PRELUDE: A Benchmark Designed to Require Global Comprehension and Reasoning over Long Contexts

Anonymous ACL submission

Abstract

We introduce PRELUDE, a benchmark for evaluating long-context understanding through the task of determining whether a character's prequel story is consistent with the canonical narrative of the original book. Our task poses a stronger demand for global comprehension and deep reasoning than existing benchmarks – as the prequels are not part of the original story, assessing their plausibility typically requires searching and integrating information that is only indirectly related. Empirically, 88% of instances require evidence from multiple parts of the narrative. Experimental results highlight the challenge of our task: in-context learning, RAG and in-domain training with state-of-the-art LLMs, and commercial DeepResearch services, lag behind humans by >15%. A further human study reveals that models often produce correct answers with flawed reasoning, leading to an over 30% gap in reasoning accuracy compared to humans. These findings underscore the substantial room for improvement in long-context understanding and reasoning.

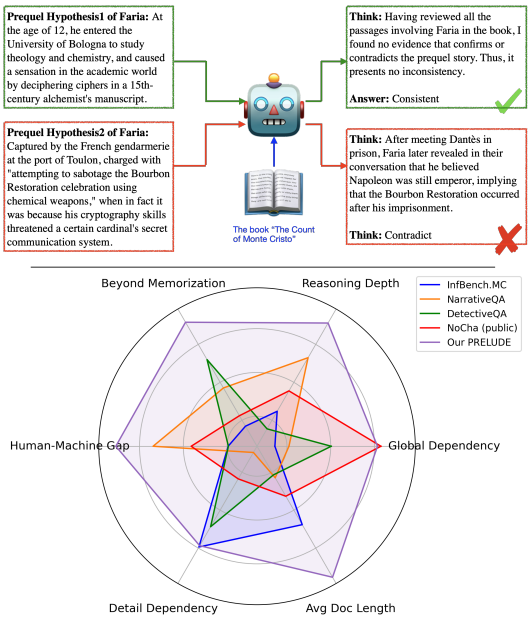

Figure 1: (Top) Two examples from our PRELUDE. (Bottom) Comparison of existing benchmarks along the different criteria for long context understanding assessment. We report the normalized measure along each criterion, with the details presented in Appendix D.

1 Introduction

The emergence of new LLM-driven applications, such as multi-document analysis (Google DeepMind, 2024; Wang et al., 2024d; Gutierrez et al., 2024), personal assistants with chat histories (Wu et al., 2024; Xu et al., 2025a), autonomous agents (Hu et al., 2025; OpenAI, 2025; Wang et al., 2025), and repository-level coding tools (Jimenez et al., 2024), has created increasing demands for robust long-context understanding and reasoning.

To better support long inputs, many techniques have been proposed, primarily focusing on efficient attention mechanisms (Xiong et al., 2021, 2023) and retrieval-augmented generation (RAG) (Lewis et al., 2020; Xu et al., 2024a; Edge et al., 2024; Gutierrez et al., 2024; Asai et al., 2023). Alongside these technical advances, there is a growing need for effectively evaluating long context understanding and reasoning capabilities. To this end, several benchmarks have recently been introduced (see Section 2). Building on this progress, recent work has extensively discussed the criteria that a strong benchmark for long-context understanding and reasoning should satisfy (Press et al., 2023; Yu et al., 2023; Liu et al., 2024; Yen et al., 2024; Fang et al., 2024; Wu et al., 2025b). To rigorously evaluate a model's capabilities in this domain, several key criteria have emerged as essential:

- **Beyond Memorization.** LLMs memorize content from pretraining, especially for popular texts (Tirumala et al., 2022; Delétang et al., 2023; Sutskever, 2023), enabling answers without true comprehension. The existence of this shortcuts blurs the line between long-context understanding and mere activation of parametric knowledge memorized during pretraining. As training data

grows, this issue worsens. As a *Necessary Condition*, a robust benchmark must prevent solutions based on memorization alone, ensuring full-context reasoning remains essential.

- **Global Dependency.** The task should require aggregating evidence scattered across the context or exhibiting global dependencies; otherwise, it reduces to a short-context problem focused on retrieval rather than true long-text understanding.
- **Depth of Reasoning.** Long-context reasoning should inherently require synthesizing multiple pieces of evidence and multi-step deduction. However, many existing benchmarks focus on shallow reasoning, such as decomposition or enumeration (*e.g.*, counting facts (Xu and Ma, 2025), or sequential sub-questions (Yang et al., 2018)), which reduces the need for global, multi-step inference and oversimplifies the task.
- **Human-Machine Gap.** To highlight essential capabilities that general-purpose intelligent systems should possess, a benchmark should show a significant gap between humans and machines. Low human accuracy or agreement typically signals poor annotation quality or high subjectivity, undermining reliable evaluation.
- **Beyond Summarization/Salience.** Often overlooked but crucial, a strong benchmark should require attention to fine-grained details beyond high-level abstraction to remain challenging and meaningful. Otherwise, it risks reducing to a summarization task that is solvable without long-context understanding (Chang et al., 2023).

Figure 1 evaluates several widely-used benchmarks for story understanding across the aforementioned dimensions. It shows that all existing benchmarks fall short in at least one aspect, with a particularly notable limitation in mitigating the effects of *Memorization* and encouraging *Deep Reasoning*. Detailed evaluation metrics and experimental settings are provided in Appendix D.

To address these limitations, we propose a novel task format that enables comprehensive assessment across all the identified criteria. The task involves presenting the model with a list of hypothetical prequels for important supporting book characters, and asking it to determine whether each prequel is consistent with the canonical story. Each hypothesis is presented as a concise bullet point (see Figure 1) summarizing a proposed setting. These hypotheses are annotated by human experts who have read the canonical stories multiple times and performed literary analysis before. This annotation process is both efficient and high-quality, ultimately yielding a dataset of $\sim$1K labeled examples. Intuitively, our task design naturally mitigates the key limitations observed in existing benchmarks:

First, the *Memorization* shortcut is alleviated by construction, as the input prequels are newly generated and do not appear in the training data of any existing LLMs. The limited performance of OpenAI's Deep Research further supports this, suggesting that it is difficult to locate human-summarized evidence on the internet to solve our task.

Second, our task encourages *global reasoning*. This is because (1) determining whether a consistent prequel aligns with the canonical story typically requires aggregating evidence across the whole character story; and (2) contradictory prequels often involve inconsistencies that span several scattered events due to the narrative structure of the original work. Empirically, our annotation analysis reveals that 88% of the examples in `PRELUDE` require non-local evidence to resolve.

Finally, our task encourages *deep reasoning*, because of the fact that the canonical story reflects non-immediate consequences of the prequels. To solve our task, LLMs must unfold the implications of a prequel and align them with the story, often requiring multi-step inference. For instance, the second example in Figure 1 involves reasoning that *Faria was arrested when Napoleon was still emperor*, and then inferring a contradiction from the fact that *the Bourbon Restoration removed Napoleon from power*. This kind of non-immediate causality resists shallow reasoning shortcuts that decomposes the problem into subquestions.

We conduct extensive experiments on our proposed task using ICL, RAG, and DeepResearch across state-of-the-art commercial and open-source LLMs. The results reveal several key findings: (1) The best-performing system lags human performance by over 15%; (2) LLMs often arrive at correct predictions with flawed reasoning, resulting in a reasoning accuracy gap of $>$30% compared to humans; (3) Our task cannot be solved by searching for information on the web, making the advanced DeepResearch systems struggle and underperform RAG; (4) Supervised training and many-shot ICL yield no performance gains, highlighting LLMs' intrinsic limitations in long-context reasoning.

These findings highlight that `PRELUDE` requires deeper long-context reasoning capabilities beyond what current methods offer, pointing to important directions for future research.

Type	Definition	Example
Contradict – Local	A detail in the original novel directly contradicts the prequel	**Book:** The Chronicles of Narnia **Char:** Eustace Scrubb **Prequel:** Born in London and raised as an only child, Eustace was taught to call his parents by their names —"Harold" and "Alberta". *The novel introduces Eustace as being born in Cambridge, England.*
Contradict – Global I	The prequel causes part of the original plot or the character's motivations to become broadly unreasonable.	**Book:** The Count of Monte Cristo **Char:** Faria **Prequel:** He joined the Jesuits and was sent to Goa, Indian. During this time, he secretly studied the ancient Indian medical text "Charaka Samhita," which laid the groundwork for his later expertise in toxicology. *(Throughout the story, Faria never demonstrates any expertise in medicine or toxicology. Moreover, he does not speak Hindi, making such an event unlikely.)*
Contradict – Global II	The prequel does not contradict the original plot but is inconsistent with the tone or stylistic setting of the novel	**Book:** Romance of the Three Kingdoms **Char:** Xiahou Yuan **Prequel:** He was taught magical speed-enhancing techniques by a mysterious hermit, which allowed him to march 1000 li in six days. *(The book is a historical novel with a largely realistic style, and it never depicts any general using magic or supernatural powers.)*
Consistent – Irrelevant	The prequel is not closely relevant to the original plot but doesn't contradict it. It is a side story of the character.	**Book:** The Count of Monte Cristo **Char:** Faria **Prequel:** At the age of 12, he entered the University of Bologna to study theology and chemistry, and caused a sensation in the academic world by deciphering the ciphers in a 15th-century alchemist's manuscript. *(Faria's academic background and influence are never depicted in the novel. However, this is consistent with his characterization as a learned man.)*
Consistent – Core	The prequel fills in missing detail without introducing contradictions.	**Book:** Harry Potter **Char:** Sirius Black **Prequel:** Rejecting pure-blood elitism, Sirius frequently showed admiration for Gryffindor and spent time with Muggles and so-called "blood traitors" to annoy his parents even before attending Hogwarts. *(Sirius is introduced as a key figure against pure-blood elitism, but little detail is provided. This fills in gaps without introducing contradictions.)*

Table 1: Definitions of our annotation labels in PRELUDE.

2 Related Work

Tasks over Synthetic Long Contexts Following the needle-in-a-haystack task that examines LLM's in-context searching (Kamradt, 2023), a line of works have since focused on LLMs' ability in tracing and utilizing information pieces in stretched long context (Hsieh et al., 2024; Li et al., 2025b; Yu et al., 2025b), while others also fuse reasoning in their task design such as sorting or relation identification, beyond mere retrieval (Kuratov et al., 2024; Wang et al., 2024a; Dong et al., 2024; Wang et al., 2024c; Lee et al., 2025). Besides these works that specifically stress-test long context utilization, other related LLM tasks could also reflect such ability, *e.g.*, many-shot in-context learning (Agarwal et al., 2024; Xu et al., 2024b; Li et al., 2025c).

Realistic Long Context Understanding Tasks Orthogonal to those synthetic stress-testing tasks, another line of works target the more natural question answering settings for realistic long context evaluation, primarily utilizing stories in various domains, such as NarrativeQA (Kočiskỳ et al., 2018), NovelQA (Wang et al., 2024b), DetectiveQA (Xu et al., 2025b), CharToM-QA (Zhou et al., 2025). Realistic long context QA has gained particular attention in many recent LLM benchmarks, such as LongBench (Bai et al., 2024a,b), XLBench (Ni et al., 2024), CLongEval (Qiu et al., 2024), LooGLE (Li et al., 2024), ∞ Bench (Zhang et al., 2024), LaRA (Li et al., 2025a), *etc*.

Document-Level Entailment Our task is closely related to fact verification over multiple documents or web sources, as exemplified by FEVER (Thorne et al., 2018) and its extensions (Wadden et al., 2020; Yin et al., 2021; Schlichtkrull et al., 2023).

Among this line of work, NoCha (Karpinska et al., 2024) is the most relevant, as it also uses book narratives as context. However, a key distinction lies in the nature of the hypotheses: NoCha uses summaries or conclusions of the original story, which often share semantic overlap with the canonical book. Therefore, this task design is vulnerable to memorization or summarization shortcuts, as shown in Figure 1. To mitigate this, NoCha uses recently published books. Yet as training corpora expand, newer LLMs inevitably become familiar with these texts, reducing the task's effectiveness. As shown in our experiments, while the public subset of NoCha has been largely conquered by LLMs, our subset comprising works from the same period or earlier remains challenging. This shows that our task is not becoming easier for models over time.

3 The Proposed PRELUDE Task

In this section, we introduce the construction process of our PRELUDE (PRequel Entailment for Long context Understanding and DEduction) task.

3.1 Task Formulation

Our task is formulated as binary classification. The input consists of a book $\mathcal{B}$ that is split into M consecutive chunks $\mathcal{B} = \{b_1, b_2, ..., b_M\}$; and a prequel p for a character c, which is a short text describing an experience of the character prior to the story of $\mathcal{B}$ happens. The task is then predict whether p aligns with $\mathcal{B}$. The labels to predict thus are $\{consistent, contradict\}$.

3.2 Why Prequels?

Our prequel entailment task is naturally a long-context understanding task and a form of everyday research task. To solve the task, a model needs to judge whether a character's prequel remains consistent with the behaviors and experiences throughout the narrative and makes counterfactual reasoning when necessary. These requirements make our task well-suited for benchmarking long-context reasoning for the following desirable properties:

- **Natural long-context reasoning:** The task requires holistic understanding of a narrative arc, including tracking a character's psychological continuity, goals, and situational influences across temporally distant events.
- **Cognitive research practices in daily life:** While formal research is often confined to scientific domains, its core cognitive components, such as gathering evidence, forming hypotheses, and drawing conclusions, are deeply embedded in daily reasoning. Our task scenario mirrors this real-life cognition, as humans frequently make similar judgments while watching films, reading novels, or engaging in social interactions.
- **Light dependency on background knowledge:** This task requires little reliance on specialized knowledge. A reader with a full understanding of the story, even as a middle school student, can often make accurate judgments. As a result, the task emphasizes fluid intelligence rather than knowledge acquired through prior learning.

3.3 Dataset Construction

Label Definitions To facilitate human annotation, we categorize the consistent and contradictory cases into fine-grained types. Definitions and representative examples are provided in Table 1.

Guidelines for Human Annotation Annotators are instructed to follow the definitions and examples provided in Table 1. They are guided by a flowchart that first identify *Contradict - Local*, *Contradict - Global I*, and *Contradict - Global II* respectively. If no contradiction is identified, they then determine whether the consistent prequel provides key missing information. During annotation, annotators are required to carefully consult the original book to identify any contradictions.

Annotation following this flowchart is generally sufficient for people familiar with the book. However, during trial annotation, we identified three issues hence introduced the following rules:

- First, judgments must be based solely on the content of the original novel. Adaptations, derivative works, or historical inspirations behind the characters should not be considered. Otherwise, annotators might incorrectly flag a contradiction on a case consistent with the novel based on, for example, a historical figure's biography.
- Second, annotators should assume that the prequel is followed immediately by the original story, with no additional text or events in between. This rule addresses a tendency among human annotators to over-interpret. For instance, when facing a clear contradiction, they might imagine that some unseen event occurred between the prequel and the original story to make it eventually consistent. This rule reduces such subjectivity.
- Finally, unless a character's statements are later explicitly revealed to be deceptive, they should be treated as factual, akin to narrative exposition. Otherwise, one could dismiss any contradiction with the canonical text as intentional deceit.

Candidate Prequel Generation We prompt DeepSeek-R1 and GPT-4o to generate prequel samples, using the prompt provided in Appendix E.1. We explicitly instruct the LLMs to generate each prequel in Markdown list format. Each bullet point is treated as an individual prequel example for annotators to label the types defined in Table 1.

Annotation Details Four annotators labeled the prequels for 40 characters across 13 books (see Appendix E.2), as shown in Table 2. These books were selected to represent diversity in genre, popularity, and original language. Two annotators are graduate students majoring in literature or related fields, while the other two major in computer science. Each annotator worked on books they were familiar with to ensure high-quality annotations.[1] Table 3 shows that humans identified contradictions in nearly half of the generated prequels.

The annotation process resulted in a total of 795 instances, with each case taking approximately 10 minutes to complete. After training, the annotators reached substantial agreement, with a Kappa score of 0.7828, though some subjectivity in interpretation remained. Most unresolved cases were due to differing interpretations of characters, ambiguities left by the original authors, or inherently fuzzy logic. Representative examples are shown in Appendix G.1, *Examples I* and *II*.

[1] By "familiar", we require that the annotator has read the book multiple times and can recall the overall plot in reasonable detail.

Name	Author	Genre	Lang.	Public	#Char	#Instances
The Count of Monte Cristo	Alexandre Dumas	Adventure fiction	English	Yes	2	54
Demi-Gods and Semi-Devils	Louis Cha	Martial art fiction	Chinese	No	3	78
The Return of the Condor Heroes	Louis Cha	Martial art fiction	Chinese	No	4	44
Investiture of the Gods	Xu Zhonglin	Mythology	Chinese	Yes	4	94
Romance of the Three Kingdoms	Luo Guanzhong	Historical fiction	Chinese	Yes	2	28
Love in the Time of Cholera	Gabriel García Márquez	Magical realism	English	No	1	15
Pinball, 1973	Haruki Murakami	Surrealism	Chinese	No	2	22
Rebecca	Daphne du Maurier	Gothic fiction	English	No	3	82
In Search of the Castaways	Jules Verne	Adventure fiction	English	Yes	4	86
The Redeemer	Jo Nesbo	Crime fiction	English	No	3	82
Drawing Sword	Du Liang	Historical fiction	Chinese	No	3	70
Dwelling Narrowness	Liu Liu	Social critique	Chinese	No	3	92
Distant Saviour	Dou Dou	Romance	Chinese	No	4	48
Total	–	–	–	–	48	919
English	–	–	English	–	13	319
Chinese	–	–	Chinese	–	25	476
Public	–	–	–	Yes	12	262
Human-Written	–	–	–	–	17	124

Table 2: Statistics of PRELUDE.

Label	Count	Label	Count
Consistent	434	Contradict	361
Core	270	Local	94
Irrelevant	164	Global	267

Table 3: Statistics of annotated labels.

Human-Written Samples We follow (Karpinska et al., 2024) to have our four annotators write pairs of consistent and contradict prequels by themselves. The annotators first wrote consistent prequels for the characters they were familiar with. Then, for each prequel, they introduced a slight perturbation to make it inconsistent with the original books. This process yielded 62 pairs of prequels. Each takes about 10 minutes for humans to write.

4 Compared Methods

We compare multiple state-of-the-art LLMs under the following settings. The implementation details can be found in Appendix F.

LLMs with Vanilla Few-Shot ICL This is the vanilla prompting approach that first presents the LLMs with the task instruction and then k examples ($k = 5$ in our case). It does not provide book context in the input so the LLMs need to rely on their inherent parametric knowledge to solve the task. We use the prompts shown in Figure 9 and 11 from appendix, but with the field *Original Excerpt* omitted from the input.

For open-source LLMs, we use the *Instruct* versions of the Qwen2.5/3 models and DeepSeek-R1. We also compare against commercial LLMs, including GPT-4o, o3-mini, and the Gemini 2.5 family, accessed via API calls.

Retrieval-Augmented LLMs We enable the LLMs to access the canonical novels via retrieval-augmented generation (RAG) (Lewis et al., 2020; Guu et al., 2020). We experiment with various embedding models and hyperparameters, as detailed in Section 5.3. Our final system uses the Qwen3-Embedding-8B model to retrieve the top 40 chunks, each with a length of 500 tokens.

In-Domain Post-Training Prior work has shown that fine-tuning with as few as 1K examples can elicit specific capabilities in LLMs, such as mathematical reasoning or general instruction following (Zhou et al., 2023; Zhao et al., 2024; Muennighoff et al., 2025). These successes rely on the assumption that such capabilities are already present in the model acquired by pretraining thus can be activated with minimal supervision. In contrast, if an LLM lacks the potential for a given capability, such low-resource training is unlikely to produce meaningful gains. Thus, the effectiveness of low-resource in-domain training can serve as a diagnostic tool to assess the intrinsic difficulty of a task, as demonstrated in (Yu et al., 2025a).

Following this idea, we reserved the human-study subset for evaluation and used the remaining ∼700 examples to fine-tune the LLMs.

Many-Shot ICL Similar to the in-domain training approach, many-shot ICL (Agarwal et al., 2024; Bertsch et al., 2025) provides a large number of examples in the input context to elicit the latent capabilities of the LLM. We use the same data split as in the in-domain training experiment.

Commercial DeepResearch Commercial deep research services are offered by several companies, notably OpenAI DeepResearch[2]. These services showcase the ability to retrieve and synthesize in-

[2]https://openai.com/index/introducing-deep-research/

Model	Overall F1-Scores			F1-Scores on Subsets				
	Macro-Avg	Consistent	Contradict	Chinese-Set	English-Set	R1-Set	GPT-Set	Public-Set
Qwen2.5-72B	55.7	66.4	45.0	54.8	57.1	54.7	56.8	59.7
+ RAG top-40	56.4	57.9	54.9	56.9	55.5	52.6	60.8	56.0
Qwen3-32B	53.5	69.7	37.4	55.3	50.7	50.7	58.3	55.9
+ RAG top-40	61.3	67.3	55.4	61.7	60.7	59.2	64.8	64.1
Qwen3-235B-A22B	57.3	**70.1**	44.5	54.9	60.9	53.9	62.9	59.1
+ RAG top-40	59.7	59.7	59.7	58.8	60.8	59.2	60.5	60.7
DeepSeek-R1	61.3	69.2	53.4	60.3	62.7	63.9	56.2	**66.2**
+ RAG top-40	59.1	51.3	66.9	63.1	53.0	57.3	61.2	61.8
GPT4o	57.8	69.9	45.8	57.2	58.9	58.8	55.1	62.7
+ RAG	62.9	61.2	64.6	62.3	**63.7**	61.3	64.5	63.2
o3-mini	53.5	68.8	38.2	52.5	54.8	53.3	52.2	56.4
+ RAG	60.0	67.0	53.0	58.5	62.3	57.2	63.7	64.0
Gemini-2.5-Flash	61.8	62.4	61.1	66.4	54.9	59.6	64.6	60.5
+ RAG top-40	57.8	48.9	66.8	62.8	50.2	57.1	58.3	52.0
Gemini-2.5-Pro	**65.1**	61.4	**68.9**	**67.1**	62.2	**64.1**	**66.4**	62.0
+ RAG top-40	60.7	53.7	67.8	61.0	60.2	58.4	63.6	59.9
Humans	81.7*	79.5*	83.9*			–		

Table 4: Comparison of different LLMs on the full set of PRELUDE. *: Experiments conducted on a subset for reference.

formation from multiple sources to generate reports using an agentic approach. We use the web interface from OpenAI, which has demonstrated strong performance across a wide range of everyday tasks.

5 Experimental Results

5.1 Human Performance

We selected 100 examples to compute human performance. Three participants who had not involved in our task annotations and have similar backgrounds to our annotators were asked to annotate examples from books they were familiar with. The results show strong performance, with an F1 score of 81.7% (an accuracy of 82%), indicating that the task is largely solvable by humans.

Upon examining the disagreements, we found that most could be resolved, as they were often due to annotators overlooking information (either from fatigue after extended work or the the fallibility of human memory). The remaining unresolved cases are of similar types to those analyzed in Section 3.3.

Our study further reveals that humans tend to adopt a DeepResearch-style approach, which involves iteratively generating hypotheses and resolving them by locating relevant supporting evidence.

5.2 Results of LLMs

Comparison across LLMs Table 4 compares state-of-the-art open-source and commercial LLMs on our task. The Gemini-2.5-Pro model shows a clear advantage over the others, yet still falls short of human performance by >15%.

Another key finding is that, except for the Gemini-2.5 models, all other LLMs tend to over-predict the *Consistent* label when not given access to the original books, resulting in unbalanced performance. This suggests that these models struggle to encode details of the novels within their parametric knowledge, underscoring the importance of incorporating canonical books in our task.

A per-book breakdown of the results from this study and the following RAG study is provided in Appendix G.2. We also report results on the standalone human-written subset (Appendix A), which shows similar trends but a slightly larger human–machine gap. Pairwise annotations further enable comparison with NoCha (Karpinska et al., 2024), where the gap is much smaller, underscoring the greater challenge of our task.

Impact of RAG Table 4 also presents the performance of various LLMs under the RAG setting. For most models, RAG improves the F1 score on the *Contradict* class. However, we also observe a tendency toward over-rejection, where LLMs predict *Contradict* more frequently with hypercritical reasoning that focuses on minor or debatable inconsistencies. One illustrative example can be found in Appendix G.1, *Example III*.

Notably, RAG results in worse performance for Gemini-2.5-Pro on both classes. This suggests that the retrieved contexts do not effectively contribute useful information for the strong Gemini-Pro model. It also highlights a broader limitation that despite recent advancements, long-context reasoning remains a persistent challenge for LLMs.

Results of In-Domain Training and Many-Shot ICL As shown in Table 5, on our held-out subset, neither in-domain fine-tuning nor many-shot

Methods	Model	Source	F1-Scores			Accuracy	
			Avg	Consistent	Contradict	Answer	Reason
No-Context	Qwen3-32B	–	$53.8_{\pm 1.6}$	$63.2_{\pm 1.7}$	$44.4_{\pm 1.5}$	$55.7_{\pm 1.7}$	–
	Qwen3-235B-A22B	–	$57.4_{\pm 3.5}$	$62.1_{\pm 3.3}$	$52.8_{\pm 3.9}$	$57.7_{\pm 3.1}$	–
	DeepSeek-R1	–	$\mathbf{65.0}_{\pm 0.7}$	$\mathbf{65.0}_{\pm 1.9}$	$65.0_{\pm 2.0}$	$65.0_{\pm 0.8}$	47*
	GPT4o	–	$58.8_{\pm 1.1}$	$64.7_{\pm 1.5}$	$52.9_{\pm 1.4}$	$59.7_{\pm 1.2}$	–
	o3-mini	–	$50.0_{\pm 2.0}$	$55.7_{\pm 0.6}$	$44.3_{\pm 3.7}$	$50.7_{\pm 1.7}$	–
	Gemini-2.5-Flash	–	$63.4_{\pm 3.7}$	$55.2_{\pm 6.0}$	$71.6_{\pm 1.6}$	$65.3_{\pm 2.9}$	–
	Gemini-2.5-Pro	–	$64.3_{\pm 1.6}$	$52.0_{\pm 2.8}$	$\mathbf{76.8}_{\pm 0.4}$	$\mathbf{68.7}_{\pm 0.9}$	43*
RAG (top-40, total length: 20K)	Qwen3-32B	Book	$60.5_{\pm 0.5}$	$\mathbf{60.0}_{\pm 2.1}$	$61.0_{\pm 3.2}$	$60.7_{\pm 0.5}$	–
	Qwen3-235B-A22B	Book	$\mathbf{63.1}_{\pm 2.5}$	$56.9_{\pm 4.2}$	$69.3_{\pm 2.3}$	$64.0_{\pm 2.2}$	45*
	DeepSeek-R1	Book	$59.1_{\pm 4.1}$	$42.4_{\pm 6.7}$	$\mathbf{75.9}_{\pm 1.7}$	$66.0_{\pm 2.8}$	47*
	GPT4o	Book	$60.2_{\pm 2.7}$	$50.8_{\pm 3.7}$	$69.6_{\pm 1.8}$	$62.0_{\pm 2.4}$	–
	o3-mini	Book	$55.7_{\pm 1.9}$	$54.6_{\pm 2.2}$	$56.7_{\pm 1.7}$	$55.7_{\pm 1.9}$	–
	Gemini-2.5-Flash	Book	$60.7_{\pm 1.5}$	$45.8_{\pm 3.3}$	$75.6_{\pm 0.4}$	$\mathbf{66.3}_{\pm 0.5}$	47*
	Gemini-2.5-Pro	Book	$55.8_{\pm 1.6}$	$37.5_{\pm 2.5}$	$74.1_{\pm 0.8}$	$63.3_{\pm 1.2}$	–
In-Domain Training	Qwen3-32B	–	$52.4_{\pm 1.5}$	$58.7_{\pm 3.0}$	$46.1_{\pm 1.5}$	$53.3_{\pm 1.9}$	–
		Book	$59.7_{\pm 0.9}$	$60.1_{\pm 0.2}$	$59.2_{\pm 2.1}$	$59.7_{\pm 0.9}$	–
Many-Shot ICL	DeepSeek-R1	–	$62.3_{\pm 2.3}$	$62.0_{\pm 1.9}$	$62.7_{\pm 2.9}$	$62.3_{\pm 2.4}$	–
	Gemini-2.5-Pro	–	$59.5_{\pm 1.4}$	$46.0_{\pm 2.6}$	$73.0_{\pm 0.3}$	$64.0_{\pm 0.8}$	–
OpenAI DeepResearch	o3	Web	62.5	58.4	66.7	63	51
Humans	–	Book	$81.7_{\pm 4.4}$	$79.5_{\pm 7.6}$	$83.9_{\pm 4.9}$	$82_{\pm 3.9}$	79

Table 5: Comparison on the human-study subset with mean$_{\pm \text{std}}$ from three runs. *: We conduct human verification on the most accurate results from the three runs. For in-domain training, we use the largest model (32B) supported on our infrastructure (8xA800). For many-shot ICL experiments, we compare with the two best-performed models in the *No-Context* block.

ICL improves performance over the baseline usage of the same LLMs respectively. It indicates that current LLMs are still fundamentally limited in the type of reasoning required for our task.

Results of DeepResearch Finally, Table 5 shows that DeepResearch performs worse than the best LLMs, both with and without RAG. It is also less effective at identifying contradictory evidence compared to most RAG-based systems. Since DeepResearch primarily relies on retrieving human-written analyses from the Internet, these results suggest that our task cannot be solved using existing external commentary or interpretations alone.

5.3 Analysis

Correct Answer with Incorrect Reasoning We manually verify the outputs of the LLMs and find that, although they correctly answer a large portion of the tasks, they often fail to arrive at the correct answers through valid reasoning. Due to the labor-intensive nature of this evaluation, we select the best outputs from systems that achieve the highest performance on at least one metric in each block of Table 5. We also include DeepSeek-R1 with RAG, as its best run yields the highest answer accuracy. The detailed rubrics can be found in Appendix B.

The final column of Table 5 and Figure 2 present the results, revealing a clear gap between answer accuracy and reasoning accuracy. A such example can be found in Appendix G.1, *Example IV*. In contrast, human annotators generally agree on their

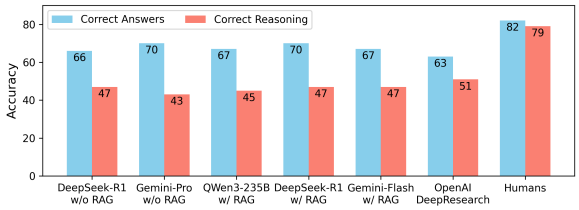

Figure 2: Accuracy the same methods when taking reasoning correctness into consideration.

reasoning, suggesting that current models still lack true comprehension in solving our task.

It is noteworthy that, despite its lower answer accuracy, DeepResearch exhibits the smallest performance drop when considering reasoning accuracy. This suggests combining a strong reasoning model with reflective mechanisms can lead to more reliable reasoning traces.

Impact of Context Length in RAG Figure 3(a) investigates the impact of retrieved context lengths. The RAG system achieves its best performance when the input length is around 20k tokens. With shorter contexts, the retriever often fails to include important evidence due to limited capacity. Conversely, with longer contexts, the excess information can overwhelm the LLM and hinder its ability to effectively utilize the extended context.

Different Retrieval Methods Table 6 provides ablation study on our retrieval method using the following variations:

- *Sorting the retrieved chunks according to their order in the books*: While this intuitively provides a more coherent context, it overlooks chunk rele-

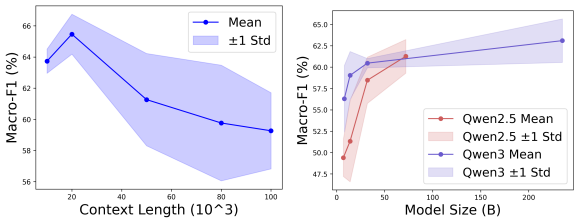

Figure 3: (a) Effect of retrieved context length on RAG performance. We use Qwen3-Embedding-8B and GPT4o to generate the results on the human study subset. (b) Study of the impact of model sizes under the best setting from (a).

Retriever	Top-k	Len	Overall F1-Scores Macro-Avg	Consis.	Contra.
No-RAG	–	–	57.8	**69.9**	45.8
Qwen3-8B	40	500	**62.9**	61.2	**64.6**
+ sorted	40	500	60.1	58.8	61.5
Qwen3-8B	20	1k	61.9	62.3	61.4
BGE-M3	40	500	61.4	61.9	60.9
BGE-M3	20	1k	59.7	59.1	60.4

Table 6: Comparison of different retrieval methods with the same context length. We use GPT4o to generate the results.

vance, resulting in decreased performance.

- *Doubling the chunk size while keeping the input length*: This causes a slight performance drop, likely due to the reduced effectiveness of embedding models on longer chunks (Wu et al., 2025a).
- *Replacing the embedding model with BGE-M3*: This slightly reduces performance.
- *Replacing with BGE-M3 while doubling the chunk size*: This results in a further decrease in performance, likely due to BGE-M3's weaker handling of long inputs.

Effect of Model Scaling Figure 3(b) shows how model performance changes with increasing model size. We experiment with both the Qwen2.5 and Qwen3 series. As shown in Table 4, Qwen models rely on RAG to incorporate additional knowledge and improve performance. Therefore, all experiments in this section are conducted under the RAG setting, with each model evaluated over three runs.

The results indicate that for both Qwen series, performance consistently improves as model size increases. However, this improvement begins to plateau beyond the 32B model. Notably, the 235B model even outperforms the larger 671B R1 model on certain metrics, suggesting that simply scaling up model size is not efficient to our task.

6 Discussions

Limited Long Context Reasoning Capability in Recent LLMs Table 4 shows that while some LLMs benefit from retrieved contexts, advanced reasoning models, such as DeepSeek-R1 and Gemini-2.5, exhibit a notable performance drop when context is provided. This gap becomes more pronounced as the base models grow stronger.

These findings suggest that recent improvements in LLMs' general reasoning capabilities do not necessarily translate to better long-context reasoning. One possible explanation is that as models become more powerful, their internal (parametric) knowledge is more efficient to solve tasks, making them prone to ignore external inputs during post-training. This highlights the need for improved training data and strategies specifically designed to encourage long-context reasoning.

Model Bias in Our Dataset Construction Method If our dataset construction method introduced bias, we would expect a model to perform worse on examples it generated itself, assuming that LLMs inherently trust their own outputs. However, the results in Table 4 show that this is not the case, indicating that our construction process does not introduce significant bias toward any particular model family. This also suggests that state-of-the-art LLMs do not inherently trust their own generations and remain susceptible to hallucination, even when evaluating content they previously produced.

Measuring Fluid Intelligence in Natural Language Space Combined with the observation that DeepResearch performs poorly on our task, it becomes evident that the task cannot be solved simply by retrieving existing information from the web. Instead, it requires generating new knowledge through reasoning based on learned rules, aligning with the notion of fluid intelligence tests (Chollet, 2019; Chollet et al., 2025; Yu et al., 2025a). Unlike prior work, our task represents the first fluid intelligence assessment conducted entirely in the natural language space.

7 Conclusion

We introduce PRELUDE, a new benchmark for evaluating long-context comprehension and reasoning in LLMs. Our task design addresses several key shortcuts present in prior long-context benchmarks. Experiments show that state-of-the-art models still fall significantly short of human performance, particularly in generating valid reasoning. PRELUDE calls for further research into robust long-context understanding and the development of models with stronger global reasoning capabilities.

Limitations

Our task empirically mitigates the shortcuts observed in prior work (as shown in Figure 1). However, due to the inherent complexity of long-context reasoning and the subjective nature of interpreting literary narratives, human performance on our task is also non-perfect. Despite the fact that this subjectivity does not affect the conclusions of our work (refer to Appendix C for more analysis), in future work, we aim to improve the annotation framework to further enhance inter-annotator consistency and reduce the annotation and human study workload of the task.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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. 2024a. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. 2023. Language modeling is compression. *arXiv preprint arXiv:2309.10668*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099, Torino, Italia. ELRA and ICCL.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2024. What is wrong with perplexity for long-context language modeling? *arXiv preprint arXiv:2410.23771*.

Google DeepMind. 2024. NotebookLM: Your AI-Powered Research Assistant. `https://notebooklm.google/`. Accessed: 2025-07-24.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Shousheng Jia Haosheng Zou, Xiaowei Lv and Xiangzheng Zhang. 2024. 360-llama-factory.

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? *arXiv preprint arXiv:2404.06654*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Bernard Ghanem, Ping Luo, and Guohao Li. 2025. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTestNeedleInAHaystack/tree/main. Accessed: 2025-07-23.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. 2025. ETHIC: Evaluating large language models on long-context tasks with high information coverage. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5497–5512, Albuquerque, New Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. 2025a. Lara: Benchmarking retrieval-augmented generation and long-context llms–no silver bullet for lc or rag routing. *arXiv preprint arXiv:2502.09977*.

Mo Li, Songyang Zhang, Taolin Zhang, Haodong Duan, Yunxin Liu, and Kai Chen. 2025b. Needlebench: Can llms do retrieval and reasoning in information-dense context?

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2025c. Long-context LLMs struggle with long in-context learning. *Transactions on Machine Learning Research*.

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 Association for Computational Linguistics*, 12.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Xuanfan Ni, Hengyi Cai, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, and Piji Li. 2024. Xl^2 bench: A benchmark for extremely long context understanding with long-range dependencies. *arXiv preprint arXiv:2404.05446*.

OpenAI. 2025. Deep research system card. Technical report, OpenAI.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Zexuan Qiu, Jingjing Li, Shijue Huang, Xiaoqi Jiao, Wanjun Zhong, and Irwin King. 2024. Clongeval: A chinese benchmark for evaluating long-context large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Ilya Sutskever. 2023. A theory of unsupervised learning.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of

large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024a. Ada-LEval: Evaluating long-context LLMs with length-adaptable benchmarks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3712–3724, Mexico City, Mexico. Association for Computational Linguistics.

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

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024c. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646, Miami, Florida, USA. Association for Computational Linguistics.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024d. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. 2025. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*.

Junjie Wu, Jiangnan Li, Yuqing Li, Lemao Liu, Liyan Xu, Jiwei Li, Dit-Yan Yeung, Jie Zhou, and Mo Yu. 2025a. Sitemb-v1.5: Improved context-aware dense retrieval for semantic association and long story comprehension.

Yuhao Wu, Yushi Bai, Zhiqing Hu, Shangqing Tu, Ming Shan Hee, Juanzi Li, and Roy Ka-Wei Lee. 2025b. Shifting long-context llms research from input to output. *arXiv preprint arXiv:2503.04723*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148.

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

Nan Xu and Xuezhe Ma. 2025. Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3344–3370.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025a. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. 2024b. Stress-testing long-context language models with lifelong ICL and task haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, and Xipeng Qiu. 2025b. DetectiveQA: Evaluating long-context reasoning on detective novels. In *Workshop on Reasoning and Planning for Large Language Models*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP 2018*, pages 2369–2380.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. *arXiv preprint arXiv:2305.10156*.

Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. 2025a. The stochastic parrot on llm's shoulder: A summative assessment of physical concept understanding. *arXiv preprint arXiv:2502.08946*.

Yifei Yu, Qian-Wen Zhang, Lingfeng Qiao, Di Yin, Fang Li, Jie Wang, Zengxi Chen, Suncong Zheng, Xiaolong Liang, and Xing Sun. 2025b. Sequential-niah: A needle-in-a-haystack benchmark for extracting sequential needles from long contexts. *arXiv preprint arXiv:2504.04713*.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *International Conference on Machine Learning*, pages 60674–60703. PMLR.

Chulun Zhou, Qiujing Wang, Mo Yu, Xiaoqian Yue, Rui Lu, Jiangnan Li, Yifan Zhou, Shunchi Zhang, Jie Zhou, and Wai Lam. 2025. The essence of contextual understanding in theory of mind: A study on question answering with story characters. *arXiv preprint arXiv:2501.01705*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

A Experiments on Human-Written Prequels

We investigate how the use of LLM-written prequels affects evaluation outcomes. To this end, we evaluate the top three systems according to Table 5 (DeepSeek-R1, Gemini-2.5-Pro, and Gemini-2.5-Flash, all without RAG) on the human-written subset and compare their performance with the results on the LLM-written subset and the human-written NoCha public subset.

Table 7 shows that, similar to the results on LLM-written prequels, all evaluated models perform poorly on the human-written ones. Our human-written prequels are slightly more challenging, likely because LLMs tend to produce more superficial prequels. The LLM-written subset therefore serves as an easier variant of the task, where strong performance indicates readiness for the more demanding human-written version. Overall, this confirms that the difficulty of our PRELUDE does not stem from the use of LLM-generated prequels, but rather from the task's inherent demand for long-context reasoning – a capability that current paradigms still struggle to master.

Finally, we would like to highlight that the human-written prequels exhibit less subjectivity, as reflected by the higher human accuracy of 88.7%. Some human errors arise from inaccurate recollections of the original books; after reviewing the annotation rationales, the human agreement increases to 91.1%, indicating a reduced level of subjectivity. This, in turn, results in a larger *Human–Machine Gap*, as defined in Appendix D.

Models	PRELUDE		NoCha
	LLM	Human	Human
Gemini-Pro	$68.7_{\pm 0.9}$	$61.3_{\pm 1.7}$	$88.9_{\pm 0.0}$
Gemini-Flash	$65.3_{\pm 2.9}$	$59.7_{\pm 1.7}$	$76.2_{\pm 1.3}$
DeepSeek-R1	$65.0_{\pm 0.8}$	$58.1_{\pm 0.7}$	$75.1_{\pm 0.7}$
Humans	$82.0_{\pm 3.9}$	$88.7_{\pm 5.0}$	~ 97.0
Human–Machine Gap	41.6%	70.8%	17.2%

Table 7: Comparing the accuracy on LLM-written and human-written subsets from our task. We report the results on NoCha (public) for reference, for which the human performance is estimated from their reported pair-accuracy.

B Details of Human Study on Reasoning Accuracy

We provide a detailed description of the evaluation process for the reasoning outputs.

Evaluators The same three annotators from the human study (Section 5.1) were recruited to evaluate examples from the books they had previously worked on. This ensures that each example's predictions from different LLMs were evaluated by the same annotator for consistency.

Annotation Guidelines The annotators were instructed to follow the steps below when evaluating the reasoning processes:

- *Identify factual errors in the premises* of the reasoning processes (e.g., the Example IV in our submission). This usually occurs when an LLM has an intuition about the predicted label but cannot recall the critical evidence, and thus hallucinates supporting information.
- *Identify irrelevant evidence as premises*, where the recalled information is true but unrelated to judging the prequel. This often occurs in the *Contradict* cases where the prequel directly contradicts facts in the book, but these relevant facts is not recalled. Instead, the LLM introduces unrelated information and artificially links it to the prequel.
- *Insufficient evidence as premises*, where the recalled pieces of evidence are correct and relevant but not sufficient to derive the predicted labels.
- *Misunderstanding of the prequel*, which is a rare case where the reasoning process indicates that the LLM has fundamentally misunderstood the content of the prequel.
- *The subjective interpretation is excluded* from reasoning errors. That says, if none of the first four types of errors are identified and the LLM arrives at the same correct answer as humans, the reasoning process is considered correct.

Observations From examining 50 examples where LLMs produced correct answers with incorrect reasoning (using the best-performing system, Gemini-Pro w/o RAG), we found that 60% of the errors were due to irrelevant evidence. Surprisingly, 30% were due to factual errors, indicating both flawed memory and a high rate of hallucination even in a strong model like this. Furthermore, since subjective interpretation occurs more often in the *Consistent* cases, and it has already been excluded, we regard most correct predictions on *Consistent* cases as reflecting correct reasoning.

C Analysis on the Impact of Subjective Interpretations

Given the relatively lower human performance, it is a common concern that the subjective nature of the task makes the evaluation susceptible to interpretation differences. In this section, we show that although subjective interpretation exists among our human annotators, it does not account for the poor results of LLMs:

First, humans are generally aware of the existence of subjective interpretations. As a result, it is observed that when humans read the reasoning provided by others, they often agree with those reasons even if they initially arrived at the same conclusion through different reasoning paths. During the annotation process, for cases where annotators disagree, they can still recognize that the disagreement stems from subjective interpretation. In contrast, this is not the case for LLMs. From our analysis in Figure 2 and Appendix B, the gaps between their answer accuracy and reasoning accuracy are primarily due to factual and logical errors in the reasoning steps, not diverse interpretations. In other words, the substantial drop in LLM performance is not caused by subjective interpretation.

Second, to further support this, we conducted additional experiments on the subset of instances where human annotators showed perfect agreement, thereby largely eliminating the influence of subjective interpretation. However, all the LLMs still performed as poorly as they did on the full set. Table 8 shows the performance changes of the top four systems.

Finally, as shown in Appendix A, on the human-written prequels where subjectivity is lower, the LLMs perform even worse. This finding indicates that the poor performance of LLMs is not primarily due to subjective human interpretations.

Models	Accuracy	
	full set	subjectivity-reduced
Gemini-Pro w/o RAG	70.0	70.7
Gemini-Flash w/o RAG	67.0	69.5
QWen3-235B w/ RAG	67.0	69.5
DeepSeek-R1 w/o RAG	66.0	68.3
Humans	82.0	100.0

Table 8: Comparing the results on the full human-study subset and that with the subjectivity reduced.

Taken together with the above, this indicates that LLMs diverge from human in a fundamentally different manner than humans disagree with each other, which stems from LLMs' lack of global reasoning and understanding. Therefore, while interpretive differences may affect the upper bound of human performance on our task, they do not undermine its validity as an evaluation benchmark.

D Assessment of Assessments: What Makes a Long-Context Benchmark Truly Meaningful?

For each criterion discussed in Section 1, we first go through the definition and examples of the criterion, then propose its delegate measure:

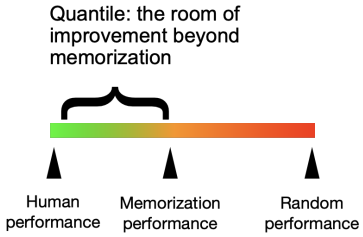

Figure 4: Illustration of the measurement for the criterion of *Beyond Memorization*.

Beyond Memorization. LLMs are known to memorize large amounts of training data (Tirumala et al., 2022; Delétang et al., 2023; Sutskever, 2023). For popular texts (*e.g.*, widely read books), models may recall content or associated analysis from pretraining, bypassing the need for actual comprehension. As training datasets continue to expand, this issue becomes increasingly problematic. Therefore, a benchmark should be designed such that it cannot be solved purely through memorized knowledge, ensuring that the full length and structure of the context remain necessary for reasoning.

• *Measurement:* We first evaluate the memorization performance of GPT-4o by having it answer questions without using RAG. We then compute its quantile within the range defined by human and random performance. This measurement is illustrated in Figure 4.

Global Dependency. The task should require aggregating information from multiple pieces of evidence that are scattered across the context or exhibit global dependencies. Otherwise, it reduces to a short-context problem solvable by retrieving the relevant passage for a given question. In such cases, the task becomes more about improving retrieval quality than long-context understanding ability.

• *Measurement:* This dimension differs slightly from the previous one in how the performance interval is defined. To evaluate how much better a model performs compared to using a single document, the

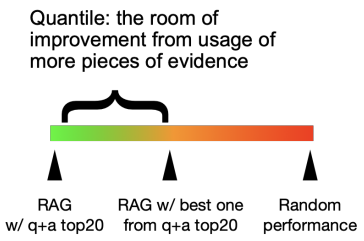

Figure 5: Illustration of the measurement for the criterion of *Global Dependency*.

reference point should be the model's performance with sufficient evidence, rather than human performance. To approximate this, we retrieve the top-20 documents using both the question and the answer, and treat this as the upper bound of performance with sufficient evidence. We then ask GPT-4o to select the best supporting document from the top-20 retrieved using both the question and answer, treating it as the **strongest single piece of evidence**. We compute the quantile of the model's RAG performance using only this selected document, relative to the interval defined by the QA-top-20 RAG performance and the random baseline.

Depth of Reasoning. By nature, long-context reasoning involves synthesizing multiple pieces of evidence across the input. The complexity of reasoning—especially multi-step deduction—is closely tied to task difficulty. Many existing benchmarks emphasize shallow reasoning, primarily requiring decomposition or enumeration (e.g., counting mentions of a fact (Xu and Ma, 2025), or multi-hop QA where the questions are often constructed as sequences of relatively simple sub-questions (Yang et al., 2018). This limits the need for global reasoning and makes tasks easier than intended.

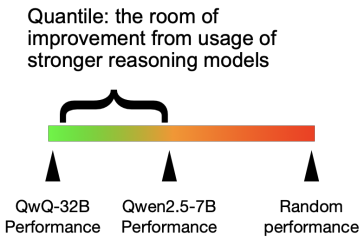

Figure 6: Illustration of the measurement for the criterion of *Depth of Reasoning*.

• *Measurement:* Similar to the previous dimension, the key in this measurement lies in identifying a representative performance interval and a suitable reference model for computing the quantile. We choose QwQ-32B as a strong reasoning model and treat its RAG performance using the top-20 documents retrieved with both the question and answer (as defined in the previous measurement) as the upper bound. We then compute the quantile of Qwen2.5-7B under the same RAG setting. These two models are selected because they both belong to the Qwen2.5 family and therefore share the same pre-training procedure. The QwQ-32B is the largest model in this family that has additionally undergone RL-based test-time inference training. Since reasoning capabilities can stem from both model scaling and RL training, their differences reflect the impact of these factors on reasoning performance. This gap reflects the potential for improvement attributable to stronger reasoning capabilities. Thus, the deeper the reasoning required by a dataset, the larger this gap is expected to be.

Note that this proxy becomes less informative if the dataset is so challenging that neither model significantly outperforms the random baseline. However, as shown in Table 9, this is not currently the case. Therefore, this measurement remains a meaningful indicator.

Beyond Summarization/Salience. This often-overlooked criterion is crucial: Tasks that can be resolved simply by generating a summary of a long input are less challenging and may no longer probe deep understanding, especially given recent advances in summarization (*e.g.*, hierarchical or iterative methods (Chang et al., 2023)). In such cases, the challenge of long-context understanding is effectively reduced to a summarization task over short segments, creating a shortcut. A high-quality benchmark should instead require attention to fine-grained details that go beyond high-level abstraction, or it can be reduced to a relatively easier task of summarization thus is less meaningful.

• *Measurement:* This dimension is measured by comparing the performance of RAG using original text chunks against RAG using chunk summaries of the same input length. We do not use quantiles here but instead directly report the relative improvement, as some datasets actually perform worse when using the original texts, indicating that the task questions primarily target salient events, making summarization sufficient for answering.

Human-Machine Gap. To highlight essential capabilities that general-purpose intelligent systems ought to possess, a meaningful benchmark should show a significant gap between humans and machines. At the same time, poor human accuracy or inter-annotator agreement typically signals low

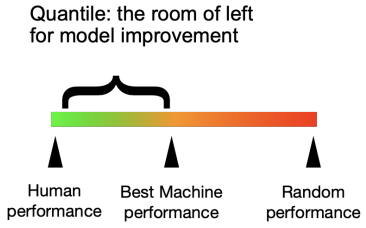

Figure 7: Illustration of the measurement for the criterion of *Human-Machine Gap*.

annotation quality or high subjectivity — both detrimental to robust evaluation.

• *Measurement:* We identify the best machine performance from the evaluations in the *Beyond Memorization* and *Beyond Summarization* dimensions and compute its quantile within the interval defined by human performance and the random baseline.

Compared Datasets. We compare our PRELUDE with representative benchmarks from prior work, including (Kočiskỳ et al., 2018; Karpinska et al., 2024; Zhao et al., 2024; Xu et al., 2025b). These widely used datasets span multiple genres such as classic and detective novels, support both English and Chinese languages, and cover a range of task formats including free-form QA, multiple choice, and true-or-false questions. For a fair comparison, we sample ∼100 questions from each benchmark: InfBench.MC (25 books), DetectiveQA (10 books), and NarrativeQA (10 books from the development set), ensuring a similar number of examples across datasets.

E Details of the Dataset

In this section we provide the detailed characters and questions used in our study in both English and Chinese dataset. The public subset will be included in our code and data released upon publication.

E.1 Prompt for Generating Prequels

We simple prompt the LLMs with no additional context to generate the prequels for human annotation. Figure 8 demonstrates the English translation of our prompt.

E.2 Selected Characters

Table 10 lists the main characters used to in our PRELUDE for annotation.

F Implementation Details

F.1 Hyperparameters of In-Domain training

We apply LoRA (Hu et al., 2021) to all available target modules for in-domain training with RAG and non-RAG settings. The training is conducted on the dataset excluding the human-study subset, resulting in a total of 695 instances. Since the average token length in the RAG setting is above 20k, we perform LoRA tuning on both settings using 360-LLaMA-Factory (Haosheng Zou and Zhang, 2024). We fine-tune each model for 2 epochs with a batch size of 1 on a single machine with 8 H20 GPUs. The dimension of LoRA's attention layer is set to 64, while α is configured to 128. In addition, the learning rate and warmup ratio are set to 1e-4 and 0.1 respectively.

F.2 Prompt Templates

We use the prompt template in Figure 9 and 11 for the RAG experiment, depending on whether the instance originates from a Chinese book or an English book. For the "No Context" experiment, we remove the original excerpt part both in the ICL examples and the actual instance.The exact content of ICL in both languages is in Figure 10 and 12.

G Additional Results

G.1 Illustrative Examples for Qualitative Study

- *Subjective Example – Fuzzy Logic*: *Example I* shows an example that our annotators felt difficult to arrive an agreement due to the fuzzy reasoning. This setting results in the fact that the character appears in the story at around 120 years old. The novel does feature several characters in their 100s who remain healthy and skilled in combat. Still, as the most advanced martial artist in the story, whether 120 is a reasonable age is a matter of fuzzy logic.
- *Subjective Example – Different Interpretation of Persona (Example II)*: In the story, the Sweeper Monk is not a formal disciple of Shaolin; he is more like a hermit. So the disagreement arises that when others mention that many Shaolin disciples had learned Xiaowuxiang Gong, whether the Sweeper Monk would interpret that as referring to himself? The annotators would project themselves into the character's perspective when making predictions, which introduces subjectivity.

Criterion	Metric	Datasets				
		InfBench.MC	DetectiveQA	NarrativeQA	NoCha (public)	PRELUDE
–	Random	25.0	25.0	7.1	50.0	50.0
–	Human Performance	100.0*	94.0	64.3	97.0	82.0
–	Human-Random Gap	75.0	69.0	57.2	47.0	32.0
Beyond Memorization	Memorization Performance	77	50.5	36.1	80.2	56
	Quantile of Memorization	30.7	63.1	49.3	35.8	**81.3**
Global Dependency	RAG (qa-top20)	85.0	73.3	35.9	73.8	59.0
	RAG (best one from qa-top20)	84.0	55.2	32.9	57.4	53.0
	Quantile of RAG (best-one)	1.7	37.4	10.6	68.9	**66.7**
Depth of Reasoning	Qwen2.5-7B w/ RAG qa-top20	68.0	68.6	24.0	60.3	52.0
	QwQ-32B w/ RAG qa-top20	86.0	81.0	42.3	66.7	56.0
	Quantile of Qwen2.5-7B	29.5	22.1	52.1	38.1	**66.7**
Beyond Summarization	RAG over chunks (max-10K)	83.0	77.4	23.9	68.3	61.0
	RAG over chunk summaries (max-10K)	80.1	77.2	27.1	73.8	59.0
	Relative Improvement	**3.6**	0.3	-11.8	-7.5	3.4
Human-Machine Gap	Best Machine Performance	83.0	77.4	36.1	80.2	61.0
	Quantile of Machine Performance	22.7	24.0	49.3	35.8	**65.6**
Average Doc Length	–	228K	97K	104K	153K	**408K**

Table 9: The detailed experimental results we used to compute the measurement of representative criteria for good long-context understanding benchmarks. Except for the Qwen2.5-7B and QwQ-32B results, we generate the results with Qwen3-Embedding-8B and GPT4o. *The paper did not provide the human performance but claimed humans can achieve near-perfectly so we trust their claim.

Name	Author	Characters
The Count of Monte Cristo	Alexandre Dumas	*Faria, Noirtier, General de Quesnel*
Demi-Gods and Semi-Devils	Louis Cha	*Bo Murong, Yellow-Browed Monk, Sweeper Monk*
The Return of the Condor Heroes	Louis Cha	*Yin Zhiping, Yelv Qi, Huo-Du, Gongsun Zhi*
Investiture of the Gods	Xu Zhonglin	*Wen Zhong, Zhao Gongming, Zhang Kui, Huang Feihu*
Romance of the Three Kindoms	Luo Guanzhong	*Xiahou Yuan, Taishi Ci*
Love in the Time of Cholera	Gabriel García Márquez	*Don Pius V Loayza, Lotario Thugut, Rosalba*
Pinball, 1973	Haruki Murakami	*The Rat, J, Naoko*
Rebecca	Daphne du Maurier	*Mrs. Van Hopper, Mrs. Danvers, Jack Favell*
In Search of the Castaways	Jules Verne	*Jacques Paganel, Tom Ayrton/Ben Joyce, Thalcave, Kai-Koumou*
The Redeemer	Jo Nesbo	*Christo Stankic, Martine Eckhoff, Jon Karlsen*
Drawing Sword	Du Liang	*Tian Moxuan, Ding Wei, Yamamoto Kazuki*
Dwelling Narrowness	Liu Liu	*Chen Sifu, Mrs. Song, Su Chun*
Distant Saviour	Dou Dou	*Liu Bing, Han Chufeng, Xiao Yawen, Lin Yufeng*
Attack on Titan	Hajime Isayama & the Fans	*Eren Kruger*

Table 10: Main characters considered in our PRELUDE.

- *Example that RAG Performs Worse*: *Example III* gives an example in which the vanilla LLM can predict correct answer but fails when equipped with RAG. The prequel is consistent because it does not violate any part of the story. Specifically, in a flashback scene from the original novel, it is revealed that Murong Bo's mother subjected him to harsh training in order to raise him as a future monarch for national restoration. When the RAG system fails to recognize that this passage describes Murong Bo's childhood, it tends to rely solely on the more prominent information in the input (e.g., Document 0) to make its judgment.
- *Example with Correct Answer but Incorrect Reasoning*: In *Example IV*, the model correctly answers the question, but bases on incorrect reasons. The actual basis for judging this as a contradiction lies in the fact that both characters are portrayed as emphasizing personal martial prowess, with no demonstration of strategic ability or any depiction of possessing military treatises in the original novel. The act of gifting The Six Secret Teachings (Liu Tao)—a text regarded in Chinese tradition as one of the most profound and authoritative works on military strategy—stands in sharp contrast to their characterization.

The model's explanation is invalid because it fails to grasp that Wu Anguo was representing Beihai's forces (under Kong Rong) in the battle at Hulao Pass. Instead, it superficially interprets the text and incorrectly assumes that Wu Anguo's sphere of activity was limited to Hulao Pass.

This reflects the same underlying issue as in the RAG failure example: LLMs are heavily influenced by the provided context and tend to over-analyze non-informative clues, often becoming overly literal and fixating on minor wording differences.

G.2 Performance Decomposition to Books

```
You are a writing assistant who is deeply familiar with the plots of various renowned literary works and highly
skilled in literary analysis.

Please write a prequel story for the character "{{ char }}" based on the plot of the novel {{ book_name }}. The
story should provide a background that does not contradict the original work. It should explain all of the
'characters major actions in the original novel, their key personality traits, and their major relationships. Your
prequel should focus on reasonably answering unresolved questions about the character that are raised in the
original novel. These include: important issues that cannot be explained using the original content alone, plot
points that appear inconsistent with the 'characters established traits or internal logic, and significant past
experiences that are mentioned but not elaborated upon.

The background story must remain coherent with the original plot and should not introduce settings or elements that
are clearly inconsistent with the style of the original work.

You only need to write an outline. The outline should summarize key elements of the 'characters backstory and major
life events. Each item in the outline should include specific story-level details, rather than general concepts or
attributes. The total length of the outline should be within 1,000 words. Please present your answer in a Markdown
list format.
```

Figure 8: The prompt we used to generate prequels for human annotation (English translation).

Example I. An example of disagreement among human annotators due to fuzzy reasoning.

Book: Demi-Gods and Semi-Devils **Char:** Sweeper Monk **Groundtruth:** Consistent

Prequel: His birth name was Li Xuankong, a collateral member of the imperial family of Later Zhou during the late Five Dynasties period. As a child, he was orphaned by war and taken in by a hermit from Mount Song, who was secretly a first-generation disciple of the Xiaoyao Sect (unaffiliated with the internal conflict within Wuya Zi's lineage).

Example II. An example of disagreement among human annotators due to subjective interpretation of character persona.

Book: Demi-Gods and Semi-Devils **Char:** Sweeper Monk **Groundtruth:** Consistent

Prequel: As a young prodigy, he simultaneously studied the incomplete version of Shaolin's Yijin Jing and the Xiaoyao Sect's Xiaowuxiang Gong, though he never formally joined any sect.

Example III. An example that RAG makes the performance worse.

Book: Demi-Gods and Semi-Devils **Char:** Murong Bo **Groundtruth:** Consistent

Prequel: On her deathbed, his mother pressed a bloodstained jade pendant engraved with the Yan dynasty dragon emblem into his hand, repeatedly stressing that "the Xianbei Murong clan and the Zhao-Song regime cannot coexist."

Model's Reason for Contradict: In the original novel, Murong Bo's restorationist education clearly stems from a paternal lineage. His mother is portrayed merely as an ordinary family member. The added element of "the mother reinforcing the restorationist ideology" has no textual basis and contradicts the Murong clan's patrilineal transmission system of "from father to son, from grandfather to grandson" (see Document 0).

Example IV. An example that LLMs make correct prediction with incorrect reasoning.

Book: Romance of the Three Kingdoms **Char:** Taishi Ci **Groundtruth:** Contradict

Prequel: At the age of seventeen, while escorting his mother on a visit to relatives, he encountered remnants of the Yellow Turban rebels. In the midst of a fierce battle, he rescued Wu Anguo, the gate captain of Beihai, who was under siege. Moved by his bravery, Wu Anguo gifted him a handwritten copy of The Six Secret Teachings by Jiang Ziya.

Model's Reason for Contradict: In the original novel, Wu Anguo's only appearance in Romance of the Three Kingdoms is in Chapter 5, where he has his wrist severed by Lü Bu during the battle at Hulao Pass. He is never mentioned again, nor is there any record of him serving as a gate captain in Beihai. This character setting not only fabricates an encounter between Wu Anguo and a teenage Taishi Ci (violating Rule 1), but also forcibly relocates Wu Anguo—a defeated general from the Hulao Pass battle—to a post in Beihai, disrupting the narrative continuity of that battle (violating Rule 2).

Book Name	#Samples	GPT4o	+ RAG	R1	+ RAG	Gemini-2.5-Pro	+ RAG
The Count of Monte Cristo	54	66.7	70.4	68.5	70.4	72.2	75.9
Demi-Gods and Semi-Devils	78	50.0	65.4	56.4	66.7	75.6	64.1
The Return of the Condor Heroes	44	56.8	61.4	65.9	65.9	72.7	77.3
Investiture of the Gods	94	59.6	62.8	61.7	67.0	69.1	69.1
Romance of the Three Kingdoms	28	64.3	46.4	64.3	75.0	57.1	46.4
Love in the Time of Cholera	15	46.7	53.3	66.7	60.0	73.3	53.3
Pinball, 1973	22	68.2	68.2	54.5	68.2	63.6	54.5
Rebecca	82	67.1	64.6	68.3	47.6	64.6	65.9
In Search of the Castaways	86	70.9	62.8	74.4	48.8	51.2	47.7
The Redeemer	82	42.7	53.7	48.8	58.5	64.6	61.0
Drawing Sword	70	64.3	65.7	64.3	60.0	67.1	62.9
Dwelling Narrowness	92	70.7	60.9	65.2	57.6	64.1	53.3
Distant Saviour	48	54.2	60.4	56.2	64.6	60.4	66.7

Table 11: Performance decomposition to books. We report accuracy on each book.

```
你是一位对各种著名文学作品剧情了如指掌、且精通文学分析的写作助手。

我正在为小说"{{ book_name "}}中的人物"{{ char "}}写该人物的前传故事，你将阅读到我写的前传中的一条剧情设定，并帮助我判断这一项剧情是否
会与原书中的剧情和设定形成矛盾。

判断规则如下：
1. 如果原书中存在一个细节与我提供的剧情设定直接构成矛盾，请回答"矛盾"；
2. 如果我提供的剧情设定导致原书中部分剧情或者人物行为动机整体不合理，请回答"矛盾"；
3. 如果我提供的剧情设定与原书中剧情不构成矛盾，但是与原书的设定风格不统一，请回答"矛盾"；
4. 如果我提供的剧情设定与原书不存在以上类型的明显矛盾，请回答"一致"；
5. 请在你的回答中给出你的判断，并提供简要的判断依据；
6. 你将被提供一些原书中可能的相关片段作为参考，但是这些内容可能不够全面，因此请结合你对该书的知识和所提供的原文片段进行回答；
7. 你应该只基于小说原书的内容进行判断，不要考虑其他相关作品、改编或者人物历史原型。
8. 在判断时，你需要假设整个故事由该剧情设定的前传故事和原书直接承接构成，两者中间没有其他故事发生。
9. 除非某个人物的台词在后面被显式证明是谎言，否则你应该将其认为是事实，同理所有旁白也应被认为是事实。

以下为该任务的一些示例（每个示例对应的相关原文较长，故略去）：
<example_1>
[example_1]
</example_1>

<example_2>
[example_2]
</example_2>

<example_3>
[example_3]
</example_3>

<example_4>
[example_4]
</example_4>

<example_5>
[example_5]
</example_5>

请根据以上示例，完成以下任务：
[前传剧情设定]
小说：{{ book_name }}
人物：{{ char }}
相关原文：
<chunks>{{ chunks }}</chunks>
标题：{{ subject }}
设定：{{ content }}

请仿照以上例子的输出格式，给出你的回答。
```

Figure 9: Chinese prompt template used for the "RAG" experiment in Table 5, with contents of `[example_i]` presented in Figure 10.

[example_1]:
[前传剧情设定]
小说：西游记
人物：金角大王与银角大王
相关原文：（省略）
标题：下界因果
设定：五百年前孙悟空踢翻八卦炉时遗落三块火砖（对应平顶山三峰），受老君密令下界回收火砖

[输出]
Think：原书中八卦炉遗落火砖形成的是火焰山，而不是平顶山。因此符合规则第一条，与原书细节矛盾。
Answer：矛盾

[example_2]:
[前传剧情设定]
小说：西游记
人物：黄狮精
相关原文：（省略）
标题：初生与灵根渊源
设定：机缘巧合下，它得获一只"黄玉狮爪"法器雏形（后被其修炼为主要武器），为日后战斗增添独特威力。

[输出]
Think：原书中黄狮精从来没有用过黄玉狮爪这种法器，他的武器是四明铲。因此该设定符合规则第二条，与原书多处剧情构成整体不一致。
Answer：矛盾

[example_3]:
[前传剧情设定]
小说：三国演义
人物：夏侯渊
相关原文：（省略）
标题：寒门崛起
设定：得异人传授神行法道术，因此日后带兵可以"三日五百，六日一千"

[输出]
Think：虽然与原书剧情不矛盾，但是原书（三国演义）整体保持写实风格，其中的将军带兵作战主要凭借个人武勇，从来没有使用法术作战的例子。因此该设定符合规则第三条，与原书的设定风格不统一。
Answer：矛盾

[example_4]:
[前传剧情设定]
小说：三国演义
人物：徐庶
相关原文：（省略）
标题：初露锋芒
设定：中平元年（184年）黄巾之乱，护送母亲逃亡途中目睹官军屠村，首次杀人（17岁）。因救被掳妇女得罪当地豪强，开始化名"单福"。

[输出]
Think：原书中有徐庶年少任侠杀人，和化名单福的描写，但是没有给出具体细节。该设定补全了原书缺少的细节，且没有构成矛盾。因此是一个与原著存在高度兼容性的剧情设定。
Answer：一致

[example_5]:
[前传剧情设定]
小说：西游记
人物：黄狮精
相关原文：（省略）
标题：短暂拜师与炼法
设定：为求更高法力，黄狮精曾私自闯入周边其他山头，与多位小妖对决，锤炼法术，也树下冤仇。

[输出]
Think：原文中在相关章节从未提过黄狮精与周边其他山头的冤仇。该设定与原书剧情脱离，但也因此不会与原书剧情构成矛盾。因此是一个与原著剧情无紧密联系的剧情设定。
Answer：一致

Figure 10: Content of the in-context chinese examples in Figure 9.

```
You are a writing assistant who is deeply familiar with the plots of various renowned literary works and highly
skilled in literary analysis.

I am writing a prequel story for the character "{{ char "}} from the novel "{{ book_name "}}. You will read a plot
point from this prequel and help me determine whether it conflicts with the original 'novels plot and setting.

Please follow the rules below when making your judgment:

1. If a detail in the original novel directly contradicts the plot point I provide, please respond with ""Contradict
;
2. If the plot point I provide causes part of the original plot or the 'characters motivations to become broadly
unreasonable, please respond with ""Contradict;
3. If the plot point does not contradict the original plot but is inconsistent with the tone or stylistic setting of
 the novel, please respond with ""Contradict;
4. If none of the above conflicts exist between the plot point and the original novel, please respond with
""Consistent;
5. In your response, clearly state your judgment and briefly explain your reasoning;
6. You will be provided some snippets from the original book for reference, which may not be comprehensive. Please
consider both the snippets and your knowledge about the book to provide the answer;
7. You should base your judgment only on the content of the original novel --do not consider adaptations, derivative
 works, fandoms, or historical inspirations behind the characters;
8. You should assume that the prequel is followed immediately by the original story, with no additional text or
events in between;
9. Unless a character's statements are later explicitly revealed to be deceptive, they should be treated as factual,
 similar for narrative expositions.

Here are some examples for reference (the corresponding original texts are long, so they are omitted):
<example_1>
[example_1]
</example_1>

<example_2>
[example_2]
</example_2>

<example_3>
[example_3]
</example_3>

<example_4>
[example_4]
</example_4>

<example_5>
[example_5]
</example_5>

Please follow the format above and complete the task below:
[Prequel Plot Point]
Novel: {{ book_name }}
Character: {{ char }}
Original Excerpt:
<chunks>{{ chunks }}</chunks>
Title: {{ subject }}
Setup: {{ content }}

Please give your answer following the format shown in the examples above.
```

Figure 11: English prompt template used for the "RAG" experiment in Table 5, with contents of `[example_i]` presented in Figure 12.

[example_1]:
[Prequel Plot Point]
Novel: The Chronicles of Narnia
Character: Eustace Scrubb
Original Excerpt: (omitted)
Title: Early Life
Setup: Born in London and raised as an only child, Eustace was taught to call his parents by their names --"Harold" and "Alberta--"rather than "Mother" and "Father".

[Output]
Think: In the original novel, the Scrubb family lives in Cambridge and Eustace was born in Cambridge England. Therefore, this detail is consistent with Rule 1 and contradicts a factual detail from the original story.
Answer: Contradict

[example_2]:
[Prequel Plot Point]
Novel: The Chronicles of Narnia
Character: Eustace Scrubb
Original Excerpt: (omitted)
Title: Adventure in Underworld
Setup: The Lady of the Green Kirtle directs Eustace to the Underland, helping him find Prince Rilian and kill the serpent responsible for the 'princes enslavement and intent on killing him.

[Output]
Think: In the story, Eustace is actually directed by her to the 'giants castle and accidentally discovers the Underland. Since the Lady of the Green Kirtle is herself the serpent, it would be impossible for her to intentionally lead them to the Underland and help them find the prince. Therefore, this scenario matches Rule 2 and introduces inconsistencies across multiple parts of the original plot.
Answer: Contradict

[example_3]:
[Prequel Plot Point]
Novel: Romance of the Three Kingdoms
Character: Xiahou Yuan
Original Excerpt: (omitted)
Title: Rise from Humble Origins
Setup: He was taught magical speed-enhancing techniques by a mysterious hermit, which allowed him to march "500 li in three days, and 1000 li in six days".

[Output]
Think: Although not contradictory to the plot, the original novel maintains a realistic --tonegenerals fight with martial skills and do not use magic. This setup violates Rule 3 by being stylistically inconsistent with the novel.
Answer: Contradict

[example_4]:
[Prequel Plot Point]
Novel: Harry Potter
Character: Sirius Black
Original Excerpt: (omitted)
Title: Belief and Value
Setup: Rejecting pure-blood elitism, Sirius frequently showed admiration for Gryffindor and spent time with Muggles and so-called "blood traitors" to annoy his parents even before attending Hogwarts.

[Output]
Think: In the original novel, Sirius is introduced as a key figure against pure-blood elitism and is sorted into Gryffindor in 1971. However, the relevant chapters do not describe these specific actions. This addition fills in missing details without introducing contradictions. Therefore, it is compatible with the original.
Answer: Consistent

[example_5]:
[Prequel Plot Point]
Novel: Journey to the West
Character: Yellow Lion Demon
Original Excerpt: (omitted)
Title: Brief Apprenticeship and Spell Training
Setup: Seeking greater power, he once secretly ventured to other nearby mountains, fought with lesser demons, honed his magic, and made enemies along the way.

[Output]
Think: The relevant chapters in the original text do not mention these rivalries. While this setup is detached from the original plot, it 'doesnt contradict it. Thus, it is a side story that remains compatible with the canon.
Answer: Consistent

Figure 12: Content of the in-context english examples in Figure 11.