingworth at a specific moment in the story (when he recognizes Hester Prynne). In their explanation, the HUMAN expert mentions considering two different passages, but ultimately selecting the correct passage because it best demonstrates the "melodrama" mentioned in the literary analysis. The expert highlights the features of the correct passage that demonstrate "melodrama": an exaggerated character description, words that connote movement, and emotional tension. All three top-performing LLMs selected the other passage the HUMAN expert was considering but ultimately eliminated, suggesting that even the best models still lack the expertise to navigate literary devices and more nuanced signals that inform expert literary interpretation.

Models can identify alternative literary evidence In Figure 2, we present an example where both HUMAN and GEMINI PRO 2.5 selected the same "incorrect" quote. The explanations reveal that both interpreted the literary analysis context as an introduction to a quote about the jury's response. However, the ground truth quotation actually supports Frankenstein's own preoccupation rather than the trial's immediate outcome. This example highlights two key insights: (1) literary evidence retrieval is inherently interpretative, meaning that multiple passages may plausibly support a given claim, and (2) LLMs, like human readers, can surface alternative yet reasonable quotations that align with certain aspects of the analysis. While models may not always select the canonical answer, their ability to propose viable alternative evidence could be valuable for assisting literary scholars in exploring multiple textual connections.

5 Related Work

Our work builds on recent papers that apply and evaluate LLMs for computational literary analysis. Prior work has explored summarization or claim verification in novels (Subbiah et al., 2024a; Kim et al., 2024; Karpinska et al., 2024) or short stories (Subbiah et al., 2024b). Other more specific tasks, mainly in the short context setting, include extracting narrative elements (Shen et al., 2024), story arcs and turning points (Tian et al., 2024), character analysis (Papoudakis et al., 2024), narrative discourse (Piper and Bagga, 2024), plot development (Huot et al., 2024; Xu et al., 2024), and creativity evaluation (Chakrabarty et al., 2024).

MODEL + SIMPLE PROMPT	Accuracy $_{(n=40)}$	Avg. Length Ratio
GEMINI PRO 2.5	57.5	3.5
O3	47.5	3.5
GEMINI PRO 1.5	22.5	2.8
O1	25.0	3.2
GPT-4O	17.5	3.9
QWEN 2.5 INSTRUCT (72B)	7.5	3.2
LLAMA 3.1 INSTRUCT (8B)	5.0	5.9
QWEN 2.5 INSTRUCT (7B)	5.0	2.8
LLAMA 3.3 INSTRUCT (70B)	0.0	5.7
MODEL + EXPLANATION PROMPT		
GEMINI PRO 2.5	**62.5**	3.0
GPT-4.1	47.5	4.8
O3	50.0	2.7
GEMINI PRO 1.5	40.0	3.3
CLAUDE SONNET 3.7	32.5	4.0
DEEPSEEK-R1	15.0	3.6
GPT-4O	22.5	3.6
QWEN 3 (32B)	20.0	2.7
QWEN 3 (8B)	5.0	2.4
O3-MINI	10.0	3.5
BASELINES		
HUMAN	55.0	**2.1**

Table 3: The average length ratios for each model, defined as the ratio of the length model generation to that of the ground truth (measured in characters).

6 Conclusion

In this work, we introduced a long-context dataset derived from RELiC for evaluating literary evidence retrieval using large language models. We constructed a high-quality test set of 292 examples, ensuring rigorous evaluation through filtering and human verification. Our results demonstrate that long-context LLMs outperform embedding-based retrieval methods and even a HUMAN expert baseline. However, closer analysis reveals that model success is not always grounded in robust interpretative abilities: models tend to overgenerate and still struggle to grasp nuances in literature. However, they can also successfully identify additional evidence for literary claims. Finally, open-weight models still lag far behind their closed-weight counterparts, suggesting that long-context capabilities alone are insufficient without strong interpretive reasoning. We release our data and evaluation framework to spur research at the intersection of NLP and literary analysis.

Limitations

As noted in the original paper, the RELiC dataset represents a limited, English-only subset of world literature, with a strong emphasis on the Western

literary canon. To foster a more inclusive and representative benchmark, we aim to extend this work to cross-lingual literary analysis and retrieval on texts from historically underrepresented literary traditions. Expanding the dataset in this way would not only enhance the linguistic and cultural diversity of literary evidence retrieval but also improve the generalizability of models across different literary frameworks.

Additionally, our human evaluation was limited in scope and scale, as it was conducted by a single annotator, one of the authors of this paper. While this approach was practical given the intensive nature of literary evidence retrieval over full novels, it inherently reflects the knowledge, biases, and interpretive lens of a single individual. In the future, we aim to broaden our evaluation by incorporating multiple expert annotators, enabling a more comprehensive and diverse assessment of model performance on the task of complex literary reasoning.

Ethical Considerations

While LLMs can assist in literary interpretation, they should not replace scholarly analysis conducted by trained experts. We acknowledge the potential for misuse of a system that can retrieve evidence for scholarly claims and emphasize that automated retrieval should be viewed as a complement to, not a substitute for, human literary scholarship.

As mentioned in the Limitations section, the primary source texts in our dataset are drawn from public domain works, which primarily reflect Western literary traditions. The results of this benchmark may not generalize to literary traditions outside the Anglophone canon, potentially reinforcing existing biases in computational literary studies.

References

Anthropic. 2025. Claude 3.7 sonnet.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google. 2025. Gemini 2.5 pro. Available at `https://deepmind.google/models/gemini/pro/`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. In *Conference on Language Modeling*.

Ane Ohrvik. 2024. What is close reading? an exploration of a methodology. *Rethinking History*, 28(2):238–260.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. O1 model card. Accessed: 2024-02-16.

OpenAI. 2025a. Gpt-4.1.

OpenAI. 2025b. o3.

OpenAI. 2025c. Openai o3-mini system card. Accessed: 2025-02-15.

Argyrios Papoudakis, Mirella Lapata, and Frank Keller. 2024. BookWorm: A dataset for character description and analysis. In *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pages 4471–4500, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. HEART-felt narratives: Tracing empathy and narrative style in personal stories with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1026–1046, Miami, Florida, USA. Association for Computational Linguistics.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024a. STORYSUMM: Evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005, Miami, Florida, USA. Association for Computational Linguistics.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024b. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELiC: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024. Fine-grained modeling of narrative context: A coherence perspective via retrospective questions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5822–5838, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset Details

We are using a publicly available dataset released by researchers who stated in their paper that they released their data to "facilitate further research in this direction." Statistics for the primary sources can be found in 4.

B LLM-Aided Data Processing

B.1 Filtering steps

These cleaning and filtering steps were applied to the entire RELiC dataset in the following order:

1. `CLEAN` (GPT-4O-MINI): We asked the model to remove any OCR artifacts and to ensure the context and ground truth quotation flow seamless and grammatically without changing any meaning. Additionally, we asked the model to remove any remaining in-line citations that revealed the page number of the ground truth quotation, which allowed us to preserve some examples that might have been caught by the `LOCATION` filter below. See the prompt in Table 8.

2. `LEAKAGE` (heuristic): If the sentences in the context preceding or following the ground truth quotation were fuzzy-matches (threshold: 95) with any text from the primary source, the example was excluded from our dataset.

3. `LIT ANALYSIS` (GPT-4O-MINI): We asked the model to classify whether each RELiC instance was an example of literary analysis. If it was not, we excluded it from our dataset. See the prompt in Table 9.

4. `LOCATION` (GPT-4O-MINI): We asked the model if the context revealed the location of the ground truth quotation. If it did, we excluded it from our dataset. See the prompt in Table 10.

5. `FIRST SENT` (heuristic): We identified cases where the ground truth quotation was the first sentence of its primary source novel. The intuition behind this filter (and the following two) was that the first sentence of a primary source might be famous or more likely to be quoted, and therefore easier to guess without needing to reason over the text or more likely to appear in training data.

Book Title	Author	Publication Year	Token Count	Word Count
Brave New World	Aldous Huxley	1932	91,472	65,278
What Maisie Knew	Henry James	1897	124,544	95,988
Ethan Frome	Edith Wharton	1911	45,038	34,926
Frankenstein	Mary Shelley	1818	95,018	75,131
The Great Gatsby	F. Scott Fitzgerald	1925	63,683	48,972
The Awakening	Kate Chopin	1899	66,574	49,932
The Scarlet Letter	Nathaniel Hawthorne	1850	112,355	83,311

Table 4: Books included in the dataset. The token count was provided as per `tiktoken` tokenization.

6. `LAST SENT` (heuristic): Similarly, we identified cases where the ground truth quotation was the last sentence of its primary source novel.

7. `OUTLIER` (heuristic): We identified cases where the ground truth quotation was cited much more frequently than any of the other quotations from the same primary source novel.

8. `EZ2MEM` (GPT-4): We asked the model to perform the RELiC task *without* providing the primary source text and identified cases where the model was able to correctly generate at least one sentence of the ground truth quotation.

Note that we did not automatically exclude any of the examples identified by the last four filters. Though we did not use them in this project, we have left the labels in our dataset to facilitate future research.

B.2 Manual validation of data

We conducted multiple validations of the RELiC data for quality control using human annotations.

Validation of LLM filters To validate our LLM-aided approach to data preprocessing, we manually annotated 100 examples of RELiC and compared our judgments to keep or reject each example to the results of the LLM + heuristic filters. The f1-score of our filtering scheme was 89.8, with our scheme identifying 57 true positives, 30 true negatives, 6 false positives, and 7 false negatives.

C Model Inference Details

API access details

- All OpenAI models were accessed via the OpenAI API (`https://platform.openai.com/`).
- All Gemini models were accessed via the Google AI API (`https://ai.google.dev/`).
- CLAUDE SONNET 3.7 was accessed via the Vertex AI/Google Cloud (`https://cloud.google.com/vertex-ai?hl=en`).
- DeepSeek was run via OpenRouter (`https://openrouter.ai/`).

The approximate total cost of running all closed-weight models for both development and evaluation was $2k.

Local inference details All Llama and Qwen models were run locally. The smaller open-weight models were run on 1 A100 GPU for a total of 14 hours, while the larger open-weight models were run on 4 A100 GPUs for a total of 16 hours.

Generation hyperparameters The OpenAI reasoning models (O1, O3-MINI, and O3) were a special case: the temperature parameter is fixed and the token generation limit includes the inaccessible reasoning tokens. As such, we set the token generation limit to 12,000 for O1 and O3-MINI. For O3, the token generation limit was set to 25,000 as recommended by OpenAI[6]. For all three OpenAI reasoning models, we used the default `medium` reasoning effort.

For all other models, the token generation limit was set to 800 for SIMPLE prompts and 1,200 for EXPLANATION prompts. Temperature was set to 0.0 for all models except for QWEN 3 (32B) and QWEN 3 (8B), where temperature was set to the default 0.6.

For a list of the models and their checkpoints, see Table 5.

D Evaluation Scheme Details

We use the `rapidfuzz` Python package for our fuzzy match evaluation with a threshold of 95 for

[6] `https://platform.openai.com/docs/guides/reasoning#allocating-space-for-reasoning`

Model	Context	Avail.	Checkpoints
o1	200k	🔒	o1-2024-12-17
o3	200k	🔒	o3-2025-04-16
o3-mini	200k	🔒	o3-mini-2025-01-31
GPT-4o	128k	🔒	gpt-4o-2024-11-20
GPT-4.1	1M	🔒	gpt-4.1-2025-01-31
Gemini Pro 1.5	1M	🔒	gemini-1.5-pro-002
Gemini Pro 2.5	1M	🔒	gemini-2.5-pro-preview-05-06
Claude Sonnet 3.7	200k	🔒	claude-3-7-sonnet@20250219
Llama 3.3 Instruct (70B)	128k	🔓	Llama-3.3-70B-Instruct
Llama 3.1 Instruct (8B)	128k	🔓	Llama-3.1-8B-Instruct
Qwen 2.5 Instruct (72B)	128k	🔓	Qwen2.5-72B-Instruct
Qwen 2.5 Instruct (7B)	128k	🔓	Qwen2.5-7B-Instruct
Qwen 3 (8B)	131k*	🔓	Qwen3-8B
Qwen 3 (32B)	131k*	🔓	Qwen3-32B
gte-Qwen2-7B-instruct (7B)	32k	🔓	gte-Qwen2-7B-instruct

Table 5: The upper rows display the evaluated LLMs, while the bottom row displays the text embedding model used for the baseline. Context lengths marked with * were extended with YaRN.

checking the existence of the model response in the primary source and a threshold of 90 for checking the overlap between the model response and the ground truth quotation. The fuzzy matching allows for small typographical differences between the primary source and the model outputs (e.g. different types of quotation marks). Thresholds were determined after manual inspection of outputs at varying thresholds.

E Use of AI Assistants

The authors used Github Copilot for coding assistance during their experiments and ChatGPT for assistance in formatting LaTeX in the writing of this paper.

Prompt (Simple)

You are provided with the full text of **[book_title]** and an excerpt of literary analysis that directly cites **[book_title]** with the cited quotation represented as <MASK>.

Your task is to carefully read the text of **[book_title]** and the excerpt of literary analysis, then select a window from **[book_title]** that most appropriately replaces <MASK>as the cited quotation by providing textual evidence for any claims in the literary analysis.

The excerpt of literary analysis should form a valid argument when <MASK>is replaced by the window from **[book_title]**.

<full_text_of_[book_title_snake_case]>**[book_sentences]** </full_text_of_[book_title_snake_case]>

<literary_analysis_excerpt>**[lit_analysis_excerpt]** </literary_analysis_excerpt>

Identify the window that best supports the claims being made in the excerpt of literary analysis. The window should contain no more than 5 consecutive sentences from **[book_title]**.

Provide your final answer in the following format:

<window>**YOUR SELECTED WINDOW** </window>

Table 6: Prompt template for literary evidence retrieval (Simple).

Prompt w/ Explanations

You are provided with the full text of **[book_title]** and an excerpt of literary analysis that directly cites **[book_title]** with the cited quotation represented as <MASK>.

Your task is to carefully read the text of **[book_title]** and the excerpt of literary analysis, then select a window from **[book_title]** that most appropriately replaces <MASK>as the cited quotation by providing textual evidence for any claims in the literary analysis.

The excerpt of literary analysis should form a valid argument when <MASK>is replaced by the window from **[book_title]**.

<full_text_of_[book_title_snake_case]>**[book_sentences]** </full_text_of_[book_title_snake_case]>

<literary_analysis_excerpt>**[lit_analysis_excerpt]** </literary_analysis_excerpt>

First, provide an explanation of your decision marking process in no more than one paragraph.

Then, identify the window that best supports the claims being made in the excerpt of literary analysis. The window should contain no more than 5 consecutive sentences from **[book_title]**.

Provide your final answer in the following format:

<explanation>**YOUR EXPLANATION** </explanation>

<window>**YOUR SELECTED WINDOW** </window>

Table 7: Prompt template for literary evidence retrieval with explanation.

Prompt for Cleaning RELiC Examples

I want to create a dataset for Natural Language Processing that consists of excerpts of literary analysis that quote directly from a primary source text.

I have already collected windows of literary analysis that quote from many different primary sources in the public domain. The part of the excerpt before the quotation is called the "prefix," and the part of the excerpt after the quotation is called the "suffix." Because I collected the data from PDFs with OCR, the prefixes and suffixes contain artifacts like in-line citations, page numbers, chapter names, headers, and footers that I need to remove.

Here is an example:

<prefix>**[prefix]** </prefix>

<ground_truth_quotation>**[ground_truth_quotation]** </ground_truth_quotation>

<suffix>**[suffix]** </suffix>

Please follow the following guidelines to help me clean and filter this example:

1. Remove all textual artifacts from the OCR process by deleting them. Artifacts include page numbers, chapter headings, footnotes, image captions, etc.

2. Correct the grammar, punctuation, and spelling of the prefix and suffix without altering the meaning so that the primary source quotation fits seamlessly and grammatically between the prefix and suffix.

3. Remove all in-line citations following quotations, especially those that say things like "(my emphasis)."

Respond with ONLY the cleaned prefix and suffix in the following format:

<clean_prefix>**CLEAN PREFIX** </clean_prefix>

<clean_suffix>**CLEAN SUFFIX** </clean_suffix>

Table 8: Prompt template for cleaning RELiC examples

Prompt for Classifying Literary Analysis

I want to create a dataset for Natural Language Processing that consists of excerpts of literary analysis that quote directly from a primary source text.

Other texts can contain quotes from primary sources, like biographies of authors. However, I only want examples of literary analysis in the dataset.

Please determine whether the following excerpt is an example of literary analysis. If it is literary analysis, respond with TRUE. Otherwise, respond with FALSE.

Here is the excerpt:

<excerpt>**[clean_prefix] [answer_quote] [clean_suffix]** </excerpt>

Respond with ONLY your answer in the following format:

<answer>**YOUR ANSWER** </answer>

Table 9: Prompt template for classifying literary analysis

Prompt for Identifying Quote Location Disclosure

I want to create a dataset for a Natural Language Processing task called "Literary Evidence Retrieval."

The input has two parts:

(1) An excerpt of literary analysis that quotes 1 to 5 full sentences from a primary source text, but the quote is replaced with "[MASK]".

(2) The full text of the primary source.

The output should be the correct quotation of 1 to 5 full sentences from the primary source—this is the ground truth.

I have already collected windows of literary analysis that quote from many different primary sources in the public domain.

The part of the excerpt before the quotation is called the "prefix," and the part after is called the "suffix."

To avoid giving the model hints, it's extremely important that the prefix and suffix do not mention the chapter where the ground truth quotation is located.

Please determine whether the following instance contains the chapter location of the ground truth quotation in either the prefix or the suffix. If it does, respond with TRUE; otherwise, respond with FALSE.

The prefix and suffix may contain the location of other quotations in the passage. I only want to identify if the location of the ground truth quotation is revealed.

Look out for phrases like "In the prologue" or "In Chapter..."

Here is the instance:

<prefix>**[prefix]** </prefix>

<ground_truth_quotation>**[ground_truth_quotation]** </ground_truth_quotation>

<suffix>**[suffix]** </suffix>

Respond with ONLY your answer in the following format:

<answer>**YOUR ANSWER** </answer>

Table 10: Prompt template for identifying quote location disclosure