



Literary Evidence Retrieval via Long-Context Language Models

Anonymous ACL submission

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 work (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. To ensure high-quality evaluation, we curate a high-quality subset of 252 examples through extensive filtering and human verification. Our experiments show that large long-context models, such as GEMINI PRO 1.5 and GPT-4O, retrieve literary evidence more effectively than state-of-the-art retrievers, yet still trail behind human experts (40% vs. 52.5% accuracy). Moreover, smaller open-weight models achieve only 2–5% accuracy, showing that they lack the reasoning abilities essential for literary interpretation. We release our dataset to facilitate future applications of LLMs to literary analysis.¹

1 Introduction

The emergence of long-context language models, which can process millions of input tokens ([Gemini Team, 2024](#)), has unlocked new AI applications for literary analysis. In this paper, we focus specifically on the task of *literary evidence retrieval*, in which a model must retrieve evidence in the form of quotations from a primary source (e.g., a novel) to support claims made by literary scholars.

[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 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, and asked to generate the missing quote (see [Table 3](#), top). This task, an integral part of the human process of literary criticism, requires complex reasoning over the entire narrative (e.g., understanding plot, subtext, and other literary devices) to make connections between the claim and the literary text.

We curate a high-quality subset of RELiC that contains 252 high-quality examples verified by human annotation, with claims that require both global reasoning over events as well as “close reading” over singular passages to solve. Our experiments reveal that while long-context language models fare better than state-of-the-art retrievers at retrieving evidence, they significantly underperform a human expert baseline: our best model, GEMINI PRO 1.5, obtains an accuracy of **40%** compared to **52.5%** for a human expert on a subset of the data. We observe a wide gap between large closed models (GEMINI PRO 1.5 GPT-4O) and smaller open-weight models like QWEN 2.5 INSTRUCT (7B) and LLAMA 3.1 INSTRUCT (8B) in fact, the latter models obtain accuracies of 2-5%. A qualitative analysis reveals that GEMINI PRO 1.5 and GPT-4O still struggle with nuance, which form 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 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

¹Data and code to be released after blind review.

MODEL + SIMPLE PROMPT	ALL (n=252)	👤 (n=40)	📖 (n=30)
GEMINI PRO 1.5 (Gemini Team, 2024)	36.1	22.5	40.9
o1 (OpenAI, 2024)	33.3	25.0	40.9
GPT-4o (OpenAI et al., 2024)	28.2	17.5	29.5
QWEN 2.5 INSTRUCT (72B) (Qwen et al., 2025)	11.9	7.5	18.2
LLAMA 3.1 INSTRUCT (8B) (Grattafiori et al., 2024)	4.8	5.0	4.5
LLAMA 3.3 INSTRUCT (70B)	2.4	0.0	2.2
QWEN 2.5 INSTRUCT (7B)	2.0	5.0	2.2
MODEL + EXPLANATION PROMPT			
GEMINI PRO 1.5	38.5	40.0	50.0
GPT-4o	24.3	22.5	31.8
o3-MINI (OpenAI, 2025)	8.3	10.0	13.6
BASELINES			
GTE-QWEN2-7B-INSTRUCT	4.8	2.5	6.8
HUMAN	-	52.5	-

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=9)		Dataset Examples (n=252)		
	TOKENS	WORDS	TOKENS	WORDS	# Ex/BOOK
MEAN	88,468	66,903	254.9	203.6	36.0
ST. DEV.	25,527	19280	66.9	52.9	19.6
MAX	124,544	95,988	492.0	385.0	52
MIN	45,038	34,926	116.0	91.0	7.0

Table 2: Summary statistics for long-context RELiC.

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 pipeline² 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 several of the primary source novels manually review 400 filtered examples. This author marked 252 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.

Test set folds: The test set contains two labeled folds of special data, (1) the 👤 HUMAN EVAL SET: 40 examples attempted by our author, and (2) the 📖 CLOSE READING SET: 30 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 selecting a quotation³. 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 MTEB⁴ at the time of writing.

³Prompts & inference details in Appendix C.

⁴<https://huggingface.co/spaces/mteb/leaderboard>

Literary Analysis Excerpt with Ground Truth Quotation from <i>The Great Gatsby</i>	
One of the more memorable scenes in <i>The Great Gatsby</i> , which Nick calls a "picture," is a distant cousin of the facetious art photograph, "George B. Wilson at the Gasoline Pump." It is a still picture of a mundane subject that reveals, with great impact, a critical insight and the emotional jolt at the discovery of bitter knowledge:	
Daisy and Tom were sitting opposite each other at the kitchen table with a plate of cold fried chicken between them and two bottles of ale. He was talking intently across the table at her and in his earnestness his hand had fallen upon and covered her own. Once in a while she looked up at him and nodded in agreement. They weren't happy, and neither of them had touched the chicken or the ale—and yet they weren't unhappy either. There was an unmistakable air of natural intimacy about the picture and anybody would have said that they were conspiring together.	
The passage could be considered "Jamesian," comparable to corresponding tableaux in <i>The Ambassadors</i> and <i>The Portrait of a Lady</i> . The picture seals Gatsby's fate and Nick's too; his doubts and hopes are over now. "Jamesian" surely, but just as surely photographic.	
<p>✓ Human explanation: The part in the suffix about "sealing Gatsby's fate" made me think this quote was towards the end and something damning for Gatsby. The "picture" described by Nick is Nick's observation that Daisy and Tom are actually fine after all of the drama of the car accident and Gatsby and Daisy revealing their relationship to Tom.</p> <p>✗ Gemini Pro 1.5 response: In one of the windows over the garage the curtains had been moved aside a little and Myrtle Wilson was peering down at the car. So engrossed was she that she had no consciousness of being observed and one emotion after another crept into her face like objects into a slowly developing picture. Her expression was curiously familiar—it was an expression I had often seen on women's faces but on Myrtle Wilson's face it seemed purposeless and inexplicable until I realized that her eyes, wide with jealous terror, were fixed not on Tom, but on Jordan Baker, whom she took to be his wife.</p> <p>✗ o3-mini response: GEORGE B. WILSON. Cars Bought and Sold—and I followed Tom inside. The interior was unprosperous and bare; the only car visible was the dust-covered wreck of a Ford which crouched in a dim corner. It had occurred to me that this shadow of a garage must be a blind and that sumptuous and romantic apartments were concealed overhead when the proprietor himself appeared in the door of an office, wiping his hands on a piece of waste. He was a blonde, spiritless man, anaemic, and faintly handsome. When he saw us a damp gleam of hope sprang into his light blue eyes.</p>	

Table 3: An example of long context RELiC where the HUMAN annotator successfully retrieves the quotation while the LLMs fail.

3.1 Human Evaluation

The author who manually validated high quality examples attempted 40 unseen test set examples from four previously-read primary source novels. The author also wrote a short justification for each choice and took approximately 8 hours in total.

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 measure overlap between the response and the ground truth quotation⁵. For the embedding baseline, we calculate recall@1.

4 Results & Analysis

Table 1 reports results for all evaluated models on the full set of 252 claims and the two folds.

Closed-source outperform open-weight LLMs

On average, the closed-source models perform much better than the open-weight models, with the best closed-source model, GEMINI PRO 1.5 outperforming the best open-weight model, QWEN 2.5 INSTRUCT (72B), 38.5% to 11.9%.

Close reading examples are surprisingly challenging Closed models show improved accuracy on the CLOSE READING subset and the embedding baseline, likely due to ground truth leakage. However, among open models, only QWEN 2.5 IN-

STRUCT (72B) benefits, highlighting weaknesses in long-context reasoning.

LLMs outperform SOTA embedding baseline

The embedding-based method achieved 4.8% accuracy, only 1.9% higher than the best recall@1 reported in the original RELiC paper nearly three years ago. LLMs' success over embeddings suggests the importance of contextualized representations, as embeddings lack the ability to process full primary source texts.

Mixed results with EXPLANATION prompt

While GEMINI PRO 1.5 improved with explanation-based reasoning, GPT-4o saw a decline. Notably, o3-MINI performed poorly, possibly due to requiring justification after generating the reasoning tokens that are inaccessible via the OpenAI API.

LLMs trail human accuracy

Even the best long-context LLM falls short of human performance: GEMINI PRO 1.5 (40%) vs. HUMAN (52.5%). In 15% of cases, the HUMAN expert succeeded where all models failed, while in no cases did more than two models succeed where the expert failed.⁶ However, the HUMAN expert required 12 minutes per example, whereas GEMINI PRO 1.5 completed inference in under 30 seconds.

LLMs struggle with deep literary reasoning

In Table 3, we present a challenging example where the HUMAN correctly identified the ground truth quotation from *The Great Gatsby*, a >45k-word novel,

⁶Only 2 cases (5%) saw two models succeed where the HUMAN failed, and 3 cases (7.5%) where one model succeeded.

⁵Further details can be found in Appendix D.

Literary Analysis Excerpt with Ground Truth Quotation from <i>Frankenstein</i>	
<p>"It may therefore be judged indecent in me to come forward on this occasion," she proclaims on the stand, "but when I see a fellow creature about to perish through the cowardice of her pretended friends, I wish to be allowed to speak, that I may say what I know of her character." Attempting to resist class lines and to speak of Justine as a "fellow creature" calls the silent cowardly, which means that Elizabeth, too, forgot her manners. Her testimony of Justine's dedication, loyalty, and nurturing cannot be heard, and the jury's response mimics Frankenstein's own preoccupation and inability to read the interactions and relationships between women in the broader sense.</p> <p>But I did not feel the inconvenience of the weather; my imagination was busy in scenes of evil and despair.</p> <p>Class and gender clearly matter in Geneva, and these factors, rather than a psychosexual rivalry, explain why Justine is punished for her ingratitude toward her benefactors. Elizabeth's testimony falls on deaf ears. What the jury perceives as important, and punishes, is indecorum—a breach of the mistress-servant relationship.</p>	
<p>✗ Human and Gemini Pro 1.5 response: A murmur of approbation followed Elizabeth's simple and powerful appeal, but it was excited by her generous interference, and not in favour of poor Justine, on whom the public indignation was turned with renewed violence, charging her with the blackest ingratitude.</p> <p>Human explanation: I selected the quote that follows Elizabeth's full testimony, which is quoted in part at the beginning of the prefix. The selected quote also mentions Justine's "ingratitude," like the suffix.</p> <p>Gemini Pro 1.5 explanation: The literary analysis focuses on Elizabeth's testimony during Justine's trial and the jury's reaction to it. The analysis highlights Elizabeth's attempt to transcend class boundaries by referring to Justine as a "fellow creature" and argues that this, along with Elizabeth's emotional testimony, is perceived as a breach of decorum and contributes to Justine's conviction. Therefore, the selected window should include Elizabeth's testimony where she calls Justine a "fellow creature" and speaks about her character, as well as the jury's immediate negative reaction, emphasizing their focus on the perceived impropriety of the situation rather than the content of the testimony.</p>	

Table 4: An example of long context RELiC where the HUMAN annotator and GEMINI PRO 1.5 both select the same quotation, but it is not the ground truth quotation.

but all LLMs fail. The analysis alludes to a casual remark in Chapter 2, but the ground truth quotation is from the end of Chapter 7. While o3-MINI was misled by the named entity in the analysis, GEMINI PRO 1.5 misinterprets the metaphorical use of "picture" in the analysis by selecting a scene in which a character's face is likened to a developing picture rather than its intended symbolic meaning: a static scene with deep implications. These errors suggest models rely on surface-level lexical cues rather than thematic significance.

Models can identify alternative literary evidence

In Table 4, we present an example where both HUMAN and GEMINI PRO 1.5 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).

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 252 examples, ensuring rigorous evaluation through filtering and human verification. Our results demonstrate that long-context LLMs outperform embedding-based retrieval methods, yet they still struggle with complex literary reasoning compared to human performance on the same task. Open-weight models perform especially poorly, suggesting that long-context capabilities alone are insufficient without strong interpretive reasoning. A fine-grained analysis revealed that LLMs struggle to grasp nuances in literary analysis, but they can also successfully identify additional evidence for literary claims. 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

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.

- Gemini Team. 2024. [Gemini 1.5: Unlocking multi-modal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.
- 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 Madry, 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. 2025. [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 5.

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 9.
2. LEAKAGE (heuristic): If the sentences in the context preceding or following the ground truth quotation were fuzzy-matches (threshold:) 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 10.
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 11.
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.
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

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 Picture of Dorian Gray	Oscar Wilde	1890	105,737	79,130
To the Lighthouse	Virginia Woolf	1927	91,799	69,463
The Awakening	Kate Chopin	1899	66,574	49,932
The Scarlet Letter	Nathaniel Hawthorne	1850	112,355	83,311

Table 5: Books included in the dataset. The token count was provided as per `tiktoken` tokenization.

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.

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

All close-source models were run using their APIs: OpenAI (<https://platform.openai.com/>) and Google AI (<https://ai.google.dev/>) at a total cost of \$1.6k. All open-weight 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.

The OpenAI reasoning models (o1 and o3-mini) 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 and used

MODEL	CONTEXT	AVAIL.	CHECKPOINTS	COST (\$)
o1	200k	🔒	o1-2024-12-17	\$
o3-mini	200k	🔒	o3-mini-2025-01-31	\$
GPT-4o	128k	🔒	gpt-4o-2024-11-20	\$
GEMINI Pro 1.5	1M	🔒	gemini-1.5-pro-002	\$
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	\$
GTE-QWEN2-7B-INSTRUCT (7B)	32k	🔒	gte-Qwen2-7B-instruct	\$

Table 6: The upper rows display the evaluated LLMs, while the bottom row displays the text embedding model used for the baseline. In the last column, we report the total spent in USD for development and evaluation of each model

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 a list of the models and their checkpoints, see Table 6.

D Evaluation Scheme Details

We use the `rapidfuzz` Python package for our fuzzy match evaluation with a threshold of 96 for checking the existence of the model response in the primary source and a threshold of 97 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)
<p>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>.</p> <p>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.</p> <p>The excerpt of literary analysis should form a valid argument when <MASK>is replaced by the window from [book_title].</p> <p><full_text_of_[book_title_snake_case]>[book_sentences] </full_text_of_[book_title_snake_case]></p> <p><literary_analysis_excerpt>[lit_analysis_excerpt] </literary_analysis_excerpt></p> <p>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].</p> <p>Provide your final answer in the following format:</p> <p><window>YOUR SELECTED WINDOW </window></p>

Table 7: Prompt template for literary evidence retrieval (Simple).

Prompt w/ Explanations
<p>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>.</p> <p>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.</p> <p>The excerpt of literary analysis should form a valid argument when <MASK>is replaced by the window from [book_title].</p> <p><full_text_of_[book_title_snake_case]>[book_sentences] </full_text_of_[book_title_snake_case]></p> <p><literary_analysis_excerpt>[lit_analysis_excerpt] </literary_analysis_excerpt></p> <p>First, provide an explanation of your decision marking process in no more than one paragraph.</p> <p>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].</p> <p>Provide your final answer in the following format:</p> <p><explanation>YOUR EXPLANATION </explanation></p> <p><window>YOUR SELECTED WINDOW </window></p>

Table 8: 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 9: 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 10: Prompt template for classifying literary analysis

Prompt for Identifying Quote Location Disclosure
<p>I want to create a dataset for a Natural Language Processing task called "Literary Evidence Retrieval."</p> <p>The input has two parts:</p> <ol style="list-style-type: none"> (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. <p>The output should be the correct quotation of 1 to 5 full sentences from the primary source—this is the ground truth.</p> <p>I have already collected windows of literary analysis that quote from many different primary sources in the public domain.</p> <p>The part of the excerpt before the quotation is called the "prefix," and the part after is called the "suffix."</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>Look out for phrases like "In the prologue" or "In Chapter..."</p> <p>Here is the instance:</p> <pre><prefix>[prefix] </prefix> <ground_truth_quotation>[ground_truth_quotation] </ground_truth_quotation> <suffix>[suffix] </suffix></pre> <p>Respond with ONLY your answer in the following format:</p> <pre><answer>YOUR ANSWER </answer></pre>

Table 11: Prompt template for identifying quote location disclosure