One Thousand and One Pairs: A "novel" challenge for long-context language models

https://novelchallenge.github.io/

Marzena Karpinska[☾] **Katherine Thai**[☾] **Kyle Lo**[✎] **Tanya Goyal**[☆] **Mohit Iyyer**[☾]
[☾]UMass Amherst [✎]Allen Institute for AI [☆]Princeton University
{mkarpinska,kbthai,miyyer}@umass.edu, kylel@allenai.org, tanyagoyal@princeton.edu

Abstract

Synthetic long-context LLM benchmarks (e.g., "needle-in-the-haystack") test only surface-level retrieval capabilities, but how well can long-context LLMs retrieve, synthesize, and reason over information across book-length inputs? We address this question by creating NoCha, a dataset of *1,001 minimally different* pairs of true and false claims about *67* recently-published English fictional books, written by human readers of those books. In contrast to existing long-context benchmarks, our annotators confirm that the largest share of pairs in NoCha require *global reasoning over the entire book* to verify. Our experiments show that while human readers easily perform this task, it is enormously challenging for all ten long-context LLMs that we evaluate: no open-weight model performs above random chance (despite their strong performance on synthetic benchmarks), while GPT-4o achieves the highest accuracy at 55.8%. Further analysis reveals that (1) on average, models perform much better on pairs that require only sentence-level retrieval vs. global reasoning; (2) model-generated explanations for their decisions are often inaccurate even for correctly-labeled claims; and (3) models perform substantially worse on speculative fiction books that contain extensive world-building. The methodology proposed in NoCha allows for the evolution of the benchmark dataset and the easy analysis of future models.

1 Introduction

The context size of large language models has increased by multiple orders of magnitude over the last year: for instance, Google's Gemini Pro 1.5 can process millions of input tokens at once. But can models *truly* utilize and reason over their claimed context? Existing long-context evaluation methods such as finding the "needle in the haystack" (NIAH) (Kamradt, 2023) measure surface-level retrieval capabilities, but do not necessarily assess performance on the more challenging task of synthesizing distant and underlying information as we show in §4.

We bridge this gap by introducing NoCha (A **No**vel **Cha**llenge), in which LLMs are prompted to verify claims written about recently-published fiction books. Claim verification has been extensively studied in the context of shorter documents (Thorne et al., 2018; Wadden et al., 2020; Fabbri et al., 2022a), but its application to book-length fictional texts presents unique challenges. Firstly, the task necessitates reasoning over both explicit information directly stated in the text and implicit information inferred from the narrative, which is often distributed throughout the entire document. Secondly, the use of recent, fictional context prevents the model from relying solely on parametric knowledge, necessitating the comprehension and interpretation of the long text.

Our data collection process aims to balance efficiency and quality. Rather than pre-selecting a set of books, we follow the approach of Kim et al. (2024) and ask annotators to self-report recently published novels they have read. The annotators then create true/false **narrative minimal pairs** that isolate a single narrative phenomenon present in their novels. Each false claim differs from the true claim in its pair *only* by the inclusion of false information regarding the same event or entity (see Figure 1). This approach offers two key advantages: (1) it minimizes the chances that the model is "correct for the wrong reason," as it must accurately predict *both* labels in the pair, and (2) it simplifies the process of quality control, as the false claim can be easily verified against its true counterpart, making it easier to identify claims that are either too similar or overly subjective. NoCha contains 1,001 narrative minimal pairs for 67 books, created at a total cost of $3,330 USD.

Accurately labeling the minimal pairs in NoCha often requires not only information retrieval from

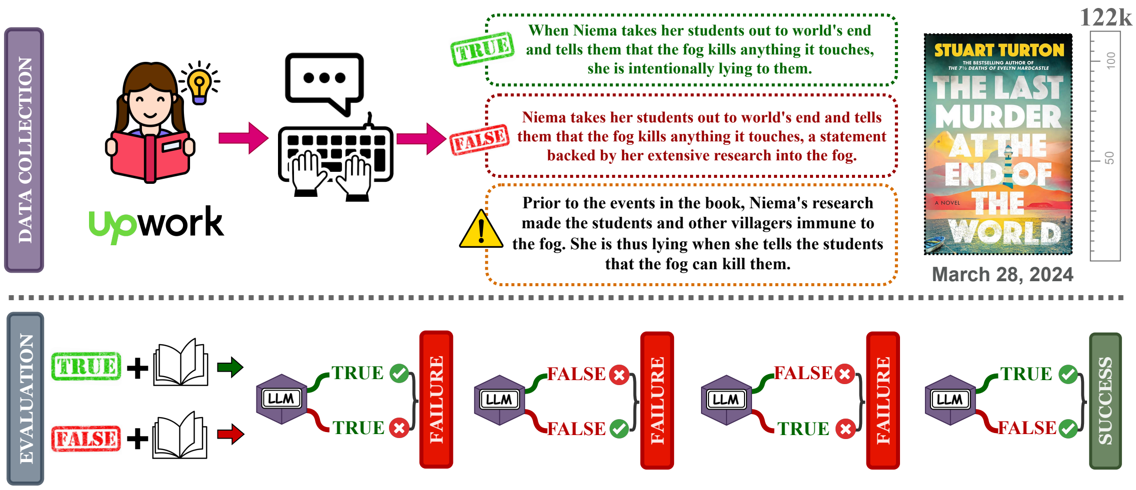


Figure 1: Overview of NoCha's data collection and evaluation pipeline. Readers create true/false claim pairs for books published between 2023 and 2024 with written justifications for the false labels. Each model is given the full book as context and evaluates one claim at a time. We measure *pair accuracy*, where the model must identify the true claim as true and the false claim as false to receive credit. This approach helps guard against label bias while also better assessing the true understanding of the text, as both claims pertain to the same events or parts of the story. Our books range from 49k to 336k tokens, and each model is tested on a subset of books that fit within its context.

the context, but also synthesis, reasoning, and inference over information from multiple parts of the document. Experiments on 5 openly-available and 6 closed-source models show that: (1) though all models struggle on NoCha, no open-weight model performs above random chance, while GPT-4o is the best-performing model overall at 55.8% accuracy; (2) on average, all models perform much better on pairs that require sentence-level retrieval than global reasoning (59.8% vs 41.6%), though their performance on these pairs in NoCha is much lower (59.8%) than on NIAH (100%) as reported in Hsieh et al. (2024); (3) model-generated explanations for their decisions are often inaccurate, even for correctly labeled claims; and (4) models perform substantially worse on books with extensive world-building, as these require more sophisticated reasoning.

To summarize, our contributions are threefold:

1. **Data:** We introduce NoCha, a dataset of 1,001 narrative minimal pairs about *recently* published fictional books, designed to evaluate the long-context language models' ability to process and reason about long narratives.

2. **Analysis:** We use NoCha to conduct a comprehensive analysis of 5 openly available and 6 closed-source models, identifying where models struggle, thereby providing valuable insights improving long-context models.[1]

3. **Methodology:** We contribute our data collection and evaluation methodology, which balances efficiency and quality, minimizing the possibility of models receiving credit for correct predictions made without full utilization of the context.

2 Data & Methods

In this section, we first motivate NoCha's design—narrative minimal pairs written by readers of books—and then describe our data collection process and evaluation methodology.

Issues with existing long-context benchmarks: A popular method to evaluate long-context models is the "needle-in-a-haystack" task (Kamradt, 2023, NIAH), which involves injecting a sentence-level piece of information at different depths of a document. NIAH was recently extended by RULER (Hsieh et al., 2024) to include various types of needles and other synthetic tasks. While this approach allows control over the exact position of the evidence, it has several limitations: (1) evidence scope is heavily restricted by the length and complexity of the needle, (2) the needle is unrelated to the continuous context, mak-

[1]In this paper, we define long-context models as those with a claimed context window of at least 128k tokens.

ing it easier to retrieve,[2] and (3) the task is synthetic, making it a poor proxy for real-world tasks. Recently, more realistic long-context benchmarks have emerged. The most similar to NoCha are L-Eval (An et al., 2024), NovelQA (Wang et al., 2024a), and ∞*bench* (Zhang et al., 2024), which are all literary QA tasks. However, all three use novels that are publicly available online, meaning they are likely part of LLMs' pretraining data along with the numerous books, articles, and summaries written about them.[3] They also lack the complex global coverage of many of our claims; for instance, while 35% of NovelQA reportedly consists of multi-hop questions, these involve simpler tasks such as aggregation (e.g., *"How many times has Bran jumped off and ran?"*) that do not require reasoning over *implicit* information in the text.

2.1 Data collection

To address these limitations, we adopt a design that includes (1) human-written examples for claim verification, an important real-world task; (2) recently-published texts to mitigate data contamination issues (Jacovi et al., 2023); (3) fictional texts to prevent over-reliance on parametric knowledge vs. in-context information; and (4) minimal pairs to guard against models being "right for the wrong reasons" (see Figure 2 for example) while also enabling easy verification to ensure dataset quality.

Collecting a corpus of recently-published fiction: We collect the 67 books in NoCha following the approach of Kim et al. (2024), who created the FABLES dataset to evaluate summarization faithfulness with similar objectives in mind (reducing data contamination and reliance solely on parametric knowledge). Concretely, we select books that are: (1) fictional narratives, (2) published in 2023 or 2024,[4] and (3) self-reported as having been read by our annotators. Our resulting dataset comprises 63 new books (33 published in 2023 and 30 in 2024) and four classic novels (see Table 12 for full list of the books). The mean length of books in our dataset is **127k** tokens and **98.5k** words (see Table 1 for statistics).[5]

	Books (*n*=67)		**Claim**$_{\{Pairs\}}$ (*n*=2002)$_{\{1001\}}$		
	Tokens	Words	Tokens	Words	# Claim/Book
Mean	127,324	98,587	23.22	18.26	14.94
St. Dev.	52,561	39,506	7.62	6.49	8.37
Max	336,288	257,445	63	57	46
Min	49,156	38,023	5	4	4

Table 1: Summary statistics for NoCha.

Annotators: We recruited 18 annotators via Upwork and 5 volunteer annotators (convenience sample) who reported having read books published in 2023 or 2024.[6] All annotators were required to read the guidelines and sign a consent form before starting the task. They were compensated \$3–\$5 USD per pair of claims, based on the agreed rate and the number of annotations, resulting in an hourly rate of approximately \$18–\$30 USD.[7] Overall, we collected 1001 claim pairs (approximately 10-15 per book) at a cost of \$2.8k USD. The protocol was reviewed and deemed *Not Human Subjects Research* by the Institutional Review Board. See §B.2 for details on the annotators and the recruitment process.

Collecting true/false pairs: To collect true/false pairs, annotators were instructed to first write a true statement about the events or characters in the book, and then create a corresponding false statement addressing the same aspect of the book such that the verification of one claim as "true" necessitates the verification of the other as "false". We trained annotators by having them write a true counterpart to a false model-generated claim.[8] Both types of pairs—those written from scratch and those where only one claim was created by the annotator—were included in the final dataset after quality control.[9]

[2]The context in Hsieh et al. (2024) may match the needle's topic, but it is not a continuous, logically connected text.

[3]While ∞*bench* swaps core entities in the texts, we found that prompting GPT-4-Turbo with excerpts from the "falsified" novels results in the model recognizing the original novel and identifying the changed entities.

[4]We will not release the full NoCha dataset because (1) the books are copyrighted, and (2) we want to prevent model providers from training on this data and compromising the benchmark. Instead, we additionally annotate a smaller subset of *classic books* that are out of copyright. This subset will serve as a publicly available sample of the data. The authors commit to updating NoCha with new books and evaluating new long-context models.

[5]All token counts provided in this paper are based on `tiktoken` (`https://github.com/openai/tiktoken`) with the `cl100k` encoding and word counts are determined by splitting the text on `whitespaces`, unless stated otherwise.

[6]The volunteer group included three of the authors, each of whom read three books of their choice and performed the annotations to better understand the difficulty of the task and the time required to complete it.

[7]The annotators were usually able to create about 6-10 claim pairs per hour.

[8]Model-generated claims were selected from FABLES (Kim et al., 2024) as well as generated by Claude-3-Opus and GPT-4-Turbo.

[9]Around 30% of the pairs in the data contain one model generated claim. In some cases, these claims were edited in

Model	Context	Avail.	Checkpoints	# Param
GPT-4o	128k	🔒	gpt-4o-2024-05-13	¯_(ツ)_/¯
GPT-4-Turbo	128k	🔒	gpt-4-turbo-2024-04-09	¯_(ツ)_/¯
Claude-3-Opus	200k	🔒	claude-3-opus-20240229	¯_(ツ)_/¯
Claude-3.5-Sonnet	200k	🔒	claude-3-5-sonnet@20240620	¯_(ツ)_/¯
Gemini Pro 1.5	1M	🔒	gemini-1.5-pro-preview-0514	¯_(ツ)_/¯
Gemini Flash 1.5	1M	🔒	gemini-1.5-flash-preview-0514	¯_(ツ)_/¯
Command R	128k	🔓	c4ai-command-r-v01	35B
Command R+	128k	🔓	c4ai-command-r-plus	104B
Gemma-10M	10M	🔓	gemma-2b-10m	2B
Phi-3-mini	128k	🔓	Phi-3-mini-128k-instruct	3.8B
LongLLaMA	256k	🔓	long_llama_3b_instruct	3B

Table 2: Evaluated models: the upper row displays all closed-source models, while the lower row lists all open-weight models (see §C for details).

The annotators were also asked to write a short explanation as to why these claims were true or false based on the content of the book. We closely monitored this process, providing feedback and seeking clarifications from the annotators.

Quality control: To ensure the quality of the claims, we hired annotators who had read the same books to validate 128 claims from five books. The annotators were paid \$1.66–\$2.77 USD per claim based on their requested rate, with the total annotation costing \$285 USD. Overall, the annotators agreed on 124 out of 128 labels (Krippendorff's $\alpha = 0.953$) (see §B.2 for details on this process). This high agreement also allow us to conclude that **human readers are very strong performers on NoCha's claim verification task (96.9%).** Finally, after collecting all pairs, two of the authors reviewed all instances, resolving unclear cases with the annotators and between themselves if a claim appeared to be subjective or incorrect.

Evidence scope: To assess the reasoning required to validate a claim pair, we obtained additional annotations on 121 claim pairs from 8 books. These annotations determine whether verification is possible based on (1) one or two sentences (similar to NIAH), (2) a single contiguous passage, or (3) global reasoning over the full book. Annotators were compensated \$2 USD per pair, totaling \$242 USD for the annotations. Overall, our annotations suggest that 12.4% of pairs can be validated with reasoning over one or two sentences, 39.7% require reasoning over a longer contiguous passage, and **the largest fraction, 47.9%, necessitate global reasoning for correct verification.**

3 Experiments

We test the long context reasoning capabilities of 5 open-weight models and 6 closed-source mod-

order to decontextualize the claim.

Evaluation Template (Main)

You are provided with a context and a statement. Your task is to carefully read the context and then determine whether the statement is true or false.

Answer TRUE if the statement is true in its entirety based on the context provided.
Answer FALSE if any part of the statement is false based on the context provided.

<context>**book text**</context>

<statement>**claim**</statement>

<question>Based on the context provided, is the above statement TRUE or FALSE?</question>

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:
<explanation>YOUR EXPLANATION</explanation>
<answer>YOUR ANSWER</answer>

Table 3: Prompt template used for `Evaluation`. This prompt was employed to evaluate *all* models. All open source models were also evaluated using prompt in Table 18.

els (see Table 2) by prompting each model in a zero-shot manner to verify a single claim given the entire book content as evidence, similar to the summarization setup of Kim et al. (2024). After pilot experiments with different prompts (see §C for a full description), we observe that the best-performing prompt asks the model to first explain its reasoning before making a final decision, similar to chain-of-thought prompting (Wei et al., 2022).

Model and prompt details: Each model was prompted with a claim and the entire book using the prompt template in Table 3.[10] All 🔒 **closed-source models** were accessed via the provider's API[11] at a total cost of ∼\$8k USD. Since we found that 🔓 **open-weight models** struggle to follow the prompt's instructions, we also experiment a simplified prompt that just asked for a true or false decision (see Table 18).[12][13] Finally, to measure whether retrieval-augmented language models per-

[10]We set the temperature to 0 and generate up to 800 tokens.

[11]We encounter several issues while generating the text. Most notably, Gemini Pro 1.5 and Gemini Flash 1.5 refuse to process some of the books returning a `prohibited content` error (likely due to copyrights). Overall, both models refused to generate the label in about 48% of cases significantly reducing the number of pairs tested for these models.

[12]Outputs generated using the simplified prompt are denoted by the subscript *simple*.

[13]We encountered some generation issues with LongLLaMA which failed to produce outputs when prompted with the prompt in Table 3. Hence, we only report the results with the simplified prompt for this model.

MODEL	PAIR ACC(correct/total)	COMMON SET ACC
GPT-4O	**55.8** (344/617)	**58.2** (206/354)
GPT-4-TURBO	40.2 (248/617)	40.1 (142/354)
CLAUDE-3-OPUS	49.4 (463/937)	50.8 (180/354)
CLAUDE-3.5-SONNET	41.0 (384/937)	40.7 (144/354)
GEMINI PRO 1.5	48.1 (247/514)	48.3 (171/354)
GEMINI FLASH 1.5	34.2 (176/515)	35.0 (124/354)
COMMAND R	19.6 (87/445)	*n/a*
COMMAND R_{simple}	22.5 (100/445)	*n/a*
COMMAND R+	17.3 (77/445)	*n/a*
COMMAND $R+_{simple}$	13.7 (61/445)	*n/a*
PHI-3-MINI	9.3 (23/247)	*n/a*
PHI-3-$MINI_{simple}$	14.5 (48/331)	*n/a*
GEMMA-10M	3.9 (39/1001)	*n/a*
GEMMA-$10M_{simple}$	7.5 (75/1001)	*n/a*
$LONGLLAMA_{simple}$	4.9 (61/937)	*n/a*
BM25+GPT-4O (*k=5*)	28.2 (282/1001)	28.9 (102/353)
BM25+GPT-4O (*k=25*)	44.1 (441/1001)	46.5 (164/353)
BM25+GPT-4O (*k=50*)	49.7 (497/1001)	51.0 (180/353)
RANDOM	25.0 (250/1001)	25.0 (88/353)

Table 4: Percentage of claim pairs identified correctly by each model (see Table 3 for the prompt; and Table 19 for the prompt employed with BM25). "COMMON SET" refers to claim pairs shared among the models. The subscript "SIMPLE" refers to the calls done with simplified prompt (Table 18). Models performing below random are marked in red. This results include also four classic novels, which were most likely in the training data. For results excluding these novels see Table 23.

MODEL	RULER (%) VANILLA NIAH	RULER (%) NIAH SUITE	NOCHA (%)
GPT-4-TURBO	100.0	89.6	40.2
COMMAND R	98.0	84.8	19.6 / 22.5_{simple}

Table 5: Performance on NIAH does not translate to NOCHA accuracy for GPT-4-TURBO and COMMAND R. The table includes published results with 128k tokens from Hsieh et al. (2024) for both the standard NIAH (Table 11) and the NIAH suite, which averages results from 8 NIAH variants (Table 13). Additionally, we provide a comparison of performance on the entire RULER benchmark versus NOCHA for overlapping models in Table 6 for reference.

MODEL	RULER (%) (128k)	NOCHA (%) (~128k)
GEMINI PRO 1.5	94.4	48.1 / $48.3_{(\sim 128k)}$
GPT-4-TURBO	81.2	40.2
COMMAND R	76.0	19.6 / 22.5_{simple}
COMMAND R+	63.1	17.3 / 13.7_{simple}
PHI-3-MINI	43.3	9.3 / 14.5_{simple}

Table 6: Performance on RULER (Hsieh et al., 2024) compared to NOCHA accuracy for overlapping models. The table includes results for the 128k tokens group from `https://github.com/hsiehjackson/RULER`.

form better or worse than the long-context setting, we also experiment with a **BM25+GPT-4o** configuration, in which BM25 (Robertson et al., 1995) is used to rank the most relevant $k \in \{5, 25, 50\}$ excerpts from the book.[14] We then prompt GPT-4O (our best performing model) with the retrieved excerpts as evidence using the prompt in Table 19.[15]

Evaluation: We report the overall PAIRWISE ACCURACY for each model. Models get credit only if they label *both* the true and false claim in a pair correctly; and no points otherwise. We report the number of correctly-verified pairs divided by the total number of pairs that the model labeled. See §C for details.

4 Results & analysis

Table 4 reports results for individual models as well as for a common set of claims shared among closed-source models.[17] Overall, GPT-4O exhibits the highest accuracy at 55.8% (58.2% for the common set), though it still misidentifies at least one claim in the pair 44.2% of the time (see Table 8 for example failure cases). The second-best performing model is CLAUDE-3-OPUS with an overall accuracy of 49.4%, followed by GEMINI PRO 1.5 with an accuracy of 48.1%. CLAUDE-3.5-SONNET, GPT-4-TURBO and GEMINI FLASH 1.5 perform worse, with accuracies of 41.0%, 40.2% and 34.2%, respectively.[18] All open-weight models perform below random (25%), ranging from 22.5% for COMMAND R (with a simplified prompt) to as low as 3.9% for GEMMA-10M. Because of their poor accuracies, we exclude open-weights models from the following analysis; see §D for details.

Good performance on "needle in the haystack" does not imply high NOCHA accuracy: In Table 5, we compare the performance of COMMAND R and GPT-4-TURBO on NOCHA with the results reported in Hsieh et al. (2024, RULER), a more complex variant of NIAH, to estimate how performance on RULER translates to our task. Despite both models achieving high performance on RULER (84.8% and 89.6%, respectively, for the longest tested context of 128k), they underperform on our

[14] Excerpts were an average of 285 words based on `whitespace` splitting with all paragraph breaks preserved.

[15] 50 excerpts equals at most 39% of the total text of any book in our dataset; additional statistics on the total percentage of a book retrieved using this approach are available in Figure 13. The total cost of this experiment was $330 USD.

[17] Open-weights models are excluded as this would significantly reduce the number of shared pairs to just over 100.

[18] See Figure 9 for example of claim on which all models failed to generate the accurate label.

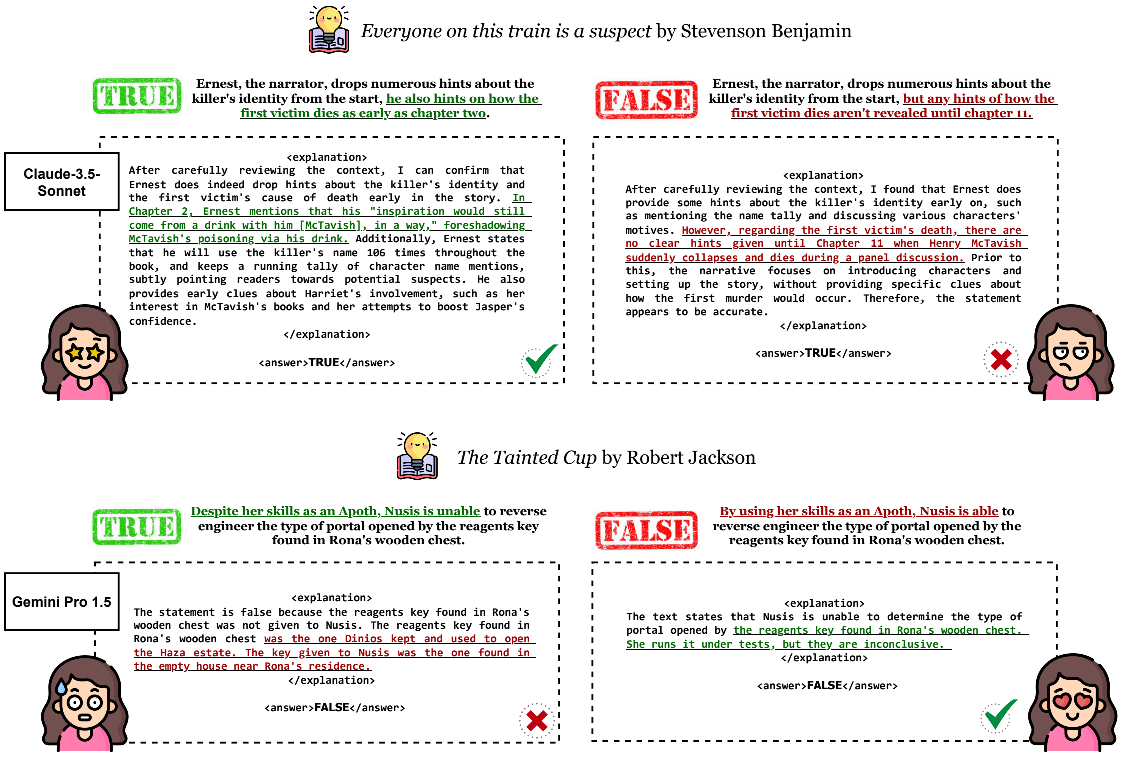


Figure 2: Examples of claim pairs where the models failed to validate one of the claims in the pair. Employing narrative minimal pairs helps us avoid awarding the model for cases where it only *appears* to produce the correct answer, when in fact the prediction was made without fully or efficiently utilizing the context. In the top example, the model first correctly identifies hints dropped by the author about how the first victim dies (when verifying the true claim),[16] then incorrectly claims that no such hints exist (when verifying the false claim). In the bottom example, the model first incorrectly claims that the key given to Nuna was not found in Rona's wooden chest (when verifying the true claim), then correctly raises no objections to the fact that the key was found in Rona's wooden chest (when verifying the false claim).

SCOPE	CORRECT	INCORRECT	ACCURACY (%)
GLOBAL	104	146	41.6
PASSAGE	101	111	47.6
SENTENCE	49	33	59.8

Table 7: Overall accuracy on claim pairs of 🔒 closed-source models on a subset of data annotated for evidence scope (see §2 for the annotation process). Pairs requiring global reasoning to verify are most difficult for models. Figure 11 shows results by model.

task, with GPT-4-TURBO achieving only 40.2% accuracy and COMMAND R performing *below* random at 19.6%.[19] This shows that synthetic retrieval-focused benchmarks are insufficient to evaluate global long-document reasoning.

[19]Note that RULER reports the results for `gpt-4-1106-preview` checkpoints for GPT-4-TURBO, while the current work uses `gpt-4-turbo-2024-04-09` checkpoints, presumably a more advanced version of that model.

NoCHA claims that require global reasoning are particularly difficult to verify: Table 7 contains further analysis of model accuracy on a subset of NoCHA annotated for the scope of evidence (see §2.1 for details). Overall, models perform worst for claim pairs requiring global reasoning (41.6%), followed by reasoning over a longer passage (47.6%), and, finally, sentence-level evidence (59.8%). While performance on sentence-level evidence is higher than in the other two setups, it is still much lower than the "needle-in-a-haystack" performance reported in Hsieh et al. (2024). This indicates that NoCHA claim pairs with sentence-level evidence are still much harder to solve than NIAH, possibly due to NIAH's out-of-context evidence injection.

Claims about speculative fiction are harder to verify: Are texts with extensive world-building more difficult for models to process than those that take place in a fictional but realistic world? We cat-

CLAIM & GOLD LABEL	MODEL'S LABEL & JUSTIFICATION	HUMAN EXPLANATION
Wyatt apologizes to Ernest because he believes McTavish left him a negative review. *Everyone on the Train is a Suspect* by Benjamin Stevenson –True	*Wyatt's apology to Ernest is mentioned in the context, but it is not because he believes McTavish left him a negative review.* ***Instead, Wyatt's apology is vague and not directly linked to the review.*** *Ernest himself is confused about the reason for the apology, and it is later revealed that Simone, Ernest's agent, was behind the negative review as part of a coded message to McTavish. Therefore, the statement that Wyatt apologizes to Ernest because he believes McTavish left him a negative review is incorrect.* –GPT-4O: False	When Wyatt apologizes, Ernest does not initially understand the reason behind it. However, it is *heavily* implied shortly after that Wyatt, McTavish's publisher, was apologizing for the bad review, which at that time everyone, including Wyatt thought was written by McTavish himself.
Emily Sparrow was pregnant with John's child and disappeared in 1966. *How to Solve Your Own Murder* by Kristen Perrin –True	*The text states that Emily Sparrow was pregnant, but* ***it is never explicitly stated who the father of the child is****. While it is implied that the father could be John, it is also implied that it could be Rutherford Gravesdown (Ford).* –GEMINI PRO 1.5: False	Throughout the book, the reader is taken back to 1966, where it is revealed that Emily was pregnant, and the father's identity was uncertain—either Lord Gravesdown or John. The baby, Laura, was later adopted and had a daughter, Annie. In the latter part of the book, set in the present day, Annie meets John, who confesses that he is her grandfather, thus confirming that he was the father of Emily's baby.
None of the writers on the train went to the Gorge excursion as they had to stay to participate in a festival event. *Everyone on the Train is a Suspect* by Benjamin Stevenson –False	*The statement is true based on the context. The passage states that while the non-festival tourists went on a scheduled day trip sailing down Katherine Gorge, "In front of our carriages were forty or so black fold-out chairs set up in the red dirt for the festival attendees."* ***This indicates that the writers stayed behind for a festival event*** *while the other passengers went on the Gorge excursion.* –CLAUDE-3-OPUS: True	While all writers invited to the festival stayed for the event, Juliet, who was also an accomplished writer and, in fact, we later learn she was originally invited as well, went on the trip.
Lost Bride Manor is haunted by various spirits, including the benevolent ghost of Clover, Sonya's grandmother and Collin's mother, who communicates through music. *Inheritance* by Nora Roberts –True	*The statement is mostly true, but contains one key inaccuracy. The context does indicate that Lost Bride Manor is haunted by various spirits, including Clover, who communicates through music. Clover is described as Sonya's grandmother and Collin's mother, which is correct. However, the statement characterizes Clover as "benevolent", but the context does not provide enough information to definitively label her as such.* ***While Clover seems to be helpful at times, playing music and aiding Sonya, the text does not explicitly state that she is entirely benevolent.*** *Given this small but significant discrepancy, the statement cannot be considered entirely true.* –CLAUDE-3.5-SONNET: False	Clover is a ghost haunting the manor where Sonya moves in. From the first day, Clover, who we learn is the ghost of Sonya's grandmother, helps Sonya by playing music to cheer her up, adding to her courage, and keeping her company. Clover also tries to protect Sonya from the evil spirit that also haunts the manor. Although it is not directly stated in the text, Clover's kind and well-meaning nature towards Sonya is evident through her actions.
Ernest, the narrator, drops numerous hints about the killer's identity from the start, but any hints of how the first victim dies aren't revealed until chapter 11. *Everyone on the Train is a Suspect* by Benjamin Stevenson –False	*To determine the accuracy of the statement, I reviewed the context for any early hints about the killer's identity and the method of the first victim's death. The prologue and early chapters contain several clues about the characters' motives and potential involvement in the murder. Specifically,* ***the prologue includes a detailed description of a potential murder scene, mentioning "the bloodied handprint on the bathroom door" and "the crimson, dripping tip of the murder weapon."*** *This indicates that hints about the method of the first victim's death are indeed revealed before chapter 11.* –GPT-4O: False	It is true that Ernest, the narrator, drops some hints about how the first victim dies from the start. However, the passage cited by the model is irrelevant. It is taken from the Prologue, where Ernest, a mystery novel writer, talks about elements that typically appear in a mystery novel's prologue. He mentions a "bloodied handprint" and "the crimson, dripping tip of the murder weapon," which are common in mystery novels prologues. However, in this novel, the first victim was poisoned, and Ernest hints at this by saying that the inspiration for this book "came from a drink" with the victim.
Levi promises Grace that he will kill anyone who tries to hurt her, and Grace calls him to help her dispose of Norman's body. *Safe and Sound* by Laura McHugh –False	*In the context, Levi does promise Grace that he would kill anyone who tried to hurt her again. However, when Grace kills Norman,... she calls Tyson, Levi's brother. The passage states: "She... wanted to make it look like she couldn't have survived. So she could just go missing."...* ***It then says "Tyson grabbed me, clamped his hands around my neck, and didn't let go. 'I dumped a body for you.'"*** *This shows that it was Tyson, not Levi, who Grace called to help dispose of Norman's body.* –GPT-4O: False	The model's justification begins correctly, but the first passage cited is unrelated to the claim and clearly a lie told by Tyson in the context of the story. Additionally, the text "Tyson grabbed me, clamped his hands around my neck, and didn't let go." never appears in the book. Tyson does say "I dumped a body for you." to Grace—which proves the claim is false—but he does not choke her before saying so or at any other point in the book.
While on Sigrid's boat, Seventeen admits to Kat that he had the same underwater dream that she had. *Assassin Eighteen* by John Brownlow –True	*The statement is true based on the context provided. In Part 8,* ***after Kat is rescued from the water by Rachel and Vilmos in the Zodiac****, she describes her near-death experience underwater where she felt she and Seventeen were entwined as one. Seventeen then admits "I had the same dream", confirming he experienced the same vision while he was drowning.* –CLAUDE-3-OPUS: True	It is true that Seventeen admits he had the same dream in Part 8 of the book. However, Kat was not rescued by Rachel and Vilmos in the Zodiac during this scene. That happens much later in the book, towards the end of Part 11. This conversation happened when Kat and Seventeen were rescued by Sigrid.
Sonya befriends Trey's sister Anna, a potter, and his cousin Owen Poole, who becomes involved in solving the mystery of the lost brides. *Inheritance* by Nora Roberts –False	*The statement is mostly true, but contains one key inaccuracy. The context shows that Sonya does befriend Trey's sister Anna, who is a potter, and helps redesign her website. Sonya also meets and interacts with Owen Poole,* ***Trey's cousin, who becomes involved in investigating the haunted manor and the mystery of the lost brides. However, the statement is incorrect in saying that Owen "becomes involved in solving the mystery of the lost brides."*** *While Owen does help investigate strange occurrences at the manor and even enters the mirror with Sonya at the end, he is not described as actively trying to solve the mystery of the lost brides specifically. Sonya is the main character focused on that mystery. Therefore, while most of the statement is accurate, this one detail makes it not entirely true.* –CLAUDE-3.5-SONNET: False	It is true that Sonya befriends Trey's sister, Anna, and Owen. However, Owen is not Trey's cousin; he is Sonya's cousin, whom she did not know about. Owen and Trey then try to help Sonya solve the mystery of the lost brides.

Table 8: **Top half:** Examples of claims where the model generated incorrect labels, along with the model-generated label and justification, and human-written explanations for the correct label of each claim. **Bottom half:** Examples of claims correctly labeled by the model but with wrong justifications. The explanations were provided by the authors, based on examples from books they read. Tags have been removed from the generated content for clarity.

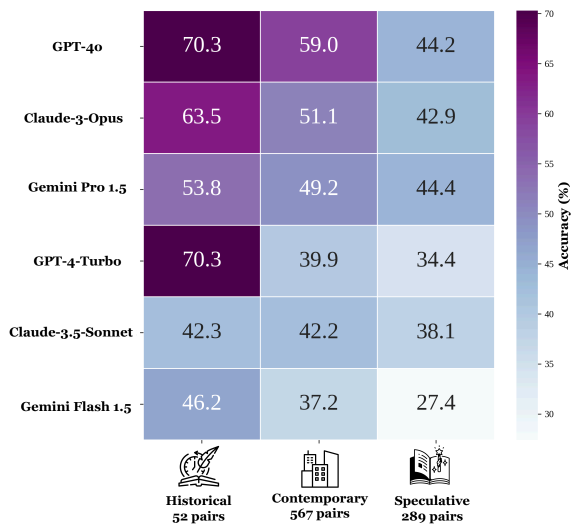


Figure 3: Performance of closed-source models on different types of novels. Note that two novels were excluded from this analysis as they could not clearly be classified in one of these categories. We provide the total number of claims in each category for reference however these numbers will vary slightly between the models due to the context-limitation and GEMINI PRO 1.5/GEMINI FLASH 1.5 refusals.

egorized 65 out of 67[20] of the books in our dataset into one of three broad categories:

- **Historical**: works set in our world before World War II, without any unrealistic elements.
- **Contemporary**: works set in our world after World War II, without any unrealistic elements.
- **Speculative**: works set in an alternate version of our world, containing both realistic and unrealistic elements, or in a completely invented universe.

The accuracy across all six closed-source models is 56.4% for historical fiction, 46.8% for contemporary fiction, and 38.8% for speculative fiction. Figure 3 shows that this pattern holds for each individual model, with accuracy being highest for historical fiction, followed by contemporary fiction, and lowest for speculative fiction. These results support the intuition that texts set in a realistic version of our world require less "work" from models to reason over than texts set in a universe that is

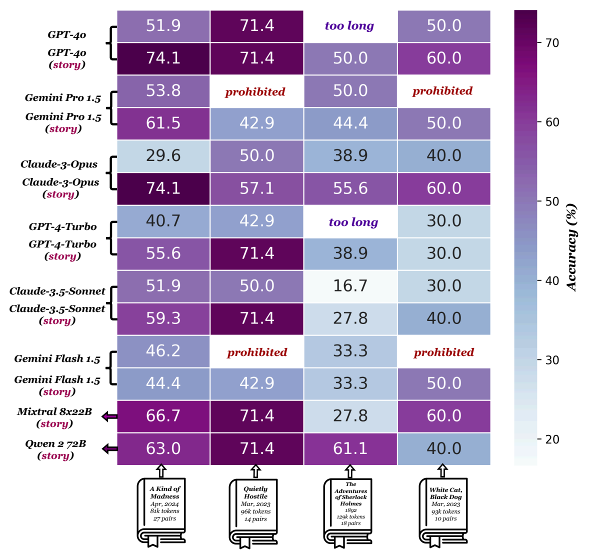


Figure 4: Model accuracy on claim pairs for different stories within collections. Accuracy is shown for (1) using the entire collection as context when prompting about a story, and (2) using only the individual story ("story") for the same set of claims. For GPT-4O and GPT-4-TURBO, one book was too long, so only the "story" context performance is presented. GEMINI PRO 1.5 and GEMINI FLASH 1.5 refused to process two books but handled the stories within, so only "story" context performance is available. We also provide performance of MIXTRAL-8X22B (65k) and QWEN-2-72B (32k)[22] on story-level input, for comparison.

defined within the text, perhaps because the models can rely more on their parametric knowledge.[21]

Irrelevant context confuses models: If we know that a claim can be verified by a short span of the book in isolation, does providing the rest of the book affect its accuracy? Figure 4 shows model performance on four *short story collections*, where the model is either given just the short story relevant to the claim (~21k tokens)[23] or the entire collection (~129k tokens). By prompting the model with the same claims, we partially control for confounding variables such as the inherent difficulty of the claim. While the Gemini models are relatively robust to added context, CLAUDE-3-OPUS's pair accuracy drops as much as 44.5% absolute when given the collection vs. the story, while CLAUDE-3.5-

[20]One book was a nonfiction collection of essays and another split the plot between modern day and the past, rendering it a combination of historical and contemporary fiction.

[21]We acknowledge that there may be confounding factors, such as the relative difficulty of writing challenging claims for different books types and the high average length of speculative fiction books (148k tokens)—though we note that all models perform better on historical (avg. length 133k tokens) than the shorter contemporary books (avg. length 115k tokens).

[23]Mean=8.5k, Min=702, Std.Dev.=4,355.

Model	Incorrect % (incorrect/total)
GPT-4o	21.7 (15/69)
GPT-4-Turbo	45.0 (27/60)
Claude-3-Opus	16.9 (13/77)
Gemini Pro 1.5	28.3 (13/46)
Gemini Flash 1.5	65.9 (27/41)

Table 9: The percentage of incorrect justifications provided for correct label by model which generated it in the analyzed subset of data.[24]

Sonnet, GPT-4-Turbo and GPT-4o exhibit smaller but still substantial drops. More generally, it is unclear whether claims about longer books (with potentially more irrelevant context) are harder to verify; see §D for more details.

Model-generated explanations are not always accurate: Do models correctly explain why a claim is true or false? Three of this paper's authors examined explanations generated for NoCha books that they read, covering a subset of 7 books (293 claims). Table 9 shows that *no* model consistently produces accurate explanations for every correctly-labeled claim, indicating a potential reliance on flawed or incomplete reasoning (see Table 8 for examples). Claude-3-Opus demonstrated the highest explanation accuracy, with 16.9% of its explanations containing errors, followed by GPT-4o and Gemini Pro 1.5 with 21.7% and 28.3% incorrect explanations respectively. GPT-4-Turbo and Gemini Flash 1.5 were the worst performing models with incorrect explanations reaching 45.0% and 65.9% respectively.[25] This is problematic when considering that users tend to rely on model explanations when verifying claims, even when the explanation is incorrect (Si et al., 2024). Further discussion on citations of evidence from the source found in model justifications can be found in §D.

[22]In fact, after the first victim dies, the narrator mentions: *"All I can tell you is what I've been telling you so far: the truth. After all, I told you Henry McTavish would be poisoned, didn't I? Well, not in those words, I suppose. But I did say the inspiration for this book would come from a drink with him."* It is also worth noting that the model incorrectly states that the first victim dies during the panel, when in fact McTavish dies the next day during breakfast.

[23]Here we employ the 32k, which can also be extended to 128k using YaRN (Peng et al., 2023).

[24]We do not analyze Claude-3.5-Sonnet's explanations because the model was run on June 22, after the completion of this analysis. Nonetheless, we have observed mistakes in this model's explanations as well, with an example provided in Table 8.

[25]As we run Claude-3.5-Sonnet on June 22, we exclude it from this analysis.

	BM25+GPT-4o		
Scope	*k=5* (%)	*k=25* (%)	*k=50* (%)
Global	25.9	29.3	41.4
Passage	22.9	43.8	45.8
Sentence	46.7	66.7	73.3

Table 10: Performance of the BM25+GPT-4o pipeline on the subset of data annotated for evidence scope.

Model	True (correct/total)	False (correct/total)
GPT-4o	77.5 (478/617)	75.9 (468/617)
GPT-4-Turbo	57.2 (353/617)	78.8 (486/617)
Claude-3-Opus	82.2 (770/937)	64.7 (606/937)
Claude-3.5-Sonnet	55.3 (518/937)	83.5 (782/937)
Gemini Pro 1.5	59.7 (307/514)	83.7 (431/515)
Gemini Flash 1.5	51.7 (266/515)	78.1 (402/515)
Command R	59.3 (264/445)	52.8 (235/445)
Command R$_{\text{simple}}$	74.4 (331/445)	44.3 (197/445)
Command R+	58.7 (261/445)	49.2 (219/445)
Command R+$_{\text{simple}}$	34.6 (154/445)	75.3 (335/445)
Phi-3-mini	34.8 (88/253)	46.5 (118/254)
Phi-3-mini$_{\text{simple}}$	69.5 (230/331)	32.0 (106/331)
Gemma-10M	20.2 (202/1001)	27.8 (278/1001)
Gemma-10M$_{\text{simple}}$	77.5 (776/1001)	18.7 (187/1001)
LongLLaMA$_{\text{simple}}$	63.8 (598/937)	19.7 (185/937)
BM25+GPT-4o (*k=5*)	33.7 (337/1001)	90.0 (901/1001)
BM25+GPT-4o (*k=25*)	57.5 (576/1001)	81.7 (818/1001)
BM25+GPT-4o (*k=50*)	65.7 (658/1001)	79.6 (797/1001)
Random	50.0 (500/1001)	50.0 (50/1001)

Table 11: Model accuracy on True and False claims for all data (see Table 3 for the prompt; and Table 19 for the prompt employed with BM25). "Common set" refers to claim pairs shared among the models. The subscript "simple" refers to the calls made with the simplified prompt (Table 18).

Can BM25 help prioritize important context? We observe that our BM25-assisted GPT-4o approach with $k = 50$ excerpts performs better than all models except for GPT-4o with the full book. For $k = 50$, an average of only 17% of a book was retrieved by BM25 and fed to GPT-4o in ranked order. Retrieval-based methods struggle on global reasoning due to out-of-order chunks, but they are effective for claims that require sentence- and, to a much lesser extent, passage-level retrieval to verify; thus, there is likely an upper bound to their NoCha accuracy (see Table 10).

Models have different predilections for predicting True vs False: Our pairs were designed so that validating one claim should enable validation of the other. However, we observe in Table 11 that some models tend to predict one label much more frequently than another. This tendency was particularly evident in Claude-3.5-Sonnet, Gemini

PRO 1.5, GEMINI FLASH 1.5, and GPT-4-TURBO, which had strong preferences for predicting False, and is in line with the observation reported for GEMINI PRO 1.5 in Levy et al. (2024). In contrast, CLAUDE-3-OPUS exhibited much higher accuracy on True labels (82.2%) compared to False (64.7%). GPT-4O was the only balanced model among the closed-source models, with accuracies of 77.5% for True and 75.9% for False.

5 Related Work

Evaluation of long-context models: Several benchmarks with inputs ranging from at least 8k up to 1M tokens have recently been introduced to evaluate long-context language models. Some of these benchmarks use synthetic tasks that can be generated programmatically or by LLMs, such as NIAH (Tay et al., 2021; Sun et al., 2021; Kamradt, 2023; Li et al., 2023; Hsieh et al., 2024; Levy et al., 2024; Liu et al., 2024; Lee et al., 2024). Others contain "realistic" tasks that typically require human annotators to devise, like traditional QA, summarization, and claim verification (Shaham et al., 2023; Bai et al., 2023; Dong et al., 2023) sometimes created by utilizing existing datasets (Hudson and Al Moubayed, 2022; Li et al., 2024). The existing long-context evaluation benchmarks with literary tasks all consist of publicly available English (Hudson and Al Moubayed, 2022; Zhang et al., 2024; Wang et al., 2024a; An et al., 2024) or Chinese language books (Yu et al., 2024). All but An et al. (2024) contain generative or multiple choice questions, including true/false questions, but unlike NOCHA, these are not minimal pairs.

Claim verification: Our work relates to prior work on claim verification, where claims are verified against a whole knowledge datastore (Thorne et al., 2018; Wadden et al., 2020; Schuster et al., 2021) or single evidence documents (Maynez et al., 2020; Falke et al., 2019). While earlier methods primarily relied on task-specific natural language inference (Laban et al., 2022; Honovich et al., 2022) or question answering models (Fabbri et al., 2022b; Wang et al., 2020) for claim verification, more recent work has explored using LLMs to evaluate the factuality of long-form model-generated text (Min et al., 2023; Wei et al., 2024; Manakul et al., 2023).

Minimal pairs: Minimal pairs, or contrast sets, are pairs of test set instances where a slight but impactful difference between the instances affects the gold label. Gardner et al. (2020) proposed contrast sets as a method for evaluating models on their intended tasks by manually perturbing test set instances. Our narrative minimal pairs were inspired by the DEMETR (Karpinska et al., 2022) dataset of minimals pairs for evaluating machine translation metrics and the BLiMP (Warstadt et al., 2020) benchmark of minimal pairs for evaluating English grammatical phenomena.

6 Conclusion and Discussion

We introduce NOCHA, a claim verification dataset designed to evaluate long context LLMs in a realistic task setting. Our design ensures that most claims necessitate global reasoning over extended contexts for verification and cannot be easily "gamed" by relying solely on parametric knowledge or generating correct answers through incorrect reasoning. Our experiments with 11 different long-context LLMs (5 open-weight, 6 closed-source) demonstrate that all models struggle on NOCHA. Furthermore, we reveal a substantial competency gap with human readers who can very easily perform this task.

Importantly, our results show that models that are "state-of-the-art" according to synthetic benchmarks like NIAH actually perform very poorly on our meticulously designed dataset. Nevertheless, we argue that synthetic datasets are useful and complementary to our realistic dataset; they allow for much higher flexibility such as easily adjusting for different context lengths or analyzing the lost-in-the-middle phenomenon. We encourage researchers to use a holistic approach and consider *both* synthetic and realistic tasks when evaluating long-context language models.

Limitations

The scope of our work is limited to novels published in English and the task of claim verification. It is unclear how the models' performance would translate to other languages, domains, or realistic tasks, and we leave that for future work.

We also acknowledge that this study and the methodology proposed are inherently and possibly prohibitively expensive due to the hiring of expert annotators and the thousands of LLM API calls. Because using the model developer's API is often the only way to access a closed-source model, the extent of our evaluation was limited to whether or not the closed-source models provided a response to our prompts. This resulted in a common set of only 354 pairs that all closed-source models labeled, making it difficult to truly compare the models to each other. Finally, while we commit to periodically updating the dataset and evaluating the models ourselves, we are unable to release the entire dataset due to copyright restrictions and to prevent model providers from training on it.

Ethical Considerations

All annotators consented to the use and publication of their annotations, which we will release for the portions of data that is *not* under copyright. Additionally, we ensured annotators received fair compensation for their contributions and respected their preferred rates. All copyrighted books were purchased using the funds that supported this work. We will *not* release the copyrighted portion of the data. Instead, we commit to consistently updating the dataset with newly published novels and evaluating the models ourselves.

Acknowledgements

We extend special gratitude to the Upwork annotators for their hard work, and to members from the UMass NLP lab for their feedback, particularly Chau Pham, Nader Akoury, Yekyung Kim, and Yapei Chang. We would also like to thank Byron Wallace for the name idea and Simeng Sun for the discussion that gave us the initial idea for this project. This project was partially supported by awards IIS-2202506 and IIS-2312949 from the National Science Foundation (NSF) as well as an award from Adobe.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mustafa Aljadery, Siddharth Sharma, and Aksh Garg. 2024. Gemma 2b - 10m context. Technical report. Accessed: 2024-05-23.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *ICLR*.

Anthropic. 2024a. Claude 3.5 Sonnet. Technical report, Anthropic. Accessed: 2024-06-23.

Anthropic. 2024b. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic. Accessed: 2024-05-23.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh.

2024. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. *Preprint*, arXiv:2407.07087.

Manuel Ciosici, Joe Cecil, Dong-Ho Lee, Alex Hedges, Marjorie Freedman, and Ralph Weischedel. 2021. Perhaps PTLMs should go to school – a task to assess open book and closed book QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6104–6111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cohere. 2024a. Command R: Retrieval-Augmented Generation at Production Scale. Technical report, Cohere. Accessed: 2024-05-23.

Cohere. 2024b. Introducing Command R+: A Scalable LLM Built for Business. Technical report, Cohere. Accessed: 2024-05-23.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv:2309.13345*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022a. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022b. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

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

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

George Hudson and Noura Al Moubayed. 2022. MuLD: The multitask long document benchmark. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3675–3685, Marseille, France. European Language Resources Association.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. *Github*.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. 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. *Preprint*, arXiv:2404.01261.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu.

2024. Can long-context language models subsume retrieval, rag, sql, and more? *ArXiv*, abs/2406.13121.

Russell V. Lenth. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.8.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Preprint*, arXiv:2402.14848.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *Preprint*, arXiv:2311.04939.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the ACL*, 12:157–173.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Shinichi Nakagawa, Paul C. D. Johnson, and Holger Schielzeth. 2017. The coefficient of determinationr2and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, 14(134):20170213.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Model release blog: GPT-4o. Technical report, OpenAI. Accessed: 2024-05-23.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *Preprint*, arXiv:2309.00071.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *NIST Special Publication SP*, 109:109.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III au2, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. *Preprint*, arXiv:2310.12558.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling. *Preprint*, arXiv:2307.03170.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. Novelqa: A benchmark for long-range novel question answering. *Preprint*, arXiv:2403.12766.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *Preprint*, arXiv:2406.17419.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024. LFED: A literary fiction evaluation dataset for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10466–10475, Torino, Italia. ELRA and ICCL.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

A Issues with current long-context evaluation benchmarks

A.1 Evaluation may be conducted on the training data

Data contamination: One of the most prominent issues in current evaluation practices is the test data. A fundamental rule in evaluating any trained model is to test it on a separate withheld *test* set, avoiding the instances the model was trained on (van der Goot, 2021; Kapoor and Narayanan, 2023). However, this has become increasingly difficult as existing test sets have almost certainly been leaked into the training data, and most available LLMs are either entirely closed-source or only open-weights, meaning we do not have access to the training instances (Sainz et al., 2023; Balloccu et al., 2024). While analyzing existing long-context benchmarks, though this issue is certainly not limited to long-context setups, we notice that many either utilize existing datasets (e.g., Bai et al., 2023; An et al., 2024) or use old texts due to copyright issues (e.g., Li et al., 2023; Wang et al., 2024a), both of which are likely included in the models' training data. Some researchers have tried to mitigate this issue by transforming the original texts, for instance, by replacing the named entities with others (Zhang et al., 2024; Dong et al., 2023). However, we observe that prompting models with such data (even fragments of these) still leads to recognition of the source materials, an outcome in line with results reported in Chang et al. (2023).

Issues with closed-book tests: A popular method used to ensure that the training data do not affect the model's performance is to perform the *closed-book* test (Roberts et al., 2020; Ciosici et al., 2021), used, for instance, in Wang et al. (2024a). Just like a student taking a test without access to the study book, the model is tested without access to the source text. The idea is that if the model has memorized the training data, it will be able to perform well even without access to the source text. However, as was recently pointed out, memorization does not have to be perfect to impact generation (Chen et al., 2024). For instance, even a model that fails a *closed-book* test might have encountered the source text in its training data and may benefit from this during the test when presented with the task along with the source text as the context. Since little is known about memorization in LLMs, the safest evaluation is to test the model on newly produced data (e.g., novels published past the model's cutoff date). However, even this approach necessitates constant updates to the benchmark.

A.2 Employing LLMs for evaluation and data creation

Unreliability of LLM-based evaluation: Another issue is the way in which model outputs are evaluated. While multiple-choice questions (including True/False evaluations) are straightforward to verify if all plausible answers are correctly marked, they are also likely to be easier for models to answer correctly. Consequently, researchers have developed datasets requiring models to generate long-form answers or perform complex tasks involving long contexts, such as summarization (Hudson and Al Moubayed, 2022; Bai et al., 2023; Zhang et al., 2024, inter alia). These setups are more challenging for the models but also harder to evaluate. Currently, no reliable metric exists for verifying long-form outputs, leading some researchers to use language models, like GPT-4o, for evaluation (e.g., Wang et al., 2024b). This method has several flaws. Firstly, the model typically evaluates the produced answer against a gold standard, which may penalize valid answers that include additional, relevant information not present in the gold answer. Secondly, inconsistent instructions to the evaluator model can lead to variability in judgements. For example, if a model is instructed to produce a score between 0 and 100, with 100 for a perfect answer (as in Wang et al., 2024b), scores in between may vary due to inconsistent deductions for similar issues across different evaluations.

Creating test data with language models: Since employing human annotators to create test instances (e.g., questions about the documents) is time- and resource-intensive (i.e., the annotators have to read long texts),[26] some researchers use language models to generate these instances. However, this approach has its own issues. For instance, using models such as GPT-4o to generate questions and answers (as in Wang et al., 2024b) may result in flawed test data or test data that is inherently

[26]Reading an average book takes at least 8-10 hours.

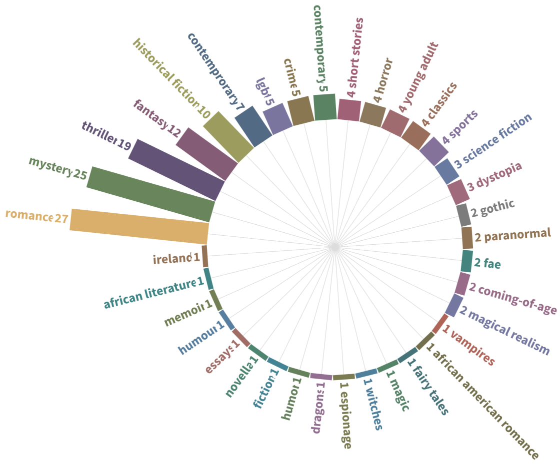


Figure 5: Genre distribution in NoCha. As a book can belong to multiple genres, such as fantasy and romance, we allow up to three labels per book.

easier for a model from the same family to solve.[27] This becomes even more complex when weaker models are used to generate test examples for relatively complex tasks. For instance, Li et al. (2023) employs GPT-3.5-Turbo to generate summaries for a cloze task, where the tested model is supposed to fill in masked named entities based on the source text. However, GPT-3.5-Turbo has been shown to be a poor summarizer, even when summarizing smaller chunks of text (Kim et al., 2024).

B Dataset

In this section of the appendix we provide more details on our corpus (§B.1) and human annotation efforts (§B.2).

B.1 Corpus

This section provides detailed information about the corpus collected for this study. Table 12 lists all the novels included in the corpus,[28] with the genre distribution illustrated in Figure 5.[29] Table 13 lists all stories included in the story collections. Additionally, we provide the the statistics for books and claims by year of publication in Table 15. Finally, we report list the statistics for stories from story collections in Table 14.

[27]In fact, on a benchmark where Q&A pairs were generated by GPT-4o, the model achieved one of the highest scores among all tested models (Wang et al., 2024b).

[28]All books in the corpus were purchased by the researchers.

[29]Please note that multiple genre tags are allowed per book, as a book can belong to more than one genre, such as both romance and mystery.

Title	Author	Gender	Genre	Pub. Date	Tokens	Words	# Pairs	Lang.
A Haunting on the Hill	Elizabeth Hand	F	horror, Gothic, paranormal	Oct 3, 2023	116,411	87,391	20	AmE
Assassin Eighteen	John Brownlow	M	thriller, crime	Apr 23, 2024	126,011	96,830	15	CanE
Come and Get It	Kiley Reid	F	contemporary	Jan 30, 2024	141,484	104,373	30	AmE
Curse of the Soul Collector	Cara Blaine	F	fantasy	Sep 20, 2023	85,012	67,495	15	AmE
Death Comes to Marlow	Robert Thorogood	M	mystery, fiction, crime	Jan 5, 2023	113,544	87,844	15	BrE
Divine Rivals	Rebecca Ross	F	romance, young adult, historical fiction	Apr 4, 2023	135,800	106,371	8	AmE
Everyone on the train is a suspect	Benjamin Stevenson	M	mystery	Oct 17, 2023	108,673	82,296	7	AusE
First Lie Wins	Ashley Elston	F	thriller, mystery	Jan 2, 2024	117,547	97,151	15	AmE
Fourth Wing	Rebecca Yarros	F	fantasy, romance, dragons	May 2, 2023	242,683	183,789	9	AmE
Funny Story	Emily Henry	F	romance, contemporary	Apr 23, 2024	138,892	104,775	14	AmE
Helpless	Kelby C. Hughes	F	fantasy	Mar 25, 2024	65,677	53,381	15	AmE
Home Is Where the Bodies Are	Jeneva Rose	F	thriller, mystery, horror	Apr 30, 2024	97,390	76,866	14	AmE
House of Flame and Shadow	Sarah J. Maas	F	fantasy, romance, fae	Jan 30, 2024	334,581	248,192	15	AmE
How to solve your own murder	Kristen Perrin	F	mystery, thriller, crime	Mar 26, 2024	130,414	104,156	10	AmE
I Hope This Doesn't Find You	Ann Liang	F	romance, young adult	Feb 6, 2024	105,934	81,432	9	AusE
Inheritance	Nora Roberts	F	paranormal, romance, mystery	Nov 21, 2023	170,979	127,511	40	AmE
Leaving	Roxana Robinson	F	romance, contemporary	Feb 13, 2024	133,317	101,039	10	AmE
Long Island	Colm Toibin	M	historical fiction, Ireland	May 7, 2024	103,361	84,339	10	IrE
Love, Lies, and Cherry Pie	Jackie Lou	F	contemporary, romance	May 7, 2024	112,179	86,994	15	AmE
Monstrous Alterations	Christopher Barzak	M	short stories, fantasy, LGBT	Sep 8, 2023	73,094	59,717	10	AmE
Only for the week	Natasha Bishop	F	African American romance	May 11, 2023	85,969	68,517	5	AmE
Pet	Catherine Chidgey	F	thriller, mystery	July 13, 2023	124,189	93,294	46	NZE
Random in Death	J. D. Robb	F	mystery, romance, crime	Jan 23, 2024	132,107	97,172	30	AmE
Romantic Comedy	Curtis Sittenfeld	F	romance, contemporary	Apr 4, 2023	115,004	88,775	12	AmE
Roux	Tamika Christy	F	romance, historical fiction, LGBT	Jan 9, 2024	121,364	89,221	15	AmE
Ruthless Vows	Rebecca Ross	F	fantasy, romance, young adult	Dec 26, 2023	161,337	127,090	15	AmE
Safe and Sound	Laura McHugh	F	thriller, mystery	Apr 23, 2024	103,054	79,243	10	AmE
Same Time Next Year	Tessa Bailey	F	romance, sports, novella	Apr 18, 2023	49,156	38,023	35	AmE
She Is a Haunting	Trang Thanh Tran	F	horror, young adult, LGBT	Feb 28, 2023	105,378	80,291	41	AmE
Six scorched roses	Carissa Broadbent	F	fantasy, romance, vampires	Mar 21, 2023	51,779	40,062	6	AmE
The Agency for Scandal	Laura Wood	F	historical fiction, romance, mystery	Jan 5, 2023	115,382	90,741	18	AmE
The Atonement Murders	Jenifer Ruff	F	mystery, thriller	Apr 14, 2023	104,258	82,134	4	AmE
The Beautiful and the Wild	Peggy Townsend	F	thriller, mystery	Nov 7, 2023	92,424	75,908	16	AmE
The book of love	Kelly Link	F	mystery, magical realism, thriller	Feb 13, 2024	272,343	209,950	10	AmE
The Bootleggers Daughter	Nadine Nettman	F	mystery, thriller	May 1 2024	97,386	75,696	15	AmE
The eye of the bedlam bride	Matt Diniman	M	fantasy, science fiction, humor	Jul 2, 2023	336,288	257,445	15	AmE
The Future	Naomi Alderman	F	science fiction, dystopia, fantasy	Nov 7, 2023	157,019	122,342	14	BrE
The Glass Woman	Alice McIlroy	F	historical fiction, Gothic, mystery	Dec 19, 2023	99,113	78,795	15	BrE
The Guest	Emma Cline	F	mystery, thriller	May 16, 2023	89,042	68,735	20	AmE
The Hanging City	Charlie Holmberg	F	fantasy, romance, magic	Aug 1, 2023	140,622	106,182	8	AmE
The Heiress	Rachel Hawkins	F	mystery, thriller	Jan 9, 2024	94,406	73,954	15	AmE
The Husbands	Holly Gramazio	F	romance, magical realism	Apr 2, 2024	130,017	100,432	15	AusE
The Last Murder at the End of the World	Stuart Turton	M	mystery, thriller, dystopia	May 21, 2024	122,364	91,280	15	BrE
The Library of Borrowed Hearts	Lucy Gilmore	F	contemporary, romance	Apr 30, 2024	132,205	105,259	15	AmE
The Limits: A Novel	Nell Freudenberger	F	contemporary	Apr 9, 2024	131,766	102,133	15	AmE
The Marriage Act	John Marrs	M	thriller, science fiction, dystopia	Jan 19, 2023	134,493	106,341	16	BrE
The Promise of Tomorrow	Mary Ellen Taylor	F	contemporary, romance	June 1 2024	116,035	89,682	15	AmE
The Prospects	KT Hoffman	F	romance, LGBT, sports	Apr 9, 2024	122,720	94,591	15	AmE
The Resort	Sarah Ochs	F	thriller, mystery	Feb 6, 2024	124,200	99,627	15	AmE
The Spy Coast	Tess Gerritsen	F	mystery, thriller, espionage	Nov 1, 2023	126,906	98,326	8	AmE
The Tainted Cup	Robert Jackson Benett	M	mystery, thriller	Feb 6, 2024	162,289	119,350	10	AmE
The Teacher	Freida McFadden	F	thriller, mystery	Feb 6, 2024	110,260	88,104	15	AmE
The White Lady	Jacqueline Winspear	F	historical fiction, mystery	Mar 21, 2023	124,585	97,094	4	BrE
Viciously Yours	Jaime Applegate	F	fantasy, romance, fae	Jan 23, 2024	85,118	65,986	7	AmE
Weyward	Emilia Hart	F	historical fiction, fantasy, witches	Feb 2, 2023	126,501	98,842	9	AusE
White Cat, Black Dog	Kelly Link	F	short stories, horror, fairy tales	Mar 2, 2023	92,826	74,154	10	AmE
Wildfire	Hannah Grace	F	romance, sports, contemporary	Oct 3, 2023	138,441	109,641	5	BrE
Yellowface	R.F Kuang	F	thriller, contemporary, mystery	May 25, 2023	113164	87648	12	AmE
You Should Be So Lucky	Cat Sebastian	F	romance, LGBT, sports	May 7, 2024	140,422	109,764	15	AmE
You, Again	Kate Goldbeck	F	romance, contemporary	Sep 12, 2023	128,445	96,681	7	AmE
Yours truly	Abby Jimenez	F	romance, contemporary	April 11, 2023	134,609	105,968	10	AmE
Quietly Hostile	Samantha Irby	F	essays, humour, memoir	May 16, 2023	95,842	76,424	14	AmE
Anne of Green Gables	L. M. Montgomery	F	classics, coming-of-age, historical	Jan 1, 1908	129,908	102,366	15	CanE
Little Women	Louisa May Alcott	F	classics, coming-of-age, historical, romance	Sep 30, 1868	235,118	185,930	15	AmE
The Adventures of Sherlock Holmes	Arthur Conan Doyle	M	classics, short stories, mystery, crime	Oct 14, 1892	129,293	104,434	18	BrE
The Great Gatsby	F. Scott Fitzgerald	M	classics, historical fiction, romance	Apr 10, 1925	61,689	48,187	15	AmE

Table 12: List of novels included in NoCha. The upper row presents all recent fictional books included in the data. The middle row displays the essay collections, while the lower row shows classical books included in the data. The genre is provided as listed on GoodReads (https://www.goodreads.com/). The language is indicated based on the author's native language or, for non-native English authors, the primary language of the country where they spent most of their time. The token count is reported as per `tiktoken` tokenization with `cl100k` encoding while the word count was determined by a `whitespace` split.

B.2 Human annotation efforts

In this section of the appendix we provide additional details about the data collection pipeline.

Annotators: The annotations were done by 18 annotators recruited on Upwork (all female) and 5 volunteer annotators, including three of the authors (four females and one male). All but four annotators are native English speakers from either the US or UK. All annotators hold university degrees, with one being a former college English professor and an author, and one holding a Ph.D. in English literature. As frequent readers, some annotators received advanced reader copies of the books, enabling the annotation of works, which were unpublished at the time of data collection.

ID	Collection	Author	Story Title	# Pairs	Story Tokens	Collection Tokens	Location	Pub. Date
1	A Kind of Madness	Uche Okonkwo	Nwunye Belgium	2	10299	80946	*beginning*	Apr 16, 2024
2	A Kind of Madness	Uche Okonkwo	Shadow	3	8549	80946	*beginning*	Apr 16, 2024
3	A Kind of Madness	Uche Okonkwo	Debris	3	1822	80946	*beginning*	Apr 16, 2024
4	A Kind of Madness	Uche Okonkwo	Long Hair	3	3464	80946	*beginning*	Apr 16, 2024
5	A Kind of Madness	Uche Okonkwo	Animals	3	8193	80946	*beginning*	Apr 16, 2024
6	A Kind of Madness	Uche Okonkwo	Milk and Oil	3	11829	80946	*middle*	Apr 16, 2024
7	A Kind of Madness	Uche Okonkwo	The Harvest	2	7145	80946	*middle*	Apr 16, 2024
8	A Kind of Madness	Uche Okonkwo	Eden	2	7408	80946	*middle*	Apr 16, 2024
9	A Kind of Madness	Uche Okonkwo	The Girl Who Lied	3	12834	80946	*end*	Apr 16, 2024
10	A Kind of Madness	Uche Okonkwo	Burning	3	9524	80946	*end*	Apr 16, 2024
1	The Adventures of Sherlock Holmes	Arthur Conan Doyle	A Scandal in Bohemia	2	10715	129293	*beginning*	Oct 14, 1892
2	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Red Headed League	1	11369	129293	*beginning*	Oct 14, 1892
3	The Adventures of Sherlock Holmes	Arthur Conan Doyle	A Case of Identity	1	8715	129293	*beginning*	Oct 14, 1892
4	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Boscombe Valley Mystery	3	11748	129293	*beginning*	Oct 14, 1892
5	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Five Orange Pips	2	9076	129293	*beginning*	Oct 14, 1892
6	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Man with the Twisted Lip	1	11401	129293	*middle*	Oct 14, 1892
7	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Adventure of Blue Carbuncle	1	9982	129293	*middle*	Oct 14, 1892
8	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Adventure of the Speckled Band	2	12070	129293	*middle*	Oct 14, 1892
9	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Adventures of the Engineers Thumb	1	10227	129293	*middle*	Oct 14, 1892
10	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Adventure of the Noble Bachelor	1	10086	129293	*end*	Oct 14, 1892
11	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Adventure of the Beryl Coronet	2	11776	129293	*end*	Oct 14, 1892
12	The Adventures of Sherlock Holmes	Arthur Conan Doyle	The Red Adventure of the Copper Beeches	1	12331	129293	*end*	Oct 14, 1892
1	Quietly Hostile	Samantha Irby	i like it!	0	1783	95842	*beginning*	May 16, 2023
2	Quietly Hostile	Samantha Irby	the last normal day	1	4188	95842	*beginning*	May 16, 2023
3	Quietly Hostile	Samantha Irby	david matthews's greatest romantic hits	1	4563	95842	*beginning*	May 16, 2023
4	Quietly Hostile	Samantha Irby	chub street diet	1	4805	95842	*beginning*	May 16, 2023
5	Quietly Hostile	Samantha Irby	my firstborn dog	1	4672	95842	*beginning*	May 16, 2023
6	Quietly Hostile	Samantha Irby	body horror!	1	4798	95842	*beginning*	May 16, 2023
7	Quietly Hostile	Samantha Irby	two old nuns having amzing [sic] lesbian sex	1	7138	95842	*beginning*	May 16, 2023
8	Quietly Hostile	Samantha Irby	qvc, ilysm	1	5974	95842	*beginning*	May 16, 2023
9	Quietly Hostile	Samantha Irby	superfan!!!!!!	1	11397	95842	*middle*	May 16, 2023
10	Quietly Hostile	Samantha Irby	i like to get high at night and think about whales	0	702	95842	*middle*	May 16, 2023
11	Quietly Hostile	Samantha Irby	oh, so you actually don't wanna make a show about a horny fat bitch with diarrhea? okay!	1	12351	95842	*middle*	May 16, 2023
12	Quietly Hostile	Samantha Irby	what if i died like elvis	1	9068	95842	*middle*	May 16, 2023
13	Quietly Hostile	Samantha Irby	shit happens	1	4960	95842	*end*	May 16, 2023
14	Quietly Hostile	Samantha Irby	food fight	0	1508	95842	*end*	May 16, 2023
15	Quietly Hostile	Samantha Irby	o brother, where art thou?	1	4865	95842	*end*	May 16, 2023
16	Quietly Hostile	Samantha Irby	how to look cool in front of teens?	1	5723	95842	*end*	May 16, 2023
17	Quietly Hostile	Samantha Irby	we used to get dressed up to go to red lobster	1	5558	95842	*end*	May 16, 2023
18	Quietly Hostile	Samantha Irby	please invite me to your party	0	2049	95842	*end*	May 16, 2023
1	White Cat, Black Dog	Kelly Link	The White Cat's Divorce	2	11697	92826	*beginning*	Mar 2, 2023
2	White Cat, Black Dog	Kelly Link	Prince Hat Underground	2	19724	92826	*beginning*	Mar 2, 2023
3	White Cat, Black Dog	Kelly Link	The White Road	2	11770	92826	*middle*	Mar 2, 2023
4	White Cat, Black Dog	Kelly Link	The Girl Who Did Not Know Fear	1	8606	92826	*middle*	Mar 2, 2023
5	White Cat, Black Dog	Kelly Link	The Game of Smash and Recovery	0	6869	92826	*middle*	Mar 2, 2023
6	White Cat, Black Dog	Kelly Link	The Lady and the Fox	1	12504	92826	*middle*	Mar 2, 2023
7	White Cat, Black Dog	Kelly Link	Skinder's Veil	2	21734	92826	*end*	Mar 2, 2023

Table 13: List of stories included in the collections. We provide token counts for both the entire collection and individual story. The location was determined by dividing collection length into three parts, where the stories which begin in the first part are marked as *beginning*, stories which begin in the second part are marked as *middle*, and stories beginning in the third part are marked as *end*.

	Stories
Mean	8,501.5
Max	21,734
Min	702
Std. Dev.	4,355.7

Table 14: Statistics of stories in story collections (in tokens). Note that there is 7–18 stories per collection.

Collecting True/False claim pairs: Before starting the annotation task, all annotators were required to read the guidelines (Figure 6) and sign a consent form (Figure 7). The collected claims underwent a rigorous review process, being checked at least three times by two of the authors: initially during the writing stage and later independently during the final quality control phase. During this phase, we also proofread all claims and contacted annotators if anything was unclear.

Collecting each pair of annotations typically required additional communication with the annotators, resulting in approximately 160 hours of work from each of the two authors involved in this process. Based on self-reported time and records in spreadsheets, we estimate that the annotators were able to produce between 6 and 10 claim pairs per hour.[30]

Overall, we collected about 15 claim pairs per book. While we initially aimed to collect more claim pairs per book, we observed that creating meaningful and challenging pairs becomes significantly more difficult beyond the first 10-15 pairs, though this number may vary slightly depending on the book. For some books, we collected fewer than 10 pairs due to the unavailability of annotators to create more pairs.

[30]The annotators we also asked to provide written justification as to why the true claim is true and the false claim is false relating to the events in the book.

	Books			Claim [Pairs]		
	2024 (n=33)	2023 (n=30)	Classics (n=4)	2024 (n=898 [449])	2023 (n=978 [489])	Classics (n=126 [63])
TOKENS (tiktoken)						
Mean	129,526	123,908	139,002	23.51	22.75	24.71
St. dev.	52,112	52,176	71,630	7.17	7.94	7.84
Max	334,581	336,288	235,118	63	57	45
Min	65,677	49,156	61,689	5	8	10
WORDS (whitespace)						
Mean	99,753	96,117	110,229	18.44	17.85	20.06
St. dev.	38,463	39,388	56,790	6.15	6.68	6.97
Max	248,192	257,445	185,930	57	46	39
Min	53,381	38,023	48,187	4	5	7

Table 15: Number of tokens and words across books and claims for books published in 2024, 2023, and classics. The number of claim pairs is indicated in [] square brackets.

Advantages of a minimal pair design We employ minimal pairs for two main reasons. Firstly, it ensures data quality, allowing us to verify the false claim against its true counterpart easily and identify cases where both claims were too similar (i.e., the false claim could potentially be true) or subjective. Secondly, we require models to correctly label *both* claims in a pair, which minimizes the chances of crediting them for correct predictions made "for the wrong reason" without proper utilization of the context.

Indeed, the results presented in the paper indicate that merely having a balanced set of labels is insufficient for accurately evaluating the models (see §4 and examples in Figure 2). For example, using balanced but unpaired data might lead us to conclude that GPT-4o achieves an overall accuracy of 76.7% (77.5% for True claims and 75.9% for False claims; see Table 11). However, these results are misleading because the model fails to validate *both* claims correctly in approximately 20% of cases, resulting in an actual accuracy of 55.8%. This is significantly lower than the human accuracy of 96.9% reported in §2, highlighting a substantial performance gap for the best model to bridge.

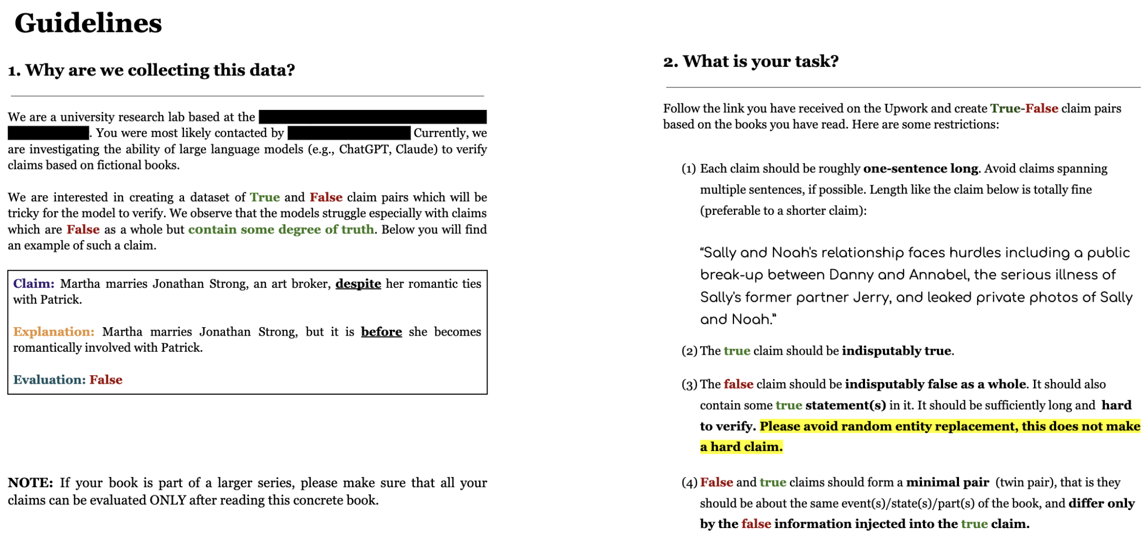
Quality control: To ensure the quality of the annotations, we reannotated a subset of the data, consisting of 128 claims made about 5 books (64 claim pairs). For this task, we hired annotators who had read the same books. These annotators were also part of the original annotation team but had worked on different books previously. They were provided with *the same* instructions as the prompt given to the model (Table 3), with the only modification being the phrase "based on the context provided" changed to "based on your book." All annotations were performed using `GoogleSpreadsheets`.

SCOPE	PAIRS (%)	TRUE (%)	FALSE (%)
GLOBAL	47.9%	38.0%	47.1%
PASSAGE	39.7%	43.8%	39.7%
SENTENCE	12.4%	18.2%	13.2%

Table 16: Percentage of pairs by scope in the annotated subset of data (*global*, *passage*, *sentence*). Additionally, we report the percentage of *claims* with specific labels for both True and False claims.

Overall, the annotators agreed on 124 out of 128 labels (62 out of 64 claim pairs). We then reviewed the two pairs where the annotators disagreed. In one pair, the disagreement was due to a sentence-level detail about whether the main characters were engaged for two or three years. The characters were together for three years but engaged for only two, a detail mentioned in the book only once. The second annotator incorrectly annotated the claim of three years as true and the two years claim as false. In the second case, the claim was about "The Great Gatsby," where the initial annotator confused parts of the book with parts of the movie. Importantly, none of the core test books have movies made based on them as they were published within the last year, so this situation was specific to this classic novel. We have since reviewed all the classic claims again to ensure this issue does not recur.

Evidence scope: For this task, we asked four annotators to annotate the scope of the evidence for two of their books each, resulting in annotations for 121 claim pairs from 8 books (approximately 15 claim pairs per book; see Figure 8 for instructions given to the annotators). The annotators first anno-



Guidelines

1. Why are we collecting this data?

We are a university research lab based at the [redacted]. You were most likely contacted by [redacted] Currently, we are investigating the ability of large language models (e.g., ChatGPT, Claude) to verify claims based on fictional books.

We are interested in creating a dataset of **True** and **False** claim pairs which will be tricky for the model to verify. We observe that the models struggle especially with claims which are **False** as a whole but **contain some degree of truth**. Below you will find an example of such a claim.

> **Claim:** Martha marries Jonathan Strong, an art broker, **despite** her romantic ties with Patrick.
>
> **Explanation:** Martha marries Jonathan Strong, but it is **before** she becomes romantically involved with Patrick.
>
> **Evaluation: False**

NOTE: If your book is part of a larger series, please make sure that all your claims can be evaluated ONLY after reading this concrete book.

(a) Guidelines: Page 1

2. What is your task?

Follow the link you have received on the Upwork and create **True-False** claim pairs based on the books you have read. Here are some restrictions:

(1) Each claim should be roughly **one-sentence long**. Avoid claims spanning multiple sentences, if possible. Length like the claim below is totally fine (preferable to a shorter claim):

"Sally and Noah's relationship faces hurdles including a public break-up between Danny and Annabel, the serious illness of Sally's former partner Jerry, and leaked private photos of Sally and Noah."

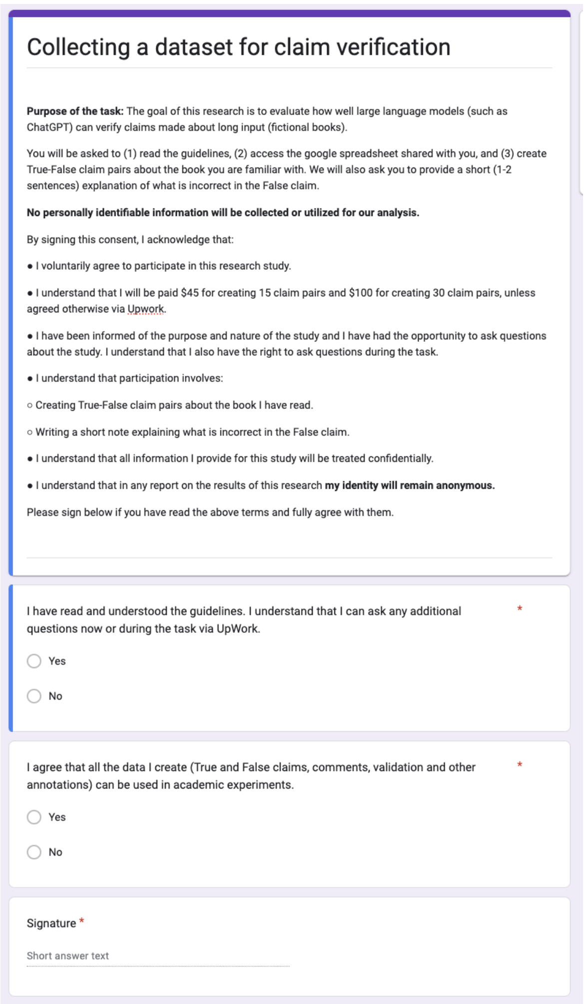
(2) The **true** claim should be **indisputably true.**

(3) The **false** claim should be **indisputably false as a whole**. It should also contain some **true statement(s)** in it. It should be sufficiently long and **hard to verify. Please avoid random entity replacement, this does not make a hard claim.**

(4) **False** and **true** claims should form a **minimal pair** (twin pair), that is they should be about the same event(s)/state(s)/part(s) of the book, and **differ only by the false information injected into the true claim.**

(b) Guidelines: Page 2

Figure 6: Guidelines provided to the annotators for the annotation task. The annotators were also provided additional examples and guidance during the data collection process.



Collecting a dataset for claim verification

Purpose of the task: The goal of this research is to evaluate how well large language models (such as ChatGPT) can verify claims made about long input (fictional books).

You will be asked to (1) read the guidelines, (2) access the google spreadsheet shared with you, and (3) create True-False claim pairs about the book you are familiar with. We will also ask you to provide a short (1-2 sentences) explanation of what is incorrect in the False claim.

No personally identifiable information will be collected or utilized for our analysis.

By signing this consent, I acknowledge that:

- I voluntarily agree to participate in this research study.
- I understand that I will be paid $45 for creating 15 claim pairs and $100 for creating 30 claim pairs, unless agreed otherwise via Upwork.
- I have been informed of the purpose and nature of the study and I have had the opportunity to ask questions about the study. I understand that I also have the right to ask questions during the task.
- I understand that participation involves:
  - Creating True-False claim pairs about the book I have read.
  - Writing a short note explaining what is incorrect in the False claim.
- I understand that all information I provide for this study will be treated confidentially.
- I understand that in any report on the results of this research **my identity will remain anonymous.**

Please sign below if you have read the above terms and fully agree with them.

I have read and understood the guidelines. I understand that I can ask any additional questions now or during the task via UpWork. *

Yes

No

I agree that all the data I create (True and False claims, comments, validation and other annotations) can be used in academic experiments. *

Yes

No

Signature *

Short answer text

Figure 7: Consent form which the annotators were asked to sign via GoogleForms before collecting the data.

tated the scope for each claim in the pair, and then the maximum scope was taken for the given pair.[31] Overall, 12.4% of pairs were labeled as requiring only one to two sentences to validate, 39.7% of pairs were labeled as requiring a longer passage, and 47.9% were labeled as requiring global reasoning. The distribution of labels for specific claims is comparable (see Table 16).

Note on evidence location: Designing a realistic experiment to verify complex claims with evidence at different depths of a book is challenging. Unlike the "needle-in-a-haystack" experiment, where the "needle" can be easily placed in different parts of the (often unrelated) context, our task involves realistic claims about the book's content, making it more problematic to control for various confounders. For instance, a claim about events at the beginning of the book might inherently be easier to verify. Similarly, it is difficult to assert with certainty that the only evidence is present at a specific depth in the book, as there could be corroborating evidence prior or later that aids verification (e.g., hints about the killer's identity), which the human reader may (reasonably) not pay attention

[31] For instance, if one claim in a pair is labeled as "passage" and the other as "global," the resulting scope for the entire pair would be "global." This is because we require the models to validate *both* claims in the pair correctly, meaning the models would have to reason over both "passage" and "global" contexts for this specific pair. It is worth noting that the labels for both claims in the pair matched 87.6% of the time.

For each claim, please tell us how much reasoning is needed to correctly validate it as either True or False. Overall, we are interested to see how much understanding/reasoning is needed to label each pair correctly (True as True and False as False). Please choose one of the following options:

(1) **sentence** - use this category if the claim can be directly validated by one or two consecutive sentences within the text. The content of these sentences alone should provide all the necessary information to affirm or refute the claim.

Example: the following fragment "Hi Sonya! Welcome to the manor. I'm Trey, Ace's son." can fully validate the following claim: "Trey, Ace's son, welcomes Sonya to the manor."

(2) **passage** - use this category if the claim requires a bit more text for validation, specifically up to three-four paragraphs. This could be a section where several sentences together provide the context needed to verify the claim.

Example: the book contains a short conversation between Sonya and Trey during which we learn everything we need to validate the following claim: "Trey, Ace's son, welcomes Sonya to the manor."

(3) **longer context** -- this category is appropriate when the information required to validate the claim is scattered throughout the text. Different aspects of the claim may be supported by different sections of the book, and piecing these together is necessary for validation.

Figure 8: Instructions given to the annotators for the annotation of claim scope.

to. Hence, we do not attempt to annotate the evidence location. Instead, we use story collections, where we can precisely identify which story the claim pertains to and where the story is located within the collection. We present the results of this experiment both in the main paper (§4) and later in the appendix (§D).

C Methods

In this section we provide additional details about our evaluation methodology.

Models: We evaluate 6 closed-source and 5 open-source long-context models (i.e., models with context window of at least 128k tokens):

- **Closed-source models:** GPT-4O (OpenAI, 2024b), GPT-4-TURBO (OpenAI, 2024a), CLAUDE-3-OPUS (Anthropic, 2024b),CLAUDE-3.5-SONNET (Anthropic, 2024a), GEMINI PRO 1.5 (Gemini Team, 2024), GEMINI FLASH 1.5 (Gemini Team, 2024). All models were accessed via provider's API.[32]

- **Open-weight models:** COMMAND R (Cohere, 2024a), COMMAND R+ (Cohere, 2024b), GEMMA-10M (Aljadery et al., 2024), PHI-3-MINI (Abdin et al., 2024), and LONGLLAMA (Tworkowski et al., 2023). COMMAND R and COMMAND R+ we accessed via `cohere` API. Other models were run on up to three A100 80GB GPUs. For comparison, in our stories experiment, we also tested MIXTRAL-8X22B (65k context window) and QWEN-2-72B (32k context window). While their context windows are too short to process entire books, they can handle stories from our collections. Both models were run using the `together.ai` API.

- **Retrieval pipeline:** We also implement a retrieval pipeline (BM25+GPT-4O), where we first retrieve the evidence (top 5, 25, and 50 paragraphs) and then prompt GPT-4O for evaluation.

Inference: We generate the labels by prompting the models with the entire book and one claim at time for verification. All closed-source models were prompted using the provider's API, while open-weight models were run on up to three A100 80GB GPUs, excluding COMMAND R and COMMAND R+, which were accessed using the `cohere` API. Additionally, we implement a BM25+GPT-4O retrieval pipeline where we first retrieve the evidence from the book and then verify the claim based on the retrieved evidence with GPT-4O.[34]

Prompts: All models were prompted with the prompt presented in Table 3. As we notice that the open-source models did not follow the instructions well, we also tested them with a simplified prompt as shown in Table 18. Finally, for BM25+GPT-4O we employed the prompt in Table 19. All generations are restricted to 800 tokens.[35]

During a pilot study, we first tested three different types of prompts on a subset of data (10 books, 176 claim pairs):

- **Answer-only:** prompting the model for the answer only (Table 20)

- **Answer-then-explanation:** prompting the model for the answer followed by an explanation (Table 21)

[32] CLAUDE-3-OPUS's generations were done partially employing `anthropic` API and partially utilizing `vertex-ai` due to the rate limit. All CLAUDE-3.5-SONNET generations were done utilizing `vertex-ai`.

[34] The estimated cost for running all data with each model is as follows: GPT-4O $640 USD, GPT-4-TURBO $1,280 USD, CLAUDE-3-OPUS $3,350 USD, CLAUDE-3.5-SONNET $670 USD, GEMINI PRO 1.5 1,385 USD, COMMAND R+ $450 USD, COMMAND R $74 USD, BM25+GPT-4O pipeline $330 USD. Note that test runs resulted in small additional costs.

[35] For some open-source models we noticed that the generation are plagued by repetitions and reduced this limit to 600 tokens.

Model	Context	Avail.	Checkpoints	# Param
GPT-4o	128k	🔒	gpt-4o-2024-05-13	¯_(ツ)_/¯
GPT-4-Turbo	128k	🔒	gpt-4-turbo-2024-04-09	¯_(ツ)_/¯
Claude-3-Opus	200k	🔒	claude-3-opus-20240229	¯_(ツ)_/¯
Claude-3.5-Sonnet	200k	🔒	claude-3-5-sonnet@20240620	¯_(ツ)_/¯
Gemini Pro 1.5	1M	🔒	gemini-1.5-pro-preview-0514	¯_(ツ)_/¯
Gemini Flash 1.5	1M	🔒	gemini-1.5-flash-preview-0514	¯_(ツ)_/¯
Command R	128k	🔓	c4ai-command-r-v01	35B
Command R+	128k	🔒	c4ai-command-r-plus	104B
Gemma-10M	10M	🔓	gemma-2b-10m	2B
Phi-3-mini	128k	🔓	Phi-3-mini-128k-instruct	3.8B
LongLLaMA	256k	🔓	long_llama_3b_instruct	3B
Mixtral-8x22B	65k	🔓	Mixtral-8x22B-Instruct-v0.1	141B
Qwen-2-72B	32k[33]	🔓	Qwen2-72B-Instruct	72B

Table 17: Evaluated models: the upper part displays all closed-source models, while the lower part lists all open-weight models. We also provide details of two shorter context models, which were tested on story-length inputs for comparison.

- **Explanation-then-answer:** prompting the model for the explanation followed by the answer (Table 3)[36]

As testing these prompts on all models would be prohibitively expensive, we conducted these experiments with GPT-4o and Claude-3-Opus.[37] We observed that Claude-3-Opus's accuracy remained constant across all three setups (39.77%), although the specific pairs it got correct varied. GPT-4o achieved the highest accuracy when prompted for the explanation followed by the answer (50%), compared to 48.3% for the answer followed by the explanation, and 49.43% for the answer-only setup. For further experiments, we employ the *explanation-then-answer* approach, as it yielded the best results despite the small differences between these three methods.

Label extraction: As we prompt the models to generate answers following a structured template, we first attempt to extract the answer from the first encountered <answer></answer> tags.[38] If this is not possible (e.g., the model did not generate the tags or produced output longer than a simple true/false within the tags), we apply the following steps:

1. Replace any occurrence of "true or false" and the text of the claim itself with an empty string.[39]

2. Replace "not true" with "false."

Evaluation Template (Simplified)

```
You are provided with a context and a statement.
Your task is to carefully read the context and
then determine whether the statement is true or
false.

Answer TRUE if the statement is true in its
entirety based on the context provided.
Answer FALSE if any part of the statement is
false based on the context provided.

<context>book text</context>

<statement>claim</statement>

<question>Based on the context provided, is the
above statement TRUE or FALSE?</question>
```

Table 18: Simplified prompt template used for Evaluation of open-source models only.

[36]We gradually refined this prompt after the first pilot run on a few examples showed that models sometimes generate unreasonably long explanations, leaving no space for the actual answer. To mitigate this issue, we added a request to provide the explanation "in at most one paragraph." Although we still observed some cases where the model's explanation was too long to generate the <answer></answer> tags, this occurred in only twice for Claude-3-Opus and once for GPT-4-Turbo. Importantly, the answer/label was always present in the explanation itself (early answer), so it was correctly extracted for that model.

[37]The total cost of this experiment, including iterative refinements to the prompt text, was $3k USD.

[37]This version can also be extended to 128k using YaRN (Peng et al., 2023).

[38]In cases where the text between the tags is longer than simply true/false, we follow the replacement pipeline and finally extract the label from the text between the tags.

[39]"True or false" often occurs when the model repeats the original question. We additionally replace the claim itself, as it can also contain words like true/false (e.g., "true friend").

Evaluation Template (BM25)

```
You are provided with excerpts of context and a statement. Your task is to carefully read the excerpts and then determine whether the statement is true or false.

Answer TRUE if the statement is true in its entirety based on the excerpts provided.
Answer FALSE if any part of the statement is false based on the excerpts provided.

<excerpt_1>excerpt_1</excerpt_1>

<excerpt_2>excerpt_2</excerpt_2>

<excerpt_(i)>...</excerpt_(i)>

<excerpt_k>excerpt_k</excerpt_k>

<statement>claim</statement>

<question>Based on the excerpts provided, is the above statement TRUE or FALSE?</question>

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:
<explanation>YOUR EXPLANATION</explanation>
<answer>YOUR ANSWER</answer>
```

Table 19: Prompt used for the BM25+GPT-4o Evaluation pipeline.

Evaluation Template (Answer-only)

```
You are provided with a context and a statement. Your task is to carefully read the context and then determine whether the statement is true or false.

Answer TRUE if the statement is true in its entirety based on the context provided.
Answer FALSE if any part of the statement is false based on the context provided.

<context>book text</context>

<statement>claim</statement>

<question>Based on the context provided, is the above statement TRUE or FALSE?</question>

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:
<answer>YOUR ANSWER</answer>
```

Table 20: Prompt used for the initial pilot study when the model is prompted to return only the answer.

3. Extract "true" if only "true" is present, "false" if only "false" is present, and the first occurrence if both are present in the generation.

If neither "true" nor "false" is present, we count it as an automatic failure.

Early response: We employ the *explanation-then-answer* template to allow the model to reason about the answer before providing its final choice. However, we observe that models sometimes still return the answer first, within the <explanation></explanation> tags, which is in line with the observations in Levy et al. (2024).[40] Overall, Claude-3-Opus and Claude-3.5-Sonnet are the most affected, with 68.4% and 55.7% of explanations containing either "true" or "false" in the first sentence, respectively (see Table 22).

Generation issues: We encounter several issues during model prompting. Most notably, Gemini Pro 1.5 and Gemini Flash 1.5 return a "prohibited content" API error for about 48.6% of claims, significantly reducing the number of claims (and claim pairs) processed by these models despite their large claimed context window. We also observe that open-source models struggle to follow the assigned template. Only Command R returns the answer with proper <answer></answer> tags 100% of the time. This compares to Command R+ at 66.3%, Phi-3-mini at 62.7%, and Gemma-10M at 52.5%. Additionally, we notice that generations by Gemma-10M are plagued with code output, often not including any instances of "true" or "false," while LongLLaMA returns empty outputs for the prompt in Table 3 (i.e., the non-simplified version of the prompt).

D Results

In this section of the appendix we provide more results of our analysis.

Table 23 reports the accuracy of each model excluding the classic novels. Overall, GPT-4o still performs the best from all the models, although its accuracy drops slightly from 55.8% to 55.1% when classic novels are excluded.

Are newer books harder to verify? Our data includes (1) classic books, which were almost certainly in the models' training data, (2) books published in 2023, which might have been included in the training data, and (3) books published in 2024, which were likely not included in the training data.[41] Hence, we look at models' performance

[40] We identify instances of an early response by segmenting the text between the <explanation></explanation> tags into sentences. In the first sentence, we remove occurrences of "true or false," "to determine if the statement is true," and similar phrases. We then check if the first sentence contains "true" or "false." Note that this excludes phrases like "this statement is incorrect." Additionally, we acknowledge that phrases such as "true nature," which may appear in the first sentence, may slightly affect this percentage as they are irrelevant to the model's answer.

[41] Given the time required to collect, clean the data, and train the model, we hypothesize that at least books published

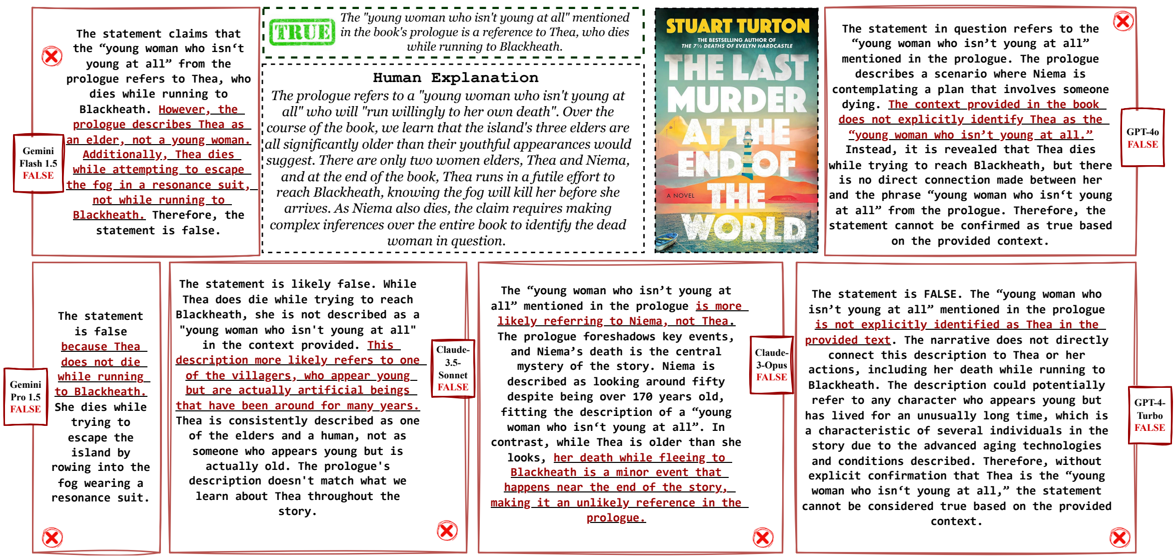


Figure 9: Example of a True claim for which all models predicted incorrect label. We provide the claim along with the annotator’s explanation for clarity.

Evaluation Template (Answer and Explanation)

You are provided with a context and a statement. Your task is to carefully read the context and then determine whether the statement is true or false.

Answer TRUE if the statement is true in its entirety based on the context provided.
Answer FALSE if any part of the statement is false based on the context provided.

<context>**book text**</context>

<statement>**claim**</statement>

<question>Based on the context provided, is the above statement TRUE or FALSE?</question>

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:
<answer>YOUR ANSWER</answer>
<explanation>YOUR EXPLANATION</explanation>

Table 21: Prompt used for the initial pilot study when the model is prompted to return the answer and then the explanation.

Model	Percentage	Count (Present/Total)
GPT-4o	1.1%	13/1234
GPT-4-Turbo	53.2%	657/1234
Claude-3-Opus	68.4%	1281/1874
Claude-3.5-Sonnet	55.7%	1044/1874
Gemini Pro 1.5	44.0%	453/1029
Gemini Flash 1.5	24.3%	250/1030

Table 22: Percentage and count of explanations where the first sentence contains an early response by the model. Open-source models are excluded as they typically do not generate structured responses or explanations.

by the publication year (see Table 24). While we do not observe a large difference in performance for books published in 2023 vs. 2024, GPT-4o and Gemini Flash 1.5’s accuracy drops slightly for the newer books, from 56.7% to 53.5% and from 37.4% to 30.7%, respectively. On the other hand, the performance of Claude-3-Opus, Claude-3.5-Sonnet, Gemini Pro 1.5, and GPT-4-Turbo seems to be higher for books published in 2024 than in 2023. Overall, these differences may be due to the inherent difficulty of the claims for a given model rather than the publication year itself. We do observe that both GPT-4o and GPT-4-Turbo perform much better on "The Great Gatsby" (the only classic book, which fits their context window) at 73.3% and 66.7%, respectively. However, no definitive claims can be made as the set of claims is too small and related to one book only.

It is unclear if claims about longer books are harder to verify: We categorized NoCha books into four buckets based on length: (1) up to 75k tokens, (2) 75k to 127k tokens, (3) 127k to 180k tokens, and (4) above 180k tokens.[42] We observed a slight drop in performance between the first and second buckets for Claude-3-Opus (56.8% to 48.5%), GPT-4o (59.3% to 55.2%), and GPT-4-

in May/June 2024 were not included in the models’ training data.

[42]We selected these numbers based on the mean length of our books (127k tokens), adjusted by adding or subtracting one standard deviation. Additional buckets were included beyond these limits on both ends.

MODEL	PAIR ACC$_{\text{(correct/total)}}$	COMMON SET ACC
GPT-4O	**55.3**$_{(333/602)}$	**57.5**$_{(195/339)}$
GPT-4-TURBO	39.5$_{(238/602)}$	38.9$_{(132/339)}$
CLAUDE-3-OPUS	49.4$_{(439/889)}$	49.9$_{(169/339)}$
CLAUDE-3.5-SONNET	42.0$_{(373/889)}$	41.3$_{(140/339)}$
GEMINI PRO 1.5	48.1$_{(224/466)}$	48.4$_{(164/339)}$
GEMINI FLASH 1.5	34.5$_{(161/467)}$	35.1$_{(119/339)}$
COMMAND R	18.8$_{(81/430)}$	*n/a*
COMMAND R$_{\text{simple}}$	22.1$_{(95/430)}$	*n/a*
COMMAND R+	17.4$_{(75/430)}$	*n/a*
COMMAND R+$_{\text{simple}}$	13.3$_{(57/430)}$	*n/a*
PHI-3-MINI	9.7$_{(23/237)}$	*n/a*
PHI-3-MINI$_{\text{simple}}$	14.2$_{(45/316)}$	*n/a*
GEMMA-10M	4.2$_{(39/938)}$	*n/a*
GEMMA-10M$_{\text{simple}}$	7.7$_{(72/938)}$	*n/a*
LONGLLAMA$_{\text{simple}}$	5.1$_{(45/889)}$	*n/a*
BM25+GPT-4O (k=5)	27.8$_{(261/938)}$	28.1$_{(95/338)}$
BM25+GPT-4O (k=25)	44.0$_{(413/938)}$	45.9$_{(155/338)}$
BM25+GPT-4O (k=50)	49.5$_{(464/938)}$	50.3$_{(170/338)}$
RANDOM	25.0$_{(234/938)}$	25.0$_{(85/339)}$

Table 23: Model accuracy on claim pairs for all data excluding *classic* novels (see Table 3 for the prompt; and Table 19 for the prompt employed with BM25). "COMMON SET" refers to claim pairs shared among the models. The subscript "SIMPLE" refers to the calls done with simplified prompt (Table 18).

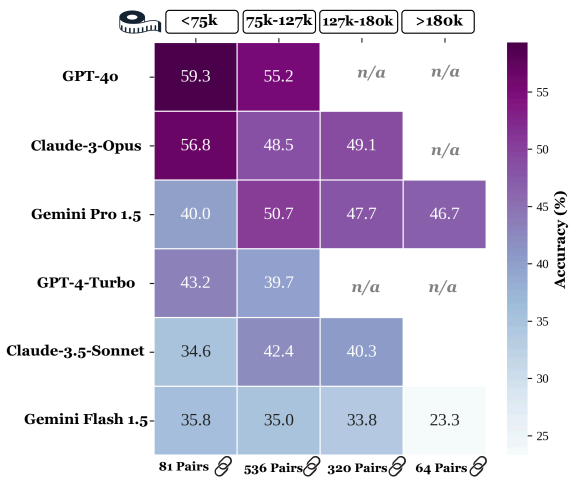


Figure 10: Model performance across different book lengths. Token counts are provided as per `tiktoken`. Length buckets were determined by taking one standard deviation from the mean on both sides, with additional buckets for values above and below this range. The number of valid pairs in each bucket is provided below.

TURBO (43.2% to 39.7%). However, these differences are small and could be influenced by factors such as the number of claims in each bucket or the complexity of the book itself (i.e., longer books tend to have more complex narratives). No such difference was observed for GEMINI FLASH 1.5 and CLAUDE-3.5-SONNET, which performance is similar or slightly better, albeit still low, for longer books. GEMINI PRO 1.5 performed worst for the shortest books (40.0%), with performance for longer books varying from 46.7% to 50.7%. See Table 25 and Figure 10 for the details. We also report Pearson correlation between the book length and model's accuracy for each model (see Table 26).

Impact of evidence scope: We report the performance of all models on the subset of data annotated for evidence scope in Table 27. Note that the number of examples varies between models due to the restrictions of the models' context windows and GEMINI PRO 1.5/GEMINI FLASH 1.5 API errors. Additionally, the performance of closed-source models is visualized in Figure 11.

MODEL	CLASSICS (correct/total)	2023 (correct/total)	2024 (correct/total)
GPT-4O	73.3 (11/15)	56.7 (194/342)	53.5 (139/260)
GPT-4-TURBO	66.7 (10/15)	39.5 (135/342)	39.6 (103/260)
CLAUDE-3-OPUS	50.0 (24/48)	48.8 (227/465)	50.0 (212/424)
CLAUDE-3.5-SONNET	22.9 (11/48)	41.9 (195/465)	42.0 (178/424)
GEMINI PRO 1.5	47.9 (23/48)	45.7 (121/265)	51.2 (103/201)
GEMINI FLASH 1.5	31.2 (15/48)	37.4 (99/265)	30.7 (62/202)
COMMAND R	40.0 (6/15)	18.5 (51/275)	19.4 (30/155)
COMMAND R$_{simple}$	33.3 (5/15)	21.5 (59/275)	23.2 (36/155)
COMMAND R+	13.3 (2/15)	18.5 (51/275)	15.5 (24/155)
COMMAND R+$_{simple}$	26.7 (4/15)	14.5 (40/275)	11.0 (17/155)
PHI-3-MINI	10.0 (1/10)	8.4 (12/143)	10.6 (10/94)
PHI-3-MINI$_{simple}$	20.0 (3/15)	13.1 (25/191)	16.0 (20/125)
GEMMA-10M	0.0 (0/63)	3.7 (18/489)	4.7 (21/449)
GEMMA-10M$_{simple}$	4.8 (3/63)	9.8 (48/489)	5.3 (24/449)
LONGLLAMA$_{simple}$	2.1 (1/48)	6.5 (30/465)	3.5 (15/424)
BM25+GPT-4O (k=5)	33.3 (21/63)	28.8 (141/489)	26.7 (120/449)
BM25+GPT-4O (k=25)	44.4 (28/63)	44.6 (218/489)	43.4 (195/449)
BM25+GPT-4O (k=50)	52.4 (33/63)	47.2 (231/489)	51.9 (233/449)

Table 24: Models accuracy on claim pairs by the publication year.

MODEL	BELOW 75K	75K-127K	127K-180K	ABOVE 180K
GPT-4O	59.3 (48/81)	55.2 (296/536)	*n/a*	*n/a*
GPT-4-TURBO	43.2 (35/81)	39.7 (213/536)	*n/a*	*n/a*
CLAUDE-3-OPUS	56.8 (46/81)	48.5 (260/536)	49.1 (157/320)	*n/a*
CLAUDE-3.5-SONNET	34.6 (28/81)	42.4 (227/536)	40.3 (129/320)	*n/a*
GEMINI PRO 1.5	40.0 (32/80)	50.7 (139/274)	47.7 (62/130)	46.7 (14/30)
GEMINI FLASH 1.5	35.8 (29/81)	35.0 (96/274)	33.8 (44/130)	23.3 (7/30)
COMMAND R	25.9 (21/81)	18.1 (66/364)	*n/a*	*n/a*
COMMAND R$_{simple}$	25.9 (21/81)	21.7 (79/364)	*n/a*	*n/a*
COMMAND R+	27.2 (22/81)	15.1 (55/364)	*n/a*	*n/a*
COMMAND R+$_{simple}$	27.2 (22/81)	10.7 (39/364)	*n/a*	*n/a*
LONGLLAMA$_{simple}$	9.9 (8/81)	5.4 (29/536)	2.8 (9/320)	*n/a*
PHI-3-MINI	13.9 (5/36)	8.5 (18/211)	*n/a*	*n/a*
PHI-3-MINI$_{simple}$	13.6 (11/81)	14.8 (37/250)	*n/a*	*n/a*
GEMMA-10M	4.9 (4/81)	4.1 (22/536)	3.4 (11/320)	3.1 (2/64)
GEMMA-10M$_{simple}$	9.9 (8/81)	7.6 (41/536)	6.9 (22/320)	6.2 (4/64)
BM25+GPT-4O (k=5)	29.6 (24/81)	28.0 (150/536)	26.6 (85/320)	35.9 (23/64)
BM25+GPT-4O (k=25)	49.4 (40/81)	42.7 (229/536)	41.6 (133/320)	60.9 (39/64)
BM25+GPT-4O (k=50)	51.9 (42/81)	47.9 (257/536)	48.1 (154/320)	68.8 (44/64)

Table 25: Models accuracy on claim pairs by the book length in tokens (`tiktoken`).

Impact of irrelevant context: As mentioned previously, creating *complex* claims with evidence at different book depths is challenging, especially when aiming to create global claims that require the model to reason over a longer context. Instead, we utilize short stories to investigate whether claims made about the same story are harder to identify when the model is prompted with the entire collec-

MODEL	CORRELATION	p-VALUE
GPT-4O	-0.02	0.9150
GPT-4-TURBO	0.05	0.7722
CLAUDE-3-OPUS	-0.09	0.4682
CLAUDE-3.5-SONNET	0.03	0.8252
GEMINI PRO 1.5	0.03	0.8773
GEMINI FLASH 1.5	-0.13	0.4698
COMMAND R	-0.38	0.0400*
COMMAND R$_{\text{simple}}$	-0.22	0.2329
COMMAND R+	-0.16	0.3865
COMMAND R+$_{\text{simple}}$	-0.53	0.0029**
PHI-3-MINI	-0.57	0.0091**
PHI-3-MINI$_{\text{simple}}$	0.12	0.5972
GEMMA-10M	-0.11	0.3959
GEMMA-10M$_{\text{simple}}$	-0.09	0.4858
LONGLLAMA$_{\text{simple}}$	-0.14	0.2760
BM25+GPT-4O (k=5)	0.11	0.3774
BM25+GPT-4O (k=25)	0.21	0.0849
BM25+GPT-4O (k=25)	0.27	0.0263*

Table 26: Pearson correlation between length and accuracy for different models. Significant correlations are marked with an asterisk.

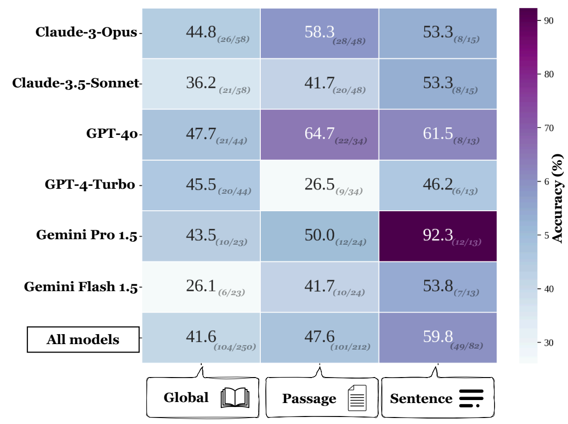


Figure 11: Performance of different closed-source models based on the scope of evidence. The count of correctly identified pairs and the total count of pairs are provided in brackets for reference.

tion versus just the story itself. The accuracies of all models for the whole collection versus individual stories are presented in Table 28. The story-only setup is indicated with a subscript "story." To see these results on common set of claim pairs, that is claim pairs fully processed by all models, see Table 29. Additionally, we report the performance by story location, either the beginning, middle or the end of the collection, relative to the performance when prompted with that story only (see Table 30).[43]

CLAUDE-3-OPUS cites evidence more often than other models: We employ a simple heuristic by looking for quotation marks in the evidence to identify how often models cite excerpts from the source. Qualitatively, we observed that CLAUDE-3-OPUS cites the source more often than other closed-source models, and the citations we verified were always present in the source. This quantitative analysis confirms our observation, as 49.6% of CLAUDE-3-OPUS's responses contain quotation marks, followed by 18.6% for both GPT-4O and GPT-4-TURBO, 11.5% for GEMINI PRO 1.5, and 10.8% for GEMINI FLASH 1.5. We also examined whether responses containing citations are more likely to be correct but did not find such a relation (see Table 31).

Effect of genre: We classify the books into three genres: historical, contemporary, and speculative.

[43]Note that we do not provide an in-depth analysis and discussion of accuracy relative to the story's location due to the small number of annotations at different depths.

MODEL	SENTENCE(correct/total)	PASSAGE(correct/total)	GLOBAL(correct/total)
GPT-4O	61.5 (8/13)	64.7 (22/34)	47.7 (21/44)
GPT-4-TURBO	46.2 (6/13)	26.5 (9/34)	45.5 (20/44)
CLAUDE-3-OPUS	53.3 (8/15)	58.3 (28/48)	44.8 (26/58)
CLAUDE-3.5-SONNET	53.3 (8/15)	41.7 (20/48)	36.2 (21/58)
GEMINI PRO 1.5	92.3 (12/13)	50.0 (12/24)	43.5 (10/23)
GEMINI FLASH 1.5	53.8 (7/13)	41.7 (10/24)	26.1 (6/23)
COMMAND R	25.0 (2/8)	23.1 (6/26)	3.7 (1/27)
COMMAND R_{simple}	37.5 (3/8)	11.5 (3/26)	22.2 (6/27)
COMMAND R+	25.0 (2/8)	19.2 (5/26)	18.5 (5/27)
COMMAND $R+_{simple}$	12.5 (1/8)	7.7 (2/26)	11.1 (3/27)
PHI-3-MINI	0.0 (0/4)	12.5 (2/16)	0.0 (0/16)
PHI-3-$MINI_{simple}$	0.0 (0/6)	15.8 (3/19)	0.0 (0/21)
GEMMA-$10M_{simple}$	6.7 (1/15)	2.1 (1/48)	1.7 (1/58)
GEMMA-10M	0.0 (0/15)	8.3 (4/48)	1.7 (1/58)
$LONGLLAMA_{simple}$	6.7 (1/15)	0.0 (0/48)	3.4 (2/58)
BM25+GPT-4O (k=5)	46.7 (7/15)	22.9 (11/48)	25.9 15/58
BM25+GPT-4O (k=25)	66.7 (10/15)	43.8 (21/48)	29.3 (17/58)
BM25+GPT-4O (k=50)	73.3 (11/15)	45.8 (22/48)	41.4 (24/58)

Table 27: Models' accuracy (%) by the scope of evidence on the annotated subset of data. Counts are provided as subscripts.

Table 32 presents results by genre for all models.

Length of the justifications: Figure 12 provides lengths of the justification provided by each model for the main prompt template (Table 3). We report the lengths in words by accuracy for each model. Note that GEMMA-10M never produces real explanation, rather sometimes repeats part of the original prompt, i.e., `<explanation>YOUR ANSWER</explanation>`.

Length of the retrieved chunks: Figure 13 shows the average percentage of the book retrieved by BM25 for varying values of k, i.e., for top 5, 25, and 50 chunks.

Performance by book: We report the performance of each model by the book title in Figure 14. Empty cells indicate cases where the model did not process the book GEMINI PRO 1.5 and GEMINI FLASH 1.5) or could not process the book because of its context window. See Table 12 for the number of claim pairs written for each book.

Statistical analysis: We conducted statistical analysis by fitting generalized linear mixed-effects models using the `glmer()` function in `R` (Bates et al., 2015).[44] The response variable was pairwise **accuracy**, a binary categorical variable ("correct" or "incorrect"). Pair IDs and annotators were modeled as random effects, with various predictors (fixed effects) included in different models:

1. **Model** - a 6-level categorical variable (GPT-4O, GPT-4-TURBO, CLAUDE-3-OPUS, CLAUDE-3.5-SONNET, GEMINI PRO 1.5, GEMINI FLASH 1.5), with analysis restricted to pairs processed by all models. GPT-4O was set as the reference level (intercept). See Table 33 and Table 34 for the results.;

2. **Length group** - a 4-level categorical variable ("below 75k", "75k-127k", "127k-180k", "above 180k"). "Below 75k" category was set as the reference level (intercept). See Table 35 and Table 36 for the results;

3. **Year** - a 3-level categorical variable ("classics", "2023", "2024") with "classics" set as the reference level. See Table 37 for the results;

[44] All models were fitted using the closed-source models data only, as all open-weights models performed below random.

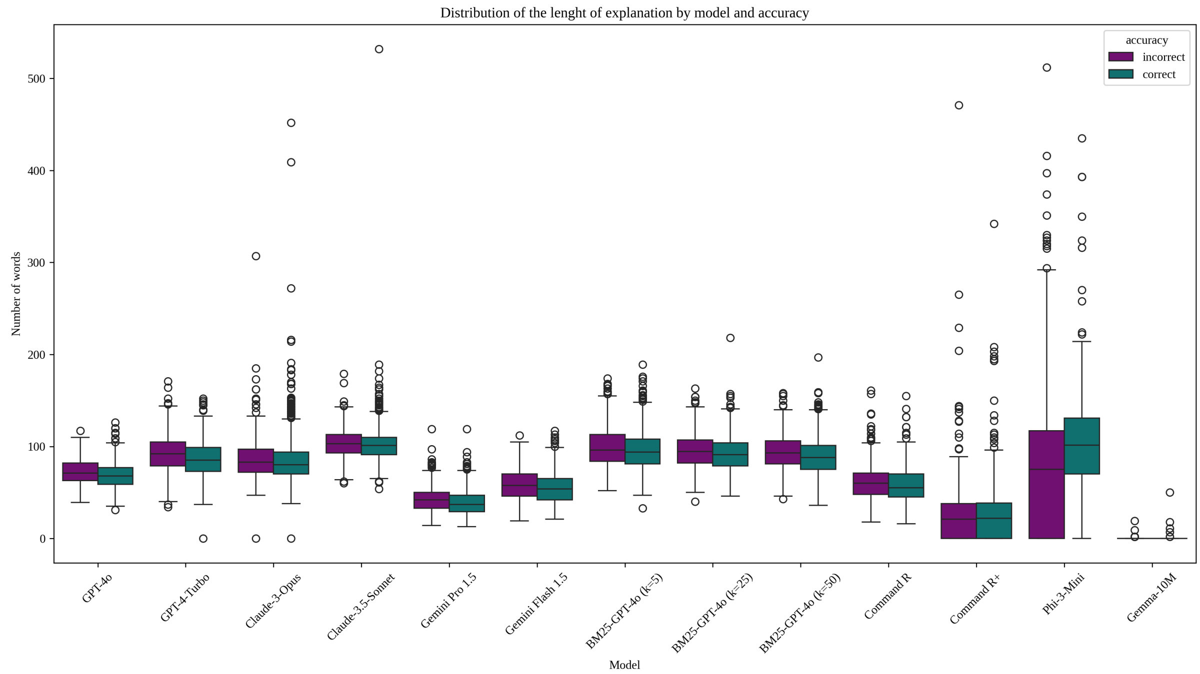


Figure 12: Boxplots of the length of justification (in words) provided for each model for correct (green) and incorrect (purple) predictions.

4. **Genre** - a 3-level categorical variable ("historical", "contemporary", "speculative") with "speculative" set as the reference level. See Table 38 and Table 39 for the results;

5. **Scope** - a 3-level categorical variable ("sentence", "passage", "global") with "sentence" set as the reference level. See Table 40 and Table 41 for the results.

All models were fitted using the `bobyqa` optimizer with a `binomial` link function. We chose mixed-effects models for two main reasons: (1) to account for repeated measures, as each model validates multiple pairs and each pair is validated by multiple models, and (2) to flexibly model pair IDs and annotators as random variables, partially controlling for the inherent difficulty of the pairs unrelated to the predictors.

For these mixed-effects models, we report two types of R^2 values:

- *Marginal* R^2, which indicates the proportion of variance explained by the fixed effects (predictors) alone.

- *Conditional* R^2, which represents the proportion of variance explained by both the fixed and random effects (Nakagawa et al., 2017).

We further conducted a post-hoc analysis using the `emmeans` package (Lenth, 2023) in R with Tukey adjustments for multiple comparisons. To obtain probabilities, we first converted log-odds to odds ratios by exponentiating the estimates, and then converted the odds ratios to probabilities which are reported in the post-hoc tables.

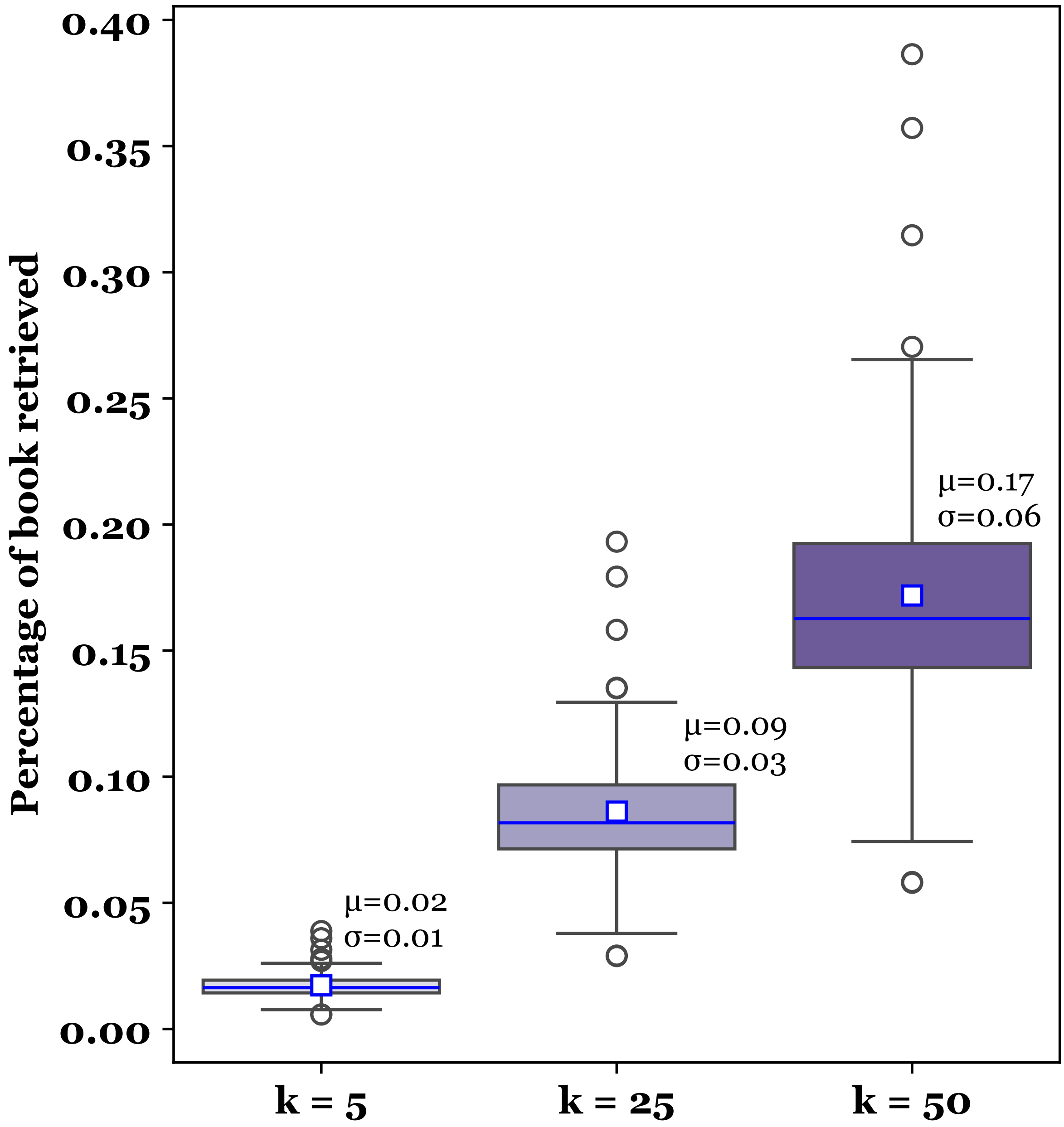


Figure 13: Average percentage of the book retrieved by BM25 for varying values of top k. For example, k = 5 means that the top 5 ranked excerpts according to BM25 were fed to GPT-4o as context for a claim.

Model	ACC (Pair)$_{\text{(correct/total)}}$	ACC (True)$_{\text{(correct/total)}}$	ACC (False)$_{\text{(correct/total)}}$
GPT-4o	56.9 (29/51)	82.4 (42/51)	72.5 (37/51)
GPT-4-Turbo	39.2 (20/51)	58.8 (30/51)	76.5 (39/51)
Claude-3-Opus	37.7 (26/69)	75.4 (52/69)	62.3 (43/69)
Claude-3.5-Sonnet	39.1 (27/69)	52.2 (36/69)	84.1 (58/69)
Gemini Pro 1.5	52.3 (23/44)	68.2 (30/44)	84.1 (37/44)
Gemini Flash 1.5	40.9 (18/44)	61.4 (27/44)	77.3 (34/44)
GPT-4o$_{\text{story}}$	65.2 (45/69)	76.8 (53/69)	88.4 (61/69)
GPT-4-Turbo$_{\text{story}}$	50.7 (35/69)	59.4 (41/69)	87.0 (60/69)
Claude-3-Opus$_{\text{story}}$	63.8 (44/69)	79.7 (55/69)	81.2 (56/69)
Claude-3.5-Sonnet$_{\text{story}}$	50.7 (35/69)	62.3 (43/69)	88.4 (61/69)
Gemini Pro 1.5$_{\text{story}}$	51.5 (35/68)	58.8 (40/68)	77.9 (53/68)
Gemini Flash 1.5$_{\text{story}}$	42.0 (29/69)	50.7 (35/69)	75.4 (52/69)
Command R+	11.8 (6/51)	68.6 (35/51)	31.4 (16/51)
Command R+$_{\text{simple}}$	19.6 (10/51)	74.5 (38/51)	39.2 (20/51)
Command R	23.5 (12/51)	80.4 (41/51)	39.2 (20/51)
Command R$_{\text{simple}}$	31.4 (16/51)	82.4 (42/51)	45.1 (23/51)
Phi-3-mini	7.9 (3/38)	45.0 (18/40)	35.9 (14/39)
Phi-3-mini$_{\text{simple}}$	9.8 (5/51)	80.4 (41/51)	29.4 (15/51)
Command R$_{\text{story}}$	39.1 (27/69)	81.2 (56/69)	53.6 (37/69)
Command R+$_{\text{story}}$	43.5 (30/69)	81.2 (56/69)	59.4 (41/69)
Phi-3-mini$_{\text{story}}$	15.4 (10/65)	26.9 (18/67)	83.3 (55/66)
Phi-3-mini$_{\text{simple-story}}$	24.2 (16/66)	62.1 (41/66)	56.7 (38/67)
Mixtral-8x22B$_{\text{story}}$	56.6 (39/69)	75.4 (52/69)	81.2 (56/69)
Qwen-2-72B$_{\text{story}}$	60.9 (42/69)	79.7 (55/69)	81.2 (56/69)
BM25+GPT-4o (k=5)	37.7 (26/69)	44.9 (31/69)	91.3 (63/69)
BM25+GPT-4o (k=25)	53.6 (37/69)	62.3 (43/69)	85.5 (59/69)
BM25+GPT-4o (k=50)	56.5 (39/69)	71.0 (49/69)	82.6 (57/69)
Random	25.0 (17/69)	50.0 (35/69)	50.0 (35/69)

Table 28: **Pairwise** accuracy, accuracy on **True**, and accuracy on **False** for claims made about collections of stories (80k-129k tokens) versus individual stories from the collection (700-21k tokens, average 8.5k). All texts fitting within the model's context window were listed. For the 128k token models, the accuracy for the entire collection excludes one book with 129k tokens, which is included in the accuracy for individual stories, as none of the stories exceeded the context window. In the case of Gemini Pro 1.5 and Gemini Flash 1.5, two collections were not processed due to being identified as copyrighted content ("prohibited content" API error). However, both models processed the stories from these collections, likely due to the length difference, though some claims were still refused due to disruptive content. We also provide accuracy on stories for Mixtral-8x22B and Qwen-2-72B, which have context window of 65k and 32k respectively, for comparison. Subscript "story" denotes results obtained with prompting the models with individual stories as the context.

MODEL	ACC PAIR(correct/total)	ACC TRUE(correct/total)	ACC FALSE(correct/total)
GPT-4O	48.0 (12/25)	76.0(19/25)	68.0$_{17/25}$
GPT-4-TURBO	44.0 (11/25)	64.0$_{16/25}$	76.0$_{19/25}$
CLAUDE-3-OPUS	32.0 (8/25)	88.0$_{22/25}$	44.0$_{11/25}$
CLAUDE-3.5-SONNET	52.0$_{13/25}$	68.0$_{17/25}$	80.0$_{20/25}$
GEMINI PRO 1.5	56.0$_{14/25}$	80.0$_{20/25}$	76.0$_{19/25}$
GEMINI FLASH 1.5	48.0$_{12/25}$	76.0$_{19/25}$	68.0$_{17/25}$
GPT-4O$_{\text{story}}$	72.0$_{18/25}$	84.0$_{21/25}$	88.0$_{22/25}$
GPT-4-TURBO$_{\text{story}}$	56.0$_{14/25}$	68.0$_{17/25}$	84.0$_{21/25}$
CLAUDE-3-OPUS$_{\text{story}}$	72.0$_{18/25}$	84.0$_{21/25}$	84.0$_{21/25}$
CLAUDE-3.5-SONNET$_{\text{story}}$	60.0$_{15/25}$	68.0$_{17/25}$	92.0$_{23/25}$
GEMINI PRO 1.5$_{\text{story}}$	64.0$_{16/25}$	72.0$_{18/25}$	88.0$_{22/25}$
GEMINI FLASH 1.5$_{\text{story}}$	48.0$_{12/25}$	68.0$_{17/25}$	76.0$_{19/25}$
MIXTRAL-8X22B$_{\text{story}}$	64.0$_{16/25}$	80.0$_{20/25}$	84.0$_{21/25}$
QWEN-2-72B$_{\text{story}}$	64.0$_{16/25}$	80.0$_{20/25}$	84.0$_{21/25}$
BM25+GPT-4O (k=5)	52.0$_{13/25}$	64.0$_{16/25}$	88.0$_{22/25}$
BM25+GPT-4O (k=5)	68.0$_{17/25}$	76.0$_{19/25}$	92.0$_{23/25}$
BM25+GPT-4O (k=5)	72.0$_{18/25}$	88.0$_{22/25}$	84.0$_{21/25}$

Table 29: **Pairwise** accuracy, accuracy on **True**, and accuracy on **False** for claims made about collections of stories (80k-129k tokens) versus individual stories from the collection (700-21k tokens, average 8.5k) on **common set** of claim pairs (i.e., pairs which were processed by all the models in all shown configurations). We also provide accuracies for MIXTRAL-8X22B and QWEN-2-72B on the same set of claims. These models were prompted with individual stories as the context which makes their results comparable with other models marked with subscript "story."

Model	Story Location within Collection		
	Beginning	Middle	End
GPT-4o	$64.0_{(16/25)}$	$50.0_{(7/14)}$	$50.0_{(6/12)}$
GPT-4o$_{\text{story}}$	$76.0_{(19/25)}$	$64.3_{(9/14)}$	$66.7_{(8/12)}$
GPT-4-Turbo	$36.0_{(9/25)}$	$35.7_{(5/14)}$	$50.0_{(6/12)}$
GPT-4-Turbo$_{\text{story}}$	$60.0_{(15/25)}$	$42.9_{(6/14)}$	$58.3_{(7/12)}$
Claude-3-Opus	$38.2_{(13/34)}$	$31.6_{(6/19)}$	$43.8_{(7/16)}$
Claude-3-Opus$_{\text{story}}$	$73.5_{(25/34)}$	$68.4_{(13/19)}$	$37.5_{(6/16)}$
Claude-3.5-Sonnet	$38.2_{(13/34)}$	$31.6_{(6/19)}$	$50.0_{(8/16)}$
Claude-3.5-Sonnet$_{\text{story}}$	$61.8_{(21/34)}$	$47.4_{(9/19)}$	$31.2_{(5/16)}$
Gemini Pro 1.5	$52.2_{(12/23)}$	$70.0_{(7/10)}$	$40.0_{(4/10)}$
Gemini Pro 1.5$_{\text{story}}$	$56.5_{(13/23)}$	$70.0_{(7/10)}$	$40.0_{(4/10)}$
Gemini Flash 1.5	$39.1_{(9/23)}$	$27.3_{(3/11)}$	$60.0_{(6/10)}$
Gemini Flash 1.5$_{\text{story}}$	$34.8_{(8/23)}$	$63.6_{(7/11)}$	$30.0_{(3/10)}$
Command R	$20.0_{(5/25)}$	$21.4_{(3/14)}$	$33.3_{(4/12)}$
Command R$_{\text{story}}$	$36.0_{(9/25)}$	$35.7_{(5/14)}$	$33.3_{(4/12)}$
Command R+	$20.0_{(5/25)}$	$0.0_{(0/14)}$	$8.3_{(1/12)}$
Command R+$_{\text{story}}$	$48.0_{(12/25)}$	$50.0_{(7/14)}$	$41.7_{(5/12)}$
Phi-3-mini	$6.2_{(1/16)}$	$8.3_{(1/12)}$	$14.3_{(1/7)}$
Phi-3-mini$_{\text{story}}$	$18.8_{(3/16)}$	$33.3_{(4/12)}$	$0.0_{(0/7)}$
Phi-3-mini$_{\text{simple}}$	$12.5_{(3/24)}$	$15.4_{(2/13)}$	$0.0_{(0/12)}$
Phi-3-mini$_{\text{simple-story}}$	$37.5_{(9/24)}$	$30.8_{(4/13)}$	$25.0_{(3/12)}$
Mixtral-8x22B$_{\text{story}}$	$61.8_{(21/34)}$	$52.6_{(10/19)}$	$50.0_{(8/16)}$
Qwen-2-72B$_{\text{story}}$	$61.8_{(21/34)}$	$63.2_{(12/19)}$	$56.2_{(9/16)}$

Table 30: Models' performance on stories at different depths of the story collection. Each collection is divided into three parts based on the number of tokens: (1) beginning (first third), (2) middle (second third), and (3) end (last third). The subscript "story" (marked with 🗎) refers to outputs obtained by prompting with the story (rather than the collection) as the context, for comparison. We also provide results for Mixtral-8x22B and Qwen-2-72B for comparison.

Model	Quotations (%)		
	All responses	Correct responses	Incorrect responses
GPT-4o	18.4% 227/1234	18.8% 178/946	17.0% 49/288
GPT-4-Turbo	18.8% 232/1234	19.5% 164/839	17.2% 68/395
Claude-3-Opus	49.8% 934/1874	49.8% 685/1376	50.0% 249/498
Claude-3.5-Sonnet	59.6% 1116/1874	62.2% 808/1300	53.7% 308/574
Gemini Pro 1.5	11.6% 119/1029	11.4% 84/738	12.0% 35/291
Gemini Flash 1.5	10.8% 111/1030	10.5% 70/668	11.3% 41/362

Table 31: Percentage of responses with identified quotations by model. Separate values are reported for claims labeled correctly and incorrectly. Note that the percentages are reported by claim, as each explanation is generated at the claim level. The counts (quotations/total) are provided in subscript. We do not report these numbers for open-weights models as the generations often do not follow the requested output format.

	GENRE		
MODEL	HISTORICAL	CONTEMPORARY	SPECULATIVE
GPT-4O	$70.3_{(26/37)}$	$59.0_{(229/388)}$	$44.2_{(72/163)}$
GPT-4-TURBO	$70.3_{(26/37)}$	$39.9_{(155/388)}$	$34.4_{(56/163)}$
CLAUDE-3-OPUS	$63.5_{(33/52)}$	$51.1_{(290/567)}$	$42.9_{(124/289)}$
CLAUDE-3.5-SONNET	$42.3_{(22/52)}$	$42.2_{(239/567)}$	$38.1_{(110/289)}$
GEMINI PRO 1.5	$53.8_{(28/52)}$	$49.2_{(127/258)}$	$44.4_{(84/189)}$
GEMINI FLASH 1.5	$46.2_{(24/52)}$	$37.2_{(96/258)}$	$27.4_{(52/190)}$
COMMAND R	$30.3_{(10/33)}$	$17.8_{(46/259)}$	$20.1_{(28/139)}$
COMMAND R_{simple}	$21.2_{(7/33)}$	$22.8_{(59/259)}$	$21.6_{(30/139)}$
COMMAND R+	$18.2_{(6/33)}$	$21.6_{(56/259)}$	$10.1_{(14/139)}$
COMMAND $R+_{simple}$	$15.2_{(5/33)}$	$13.1_{(34/259)}$	$15.1_{(21/139)}$
PHI-3-MINI	$10.0_{(1/10)}$	$9.5_{(14/147)}$	$9.0_{(8/89)}$
PHI-3-$MINI_{simple}$	$20.0_{(3/15)}$	$15.8_{(29/183)}$	$12.6_{(15/119)}$
GEMMA-10M	$1.5_{(1/67)}$	$3.2_{(18/567)}$	$4.4_{(15/338)}$
GEMMA-$10M_{simple}$	$4.5_{(3/67)}$	$7.2_{(41/567)}$	$7.7_{(26/338)}$
$LONGLLAMA_{simple}$	$0.0_{(0/52)}$	$5.1_{(29/567)}$	$4.2_{(12/289)}$
BM25+GPT-4O (k=5)	$35.8_{(24/67)}$	$28.6_{(162/567)}$	$26.0_{(88/338)}$
BM25+GPT-4O (k=25)	$50.7_{(34/67)}$	$42.9_{(243/567)}$	$44.4_{(150/338)}$
BM25+GPT-4O (k=50)	$56.7_{(38/67)}$	$50.8_{(288/567)}$	$45.9_{(155/338)}$

Table 32: Model performance by genre: *historical* (pre-WWII), *contemporary* (post-WWII), and *speculative* (fantasy/SF/ghosts).

`model <- glmer(accuracy ~ model + (1|id) + (1|annotator), common_set_data)`				
PREDICTORS	ODDS RATIOS	CI (95%)	p-VALUE	
(Intercept)	1.73	1.18 – 2.55	0.005	**
GPT-4-TURBO	0.33	0.23 – 0.48	<0.001	***
CLAUDE-3-OPUS	0.64	0.45 – 0.91	0.014	*
CLAUDE-3.5-SONNET	0.34	0.24 – 0.50	<0.001	***
GEMINI PRO 1.5	0.55	0.38 – 0.78	0.001	***
GEMINI FLASH 1.5	0.24	0.17 – 0.35	<0.001	***
RANDOM EFFECTS				
σ^2 (*residual variance*)	3.29			
τ_{00} (id) (*variance of random intercepts*)	2.54			
τ_{00} (annotator) (*variance of random intercepts*)	0.19			
ICC	0.45			
N (ID)	354			
N (ANNOTATOR)	15			
OBSERVATIONS	2124			
R^2 (*marginal*)	0.036			
R^2 (*conditional*)	0.473			

Table 33: Summary of generalized linear mixed model with **model** as the predictor of **accuracy**: `model <- glmer(accuracy ~ model + (1|id) + (1|annotator), data)`. GPT-4O was set as the reference level (intercept). Note that while we observe significant differences between the models' performance, the marginal R^2 is low, suggesting that model type alone does not explain the majority of the variance in the data. See Table 34 for post-hoc analysis.

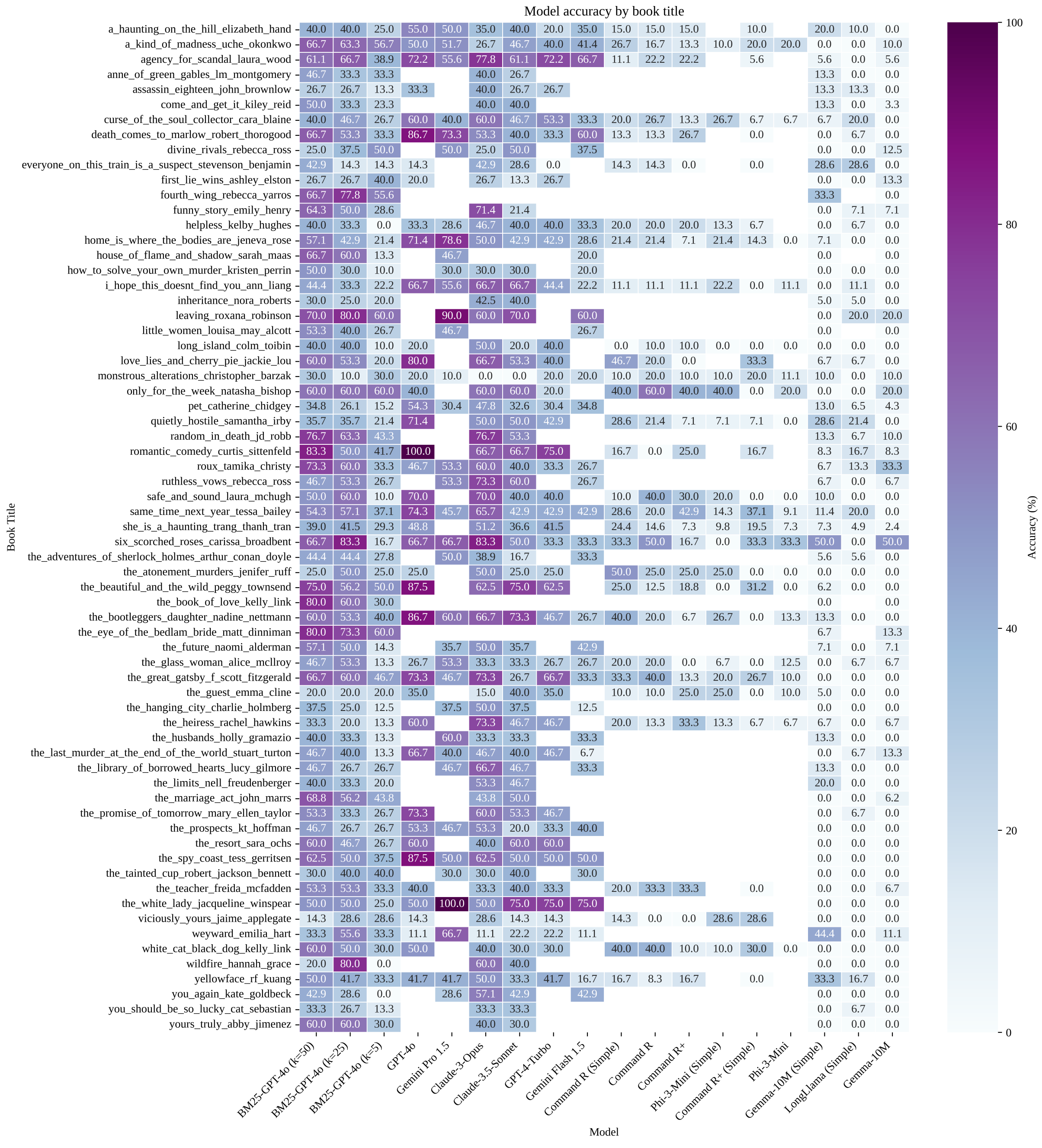


Figure 14: Heatmap of model performance by book title. Empty cells indicate cases where the model did not process the book (GEMINI PRO 1.5 and GEMINI FLASH 1.5) or could not process the book because of its context window.

Contrast	Estimate	SE	Odds Ratio	Probability	p-value	
GPT-4o - Claude-3-Opus	0.4480	0.1820	1.5652	0.6102	0.1355	
GPT-4o - Claude-3.5-Sonnet	1.0654	0.1849	2.9021	0.7437	<0.001	***
GPT-4o - Gemini Flash 1.5	1.4229	0.1888	4.1492	0.8058	<0.001	***
GPT-4o - Gemini Pro 1.5	0.6013	0.1822	1.8244	0.6459	0.0124	*
GPT-4o - GPT-4-Turbo	1.1005	0.1852	3.0056	0.7503	<0.001	***
Claude-3-Opus - Claude-3-Opus	0.6174	0.1820	1.8541	0.6496	0.0090	**
Claude-3-Opus - Gemini Flash 1.5	0.9749	0.1855	2.6509	0.7261	<0.001	***
Claude-3-Opus - Gemini Pro 1.5	0.1533	0.1801	1.1656	0.5382	0.9578	
Claude-3-Opus - GPT-4-Turbo	0.6524	0.1822	1.9202	0.6576	0.0046	**
Claude-3.5-Sonnet - Gemini Flash 1.5	0.3575	0.1852	1.4297	0.5884	0.3833	
Claude-3.5-Sonnet - Gemini Pro 1.5	-0.4642	0.1816	0.6287	0.3860	0.1084	
Claude-3.5-Sonnet - GPT-4-Turbo	0.0350	0.1826	1.0356	0.5088	1.0000	
Gemini Flash 1.5 - Gemini Pro 1.5	-0.8216	0.1849	0.4397	0.3054	<0.001	***
Gemini Flash 1.5 - GPT-4-Turbo	-0.3224	0.1853	0.7244	0.4201	0.5050	
Gemini Pro 1.5 - GPT-4-Turbo	0.4992	0.1819	1.6474	0.6223	0.0667	.

Table 34: Post-hoc comparisons of models for **accuracy** (Table 33) using Tukey adjustments for multiple comparisons. The probability values refer to the likelihood that the first model in each contrast is more accurate than the second model (i.e., a value of 0.5 suggests that both models are comparable in terms of accuracy).

`model <- glmer(accuracy ~ length_group + (1|pairID) + (1|annotator), data)`

Predictors	Odds Ratios	CI (95%)	p-value	
(Intercept)	0.56	0.34 – 0.94	0.028	*
75k–127k	1.45	0.87 – 2.43	0.154	
127k–180k	1.26	0.74 – 2.14	0.391	
above 180k	0.75	0.28 – 1.96	0.551	
Random Effects				
σ^2 (*residual variance*)	3.29			
τ_{00} (id) (*variance of random intercepts*)	2.15			
τ_{00} (annotator) (*variance of random intercepts*)	0.24			
ICC	0.42			
N (id)	967			
N (annotator)	22			
Observations	4137			
R^2 (*marginal*)	0.004			
R^2 (*conditional*)	0.423			

Table 35: Summary of generalized linear mixed model with **length group** as the predictor of **accuracy**: `model <- glmer(accuracy ~ length_group + (1|pairID) + (1|annotator), data)`. The "below 75k" group was set as the reference level (intercept). Note that the marginal R^2 is very low, indicating that length group alone does not explain the majority of the variance in the data. See Table 36 for post-hoc analysis.

Contrast	Estimate	SE	Odds Ratio	Probability	p-value
Below 75k - 127k-180k	-0.2312	0.2698	0.7936	0.4424	0.8267
Below 75k - 75k-127k	-0.3739	0.2620	0.6880	0.4076	0.4823
Below 75k - Above 180k	0.2938	0.4925	1.3416	0.5729	0.9331
127k-180k - 75k-127k	-0.1427	0.1652	0.8670	0.4644	0.8235
127k-180k - Above 180k	0.5251	0.4461	1.6906	0.6283	0.6415
75k-127k - Above 180k	0.6677	0.4464	1.9498	0.6610	0.4399

Table 36: Post-hoc comparisons of length groups for **accuracy** (Table 35) using Tukey adjustments for multiple comparisons. The probability values refer to the likelihood that the verification for the first group in each contrast is more accurate than the verification for the second group.

`model <- glmer(accuracy ~ year + (1|pairID) + (1|annotator), data)`			
PREDICTORS	ODDS RATIOS	CI (95%)	*p*-VALUE
(Intercept)	0.65	0.35 – 1.20	0.170
2023	1.17	0.64 – 2.15	0.321
2024	1.17	0.63 – 2.16	0.344
RANDOM EFFECTS			
σ^2 (*residual variance*)	3.29		
τ_{00} (id) (*variance of random intercepts*)	2.16		
τ_{00} (annotator) (*variance of random intercepts*)	0.22		
ICC	0.42		
N (ID)	967		
N (ANNOTATOR)	22		
OBSERVATIONS	4137		
R^2 (*marginal*)	0.000		
R^2 (*conditional*)	0.420		

Table 37: Summary of generalized linear mixed model with **year** as the predictor of **accuracy**: `model <- glmer(accuracy ~ year + (1|pairID) + (1|annotator), data)`. The "classics" year group was set as the reference level (intercept). Note that the marginal R^2 is 0.000, indicating that year alone does not explain the majority of the variance in the data.

`model <- glmer(accuracy ~ genre + (1|id) + (1|annotator), data = filtered_data)`				
PREDICTORS	ODDS RATIOS	CI (95%)	*p*-VALUE	
(Intercept)	0.55	0.40 – 0.74	<0.001	***
CONTEMPORARY	1.53	1.11 – 2.11	0.010	*
HISTORICAL	2.18	1.20 – 3.96	0.010	*
RANDOM EFFECTS				
σ^2 (*residual variance*)	3.29			
τ_{00} (id) (*variance of random intercepts*)	2.10			
τ_{00} (annotator) (*variance of random intercepts*)	0.16			
ICC	0.41			
N (ID)	938			
N (ANNOTATOR)	22			
OBSERVATIONS	3991			
R^2 (*marginal*)	0.010			
R^2 (*conditional*)	0.412			

Table 38: Summary of generalized linear mixed model with **genre** as the predictor of **accuracy**: `model <- glmer(accuracy ~ genre + (1|id) + (1|annotator), data = filtered_data)`. The "speculative" genre was set as the reference level (intercept). We also excluded two books which felt into "historical and contemporary" and "essays" categories. Note that the marginal R^2 is low, indicating that genre alone does not explain the majority of the variance in the data. The post-hoc analysis for this model is presented in Table 39.

Contrast	Estimate	SE	Odds Ratio	Probability	p-value	
Speculative - Contemporary	-0.4234	0.1648	0.6548	0.3957	0.0275	*
Speculative - Historical	-0.7802	0.3042	0.4583	0.3143	0.0278	*
Contemporary - Historical	-0.3568	0.2874	0.6999	0.4117	0.4287	

Table 39: Post-hoc comparisons of genres for **accuracy** (Table 38) using Tukey adjustments for multiple comparisons. The probability values refer to the likelihood that the validation of books of the first genre in each contrast is more accurate than the validation of books of the second genre (i.e., a value of 0.5 suggests that both genres are comparable in terms of accuracy).

`model <- glmer(accuracy ~ scope + (1|id) + (1|annotator), data = filtered_data)`				
Predictors	Odds Ratios	CI (95%)	p-value	
(Intercept)	2.12	0.74 – 6.11	0.163	
Passage	0.35	0.14 – 0.87	0.024	*
Global	0.32	0.13 – 0.78	0.012	*
Random Effects				
σ^2 (*residual variance*)	3.29			
τ_{00} (id) (*variance of random intercepts*)	1.19			
τ_{00} (annotator) (*variance of random intercepts*)	0.53			
ICC	0.34			
N (id)	121			
N (annotator)	4			
Observations	544			
R^2 (*marginal*)	0.030			
R^2 (*conditional*)	0.363			

Table 40: Summary of generalized linear mixed model with **scope** as the predictor of **accuracy**: `model <- glmer(accuracy ~ scope + (1|id) + (1|annotator), data = filtered_data)`. The "sentence" scope was set as the reference level (intercept). Note that the marginal R^2 is low, indicating that scope alone does not explain the majority of the variance in the data. See Table 41 for the post-hoc analysis.

Contrast	Estimate	SE	Odds Ratio	Probability	p-value	
Sentence - Global	1.1322	0.4506	3.1026	0.7563	0.0321	*
Sentence - Passage	1.0488	0.4642	2.8542	0.7405	0.0617	.
Global - Passage	-0.0834	0.3139	0.9199	0.4792	0.9618	

Table 41: Post-hoc comparisons of scopes for **accuracy** (Table 40) using Tukey adjustments for multiple comparisons. The probability values refer to the likelihood that the validation of claim pairs in the first scope in each contrast is more accurate than the validation of claim pairs in the second scope (i.e., a value of 0.5 suggests that both scopes are comparable in terms of accuracy).