DETECTIVEQA: EVALUATING LONG-CONTEXT REASONING ON DETECTIVE NOVELS

Zhe Xu[1]*, Jiasheng Ye[1]*, Xiaoran Liu[1]*, Xiangyang Liu[1]*, Tianxiang Sun[1], Zhigeng Liu[1], Qipeng Guo[3], Linlin Li[2], Qun Liu[2], Xuanjing Huang[1], Xipeng Qiu[†,3]
[1]School of Computer Science, Fudan University,
[2]Huawei Noah's Ark Lab, [3]Shanghai AI Laboratory
{zxu24,jsye23,xrliu24}@m.fudan.edu.cn,

ABSTRACT

Recently, significant efforts have been devoted to enhancing the long-context capabilities of Large Language Models (LLMs), particularly in long-context reasoning. To facilitate this research, we propose **DetectiveQA**, a dataset specifically designed for narrative reasoning within long contexts. We leverage detective novels, averaging over 100k tokens, to create a dataset containing 1200 human-annotated questions in both Chinese and English, each paired with corresponding reference reasoning steps. Furthermore, we introduce a step-wise reasoning metric, which enhances the evaluation of LLMs' reasoning processes. We validate our approach and evaluate the mainstream LLMs, including GPT-4, Claude, and LLaMA, revealing persistent long-context reasoning challenges and demonstrating their evidence-retrieval challenges. Our findings offer valuable insights into the study of long-context reasoning and lay the base for more rigorous evaluations. The evaluation code are publicly accessible via our GitHub repository

1 INTRODUCTION

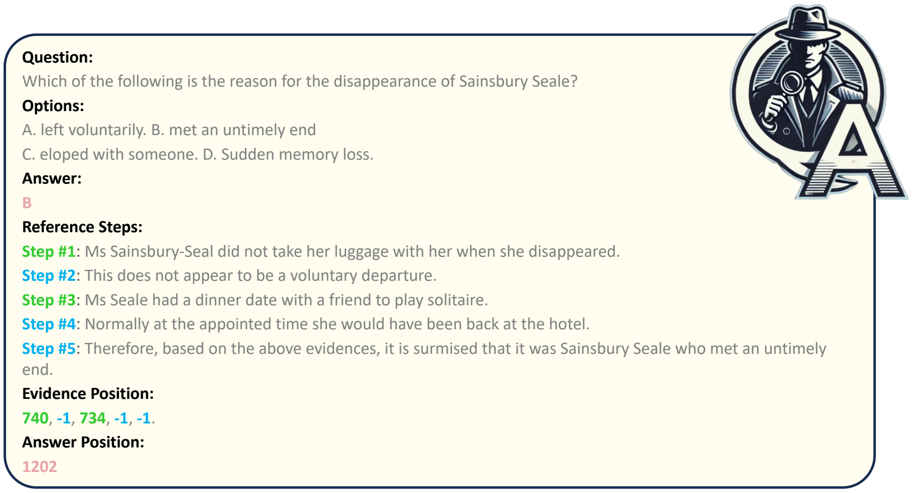
The long-context capabilities of Large Language Models (LLMs) (OpenAI, 2023; Anthropic, 2024a; Touvron et al., 2023; Sun et al., 2024; Bai et al., 2023a; Cai et al., 2024; Zeng et al., 2023b), particularly long-context reasoning(Kociský et al., 2018; Sprague et al., 2024; Wang et al., 2024; Karpinska et al., 2024), are a key competitive advantage in the current landscape. Recently, OpenAI released the O1 model (OpenAI, 2024), which not only supports a context length of 128k but can also generate extensive reasoning chains, effectively solving complex reasoning problems in intricate scenarios. As the long-context reasoning capabilities of LLMs improve, there is a growing demand for more challenging and realistic long-context reasoning evaluations(Sprague et al., 2024; Kuratov et al., 2024; Zhang et al., 2024; Wang et al., 2024; Karpinska et al., 2024). Among these, narrative reasoning based on detective novels provides a sufficiently realistic and challenging setting(Sprague et al., 2024; Gu et al., 2024). As Ilya, the core developer of GPT-4, has stated, if LLMs possess real understanding when presented with *a detective novel with a complicated plot, a storyline, different characters, numerous events, and mysteries like clues*, LLMs should be able to predict who commits the crime at the last page of the book based on the context[1].

Inspired by his words, we propose **DetectiveQA**, a long-context narrative reasoning dataset with three features. First, DetectiveQA provides ***detailed annotation information***. The dataset includes 1200 reasoning questions from English and Chinese detective novels, with an average length exceeding 100k tokens. As shown in Figure 1, we offer not only the questions, options, and answers but also the reference steps, which are the reference reasoning chains for the question, taken by detectives(Wang et al., 2024). Importantly, these reference steps include the *explicit evidence*, the evidence in the text, and the *implicit evidence*, the inference made by detectives. Our DetectiveQA has the highest average reasoning step number compared with other reasoning datasets as shown in

[*] Equal contribution.

[†] Corresponding author.

[1] Ilya Sutskever — GPT4 predicts the next word better — Now upgraded to the more powerful GPT4o https://www.youtube.com/watch?v=1OsHC1vbpc0



Question:

Which of the following is the reason for the disappearance of Sainsbury Seale?

Options:

A. left voluntarily. B. met an untimely end

C. eloped with someone. D. Sudden memory loss.

Answer:

B

Reference Steps:

Step #1: Ms Sainsbury-Seal did not take her luggage with her when she disappeared.

Step #2: This does not appear to be a voluntary departure.

Step #3: Ms Seale had a dinner date with a friend to play solitaire.

Step #4: Normally at the appointed time she would have been back at the hotel.

Step #5: Therefore, based on the above evidences, it is surmised that it was Sainsbury Seale who met an untimely end.

Evidence Position:

740, **-1**, **734**, **-1**, **-1**.

Answer Position:

1202

Figure 1: An example of annotation in DetectiveQA. We highlight the *explicit evidence* of reasoning in blue and *implicit evidence* in green. The whole reference steps include both. In contrast, in the *Evidence Position* field, the part corresponding to the explicit evidence will be the paragraph index in the novel, while that corresponding to the implicit evidence will be -1.

Table 1. We also specify the paragraph locations of the evidence and indicate where the detectives give the answer.

Furthermore, DetectiveQA features ***step-wise reasoning metric***. Besides the stable assessment results through multiple-choice questions, we design an LLM-judged metric to evaluate whether LLM's reasoning processes aligns with the steps taken by detectives. This approach offers a more challenging and feasible method for assessing long-context reasoning(An et al., 2023). Finally, DetectiveQA can provide an ***in-depth analysis*** for long-context reasoning. We adjust the evaluation results and the context used during the assessment to explore the LLM's evidence retrieval capability(Kamradt, 2023; Li et al., 2024), the impact of data contamination(Li et al., 2023; Karpinska et al., 2024), and the differences in reasoning abilities between long and short contexts. To sum up, our contributions can be summarized as follows.

- We propose DetectiveQA, a human-annotated evaluation of narrative reasoning in long contexts, averaging over 100k tokens, with 1200 questions in English and Chinese, each paired with reference steps, averaging over 8 steps.
- Furthermore, we introduce a step-wise reasoning metric that evaluates the LLM's reasoning process, which evaluates the evidence retrieval capability of long-context LLMs and, more importantly, reflects the logical coherence of their reasoning process.
- We evaluate mainstream LLMs, including GPT-4, Claude, and LLaMA, revealing challenges in long-context reasoning. We also identify data contamination issues and differences in reasoning between long and short contexts. Our findings provide valuable insights for research on long-context LLMs and long-context reasoning assessments.

2 Related Work

Long-Context Reasoning Classical long-context benchmark (Bai et al., 2023b; Zhang et al., 2024; Kamradt, 2023) primarily focuses on tasks like QA, summarization, and retrieval, lacking an evaluation of long-context reasoning in real-world scenarios. Traditional long-context reasoning tasks, such as NarrativeQA (Kociský et al., 2018), cover limited clues, while HotpotQA (Yang et al., 2018), despite pioneering multi-hop reasoning, relies on synthetic data that does not provide a realistic context. Recently, NovalQA (Wang et al., 2024) and NoCha (Karpinska et al., 2024) have

Table 1: Comparison of DetectiveQA with other reasoning datasets. To our knowledge, no previous dataset encompasses all of these qualities. ∼ denotes datasets that partially qualify for the property.

Dataset	**Reasoning**	**Natural Long**	**Real-World**	**Process Evaluation**	**Reasoning Step Num**
NarrativeQA (Kociský et al., 2018)	✓	✓	✓	✗	1
HotpotQA (Yang et al., 2018)	✓	✗	✗	✗	2
BABILong (Kuratov et al., 2024)	✓	✗	✗	✗	2.2
NovelQA (Wang et al., 2024)	✓	✓	∼	∼	-
NoCha (Karpinska et al., 2024)	✓	✓	∼	✗	-
MuSR (Sprague et al., 2024)	✓	✗	✓	✗	-
DetectBench (Gu et al., 2024)	✓	✗	✓	∼	-
DetectiveQA (Ours)	✓	✓	✓	✓	8.5

offered more challenging long-context reasoning evaluations through QA in long novels; however, their question designs are not sufficiently natural and are relatively rare in real-world situations.

Among various genres of novels, detective novels are widely regarded as the most distinctive in terms of reasoning features (Gu et al., 2024; Sprague et al., 2024; Del & Fishel, 2023). Detective reasoning questions can authentically reflect LLM's understanding of context and its reasoning capability in real-world scenarios. Although detective novels have seen successful applications in short-context evaluations, such as MuSR (Sprague et al., 2024) and DetectBench (Gu et al., 2024), it has not yet been applied to long-context reasoning evaluation. In response to this gap, we propose DetectiveQA and provide detailed annotation information for classic long-form detective novels.

Reasoning Metrics To measure the reasoning capability of long-context LLMs, the metric focused on the quality of the reasoning process is necessary. However, the commonly used ROUGE metric (Lin, 2004) generally fails to do so (An et al., 2023). Therefore, G-Eval utilizes LLMs like GPT-4 to assess the quality of NLG outputs (Liu et al., 2023), showing a higher correlation with human judgments in summarization and dialogue. Additionally, for mathematical reasoning (Mondorf & Plank, 2024), ReasonEval (Xia et al., 2024) introduces a method for evaluating the reasoning steps in math problems, emphasizing the validity and redundancy of each step and still using LLMs for automatic assessment. However, these studies have not addressed the reasoning process evaluation in narrative reasoning, particularly in long-context reasoning. For long-context reasoning, the most commonly used metrics remain multiple-choice accuracy (Wang et al., 2024) or output matching (Kuratov et al., 2024). In response to this, we draw inspiration from other process evaluations and, considering the characteristics of detective novels, propose an LLM-judged step-wise reasoning metric.

3 Creating DetectiveQA

3.1 Data Source

Detective novels are valuable for studying language models' ability to handle long contexts due to their reasoning-heavy content. Therefore, we consider detective novels as promising candidates to be data sources of our benchmark. However, many prioritize storytelling over rigorous reasoning. Fortunately, we find a group of detective novels categorized as orthodox school (Saito, 2007), which emphasize logical puzzles and provide readers with the same evidence as the detective, making them suitable for our benchmark. Therefore, we collect orthodox detective novels as sources of long context and use questions related to the puzzles in the novels to test the language models.

To ensure a smooth gradient of difficulty, we collect novels ranging from 100k to 250k words. We also limit our collection to Chinese and English versions, aligning with the language proficiency of the research team and data annotators.

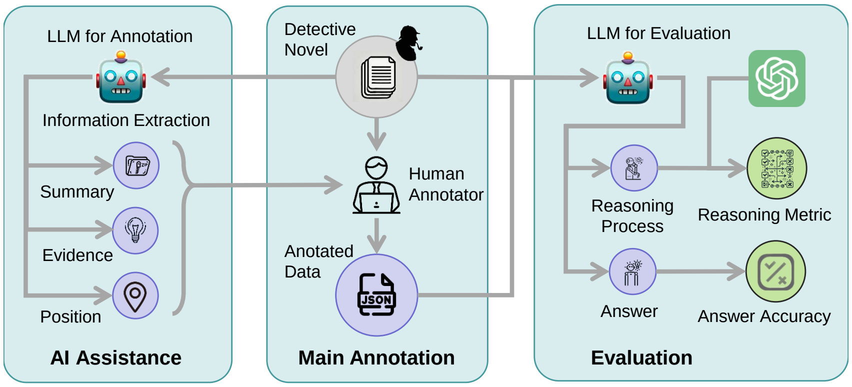


Figure 2: Illustration of DetectiveQA. The center shows the main annotation process, where human annotators annotate reasoning problems based on various information. On the left, AI-assisted information extraction offers summaries to help annotators quickly understand novels and locate key information. The right side, the most critical part, involves evaluating models using DetectiveQA, where reasoning metric and answer accuracy are measured.

3.2 Data Annotation

Data annotation by hiring workers to read long novels (Wu et al., 2021) and generate questions is time-consuming and labor-intensive, taking an average of 3.5 hours to read a 100k-word novel, making it challenging to scale, thus prompting the need for a more efficient alternative.

Our solution is to build an agent workflow using existing LLMs with strong long-context capabilities to *assist* human annotators. The key insight is that, for a complete detective novel, identifying reasoning questions and their corresponding answers can be viewed as an information extraction task, which is a simpler problem that state-of-the-art LLMs have shown promising performance (Zhang et al., 2024). The specific workflow involves the following steps:

(i) **Novel Comprehension.** We first let the LLM summarize the novel to help the annotator grasp the overall story quickly. We input the entire novel again and ask the model to extract the detective's reasoning within the novel. Allow the annotator to understand the reasoning of the novel better.

(ii) **Question proposition.** Next, Human annotators propose reasoning questions based on the reasoning process extracted.

(iii) **Human refinement.** To aid human annotators in verifying the extracted content, we prepend indices to each paragraph in the novel and instruct the model to output the locations of the extracted reference steps. To ensure data quality, the human annotator performs discriminative filtering and refinement of the extracted reasoning with reference to the corresponding location in the original text.

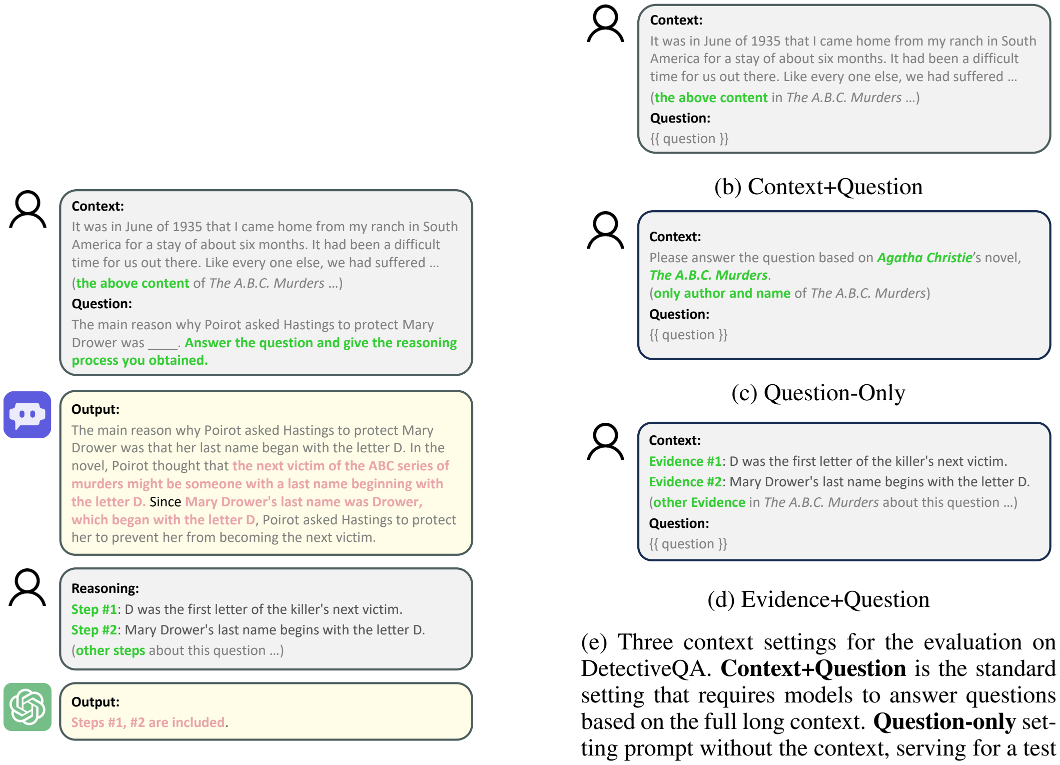
The above process enhances annotation efficiency while ensuring quality. Please refer to Appendix A for annotation details and Section 5.1 for data validation.

4 Evaluating with DetectiveQA

4.1 Metric Settings

In this section, we present the evaluation metrics for DetectiveQA, including multiple-choice accuracy for a stable assessment of reasoning results and the step-wise reasoning metric for a detailed evaluation of the reasoning process.

Multiple-Choice Accuracy Similar to previous evaluations based on multiple-choice questions (Hendrycks et al., 2021; Huang et al., 2023; An et al., 2023), we provide long-context LLM a question with four options and require LLM to output a letter corresponding to the selected option. At this point, we calculate the percentage of correctly answered questions as the score.



Context:
It was in June of 1935 that I came home from my ranch in South America for a stay of about six months. It had been a difficult time for us out there. Like every one else, we had suffered ...
(**the above content** of *The A.B.C. Murders* ...)
Question:
The main reason why Poirot asked Hastings to protect Mary Drower was ____. **Answer the question and give the reasoning process you obtained.**

Output:
The main reason why Poirot asked Hastings to protect Mary Drower was that her last name began with the letter D. In the novel, Poirot thought that **the next victim of the ABC series of murders might be someone with a last name beginning with the letter D.** Since **Mary Drower's last name was Drower, which began with the letter D**, Poirot asked Hastings to protect her to prevent her from becoming the next victim.

Reasoning:
Step #1: D was the first letter of the killer's next victim.
Step #2: Mary Drower's last name begins with the letter D.
(**other steps** about this question ...)

Output:
Steps #1, #2 are included.

Context:
It was in June of 1935 that I came home from my ranch in South America for a stay of about six months. It had been a difficult time for us out there. Like every one else, we had suffered ...
(**the above content** in *The A.B.C. Murders* ...)
Question:
{{ question }}

(b) Context+Question

Context:
Please answer the question based on ***Agatha Christie***'s novel, ***The A.B.C. Murders***.
(**only author and name** of *The A.B.C. Murders*)
Question:
{{ question }}

(c) Question-Only

Context:
Evidence #1: D was the first letter of the killer's next victim.
Evidence #2: Mary Drower's last name begins with the letter D.
(**other Evidence** in *The A.B.C. Murders* about this question ...)
Question:
{{ question }}

(d) Evidence+Question

(a) Illustration on the reasoning metrics. The metric evaluates the ability of the models to recall evidence by comparing the models' answers to the reference steps using GPT4.

(e) Three context settings for the evaluation on DetectiveQA. **Context+Question** is the standard setting that requires models to answer questions based on the full long context. **Question-only** setting prompt without the context, serving for a test of data contamination. **Evidence+Question** provides the golden evidence instead of the long context, aiming to ablate the reasoning ability of the models without long context.

Figure 3: Reasoning metrics with its illustration and three settings.

Step-wise Reasoning Metric For narrative reasoning in DetectiveQA, simply assessing the answer is insufficient; it is important to measure the logical coherence, namely whether LLM presents sufficient evidence and articulates a complete reasoning process behind the correct answer. However, the automated evaluation of long outputs from LLMs remains challenging(An et al., 2023). Fortunately, thanks to the annotation, where each question is linked to corresponding reference steps, we can assess the logical coherence of LLMs' reasoning processes with how many correct reference steps are included, as shown in Figure 3a. Therefore, we introduce a step-wise reasoning metric, the average score across all questions, to reflect the LLM's reasoning performance on DetectiveQA.

To access the containment relationship, we use GPT-4 to review and count the reference steps provided in its responses. The specific prompt used is detailed in Appendix B. Importantly, unlike traditional multi-target information retrieval(Kamradt, 2023; Li et al., 2024) or multi-hop reasoning tasks(Yang et al., 2018; Kociský et al., 2018), DetectiveQA also evaluates whether the model can provide implicit evidence beyond the evidence present in the context, making it a more challenging and realistic evaluation. We can also visualize evidence retrieval with heatmaps in NIAH(Kamradt, 2023), which will be discussed in Section 5.2.

4.2 Context Settings

In addition to two metric settings, we provide three context settings, as shown in Figure 3e, to analyze different issues in LLMs' long-context reasoning. **1.Context+Question** Concatenating the complete context before the answer appears in a detective novel, serving as the basic setting to test LLMs' long-context reasoning capability. **2.Question-Only** Only the title and author of the detective novel are provided before the question. This approach addresses potential data contamination issues since our annotated detective novels are classics and may be included in the LLM's pre-training data. By comparing this setting with the previous one, we can assess whether long-context reasoning truly

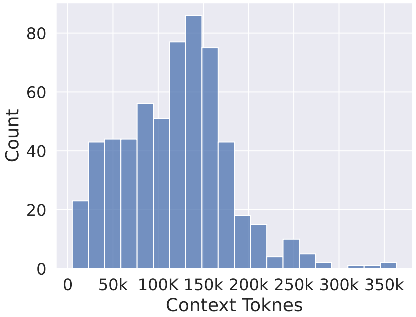


Figure 4: The distribution of the context tokens of samples in DetectiveQA. The novel content for each question is truncated before the answer appears.

relies on contextual information. **3.Evidence+Question** Only relevant evidence from the context is concatenated before the question, transforming long-context reasoning into short-context reasoning. This allows us to compare long-context and short-context reasoning across different LLMs. It is important to note that only the explicit evidence from the context is provided; the implicit evidence from the detective reference steps is not included.

5 Experiment

We conduct experiments to validate the data quality of DetectiveQA and study the capability of prominent large language models with it.

5.1 Validation Results

Statistics. We examine the statistics of DetectiveQA to ensure the distribution of both questions and answers satisfies our desiderata. We confirm that **(1)** DetectiveQA features long context, whose context lengths are distributed from 5k to 363k (Figure 4), with an average of 118k tokens (Table 2); **(2)** The number of clues distributed at each depth exceeded 100, and the exact distribution is shown in the Appendix C; and **(3)** the questions are reasoning-intensive, whose answers involve an average of 8.48 reference steps or near 400 tokens (Table 2). These results support DetectiveQA as a qualified and challenging evaluation for reasoning over long contexts.

Table 2: Statistics of evidence and inferences in DetectiveQA.Our dataset also has richer corpus information in terms of responses. We count the lengths in words. And using the GPT4 tokenizer.

	Max.	**Min.**	**Avg.**
context tokens	363k	5k	118k
reference steps	27	3	8.48
reference tokens	1448	96	394.95

Validation on data annotations. Three authors of this work double-check the data to validate the annotation quality. Specifically, we sample annotated data from 10 novels and the three authors inspect (1) whether the **questions** are reasoning questions whose answers can be derived from the novel; (2) whether the **reference steps** are consistent with the original novel and all the evidence exists in the novel; and (3) whether the annotated **answers** are correct. The results in Table 3 indicate that the quality of almost all the data satisfies our requirement, thus supporting the validity of the outcome dataset. More supporting evidence will be provided in the Appendix C.

Validation on metrics. We verify the reliability of LLM as a judge for the step-wise reasoning metric. Specifically, we manually annotate the reasoning process of 100 answers and calculate the correlation between the judgment from the human judges and GPT4. As shown in Table 4, GPT4 shows high agreement with human judges. In comparison, Rouge, an n-gram-based automatic

Table 3: We asked three authors to score each of the ten novels using agent workflow to accelerate the labeling of questions in terms of question reliability, answer accuracy, and reasonableness of the reference step, and used a voting mechanism to arrive at a final score, and computed the Jaccard correlation between this result and the full 1-sequence.

Validation part	**question**	**reasoning**	**answer**
Jaccard similarity	0.97	0.91	0.94

metric for content recall, despite being much cheaper, fails to correlate with human evaluation well (Table 5). Therefore, we consider it appropriate to apply GPT4 as an automatic evaluation of the reasoning process.

Table 4: Using whether evidence is present in the chain of reasoning as a discrete judgment problem, we analyze the consistency of human judgments and GPT4 judgments using human judgments as the gold standard.

Human vs GPT4	accuracy	kappa
Agreement	0.92	0.83

Overall, the above validation on both data and evaluation metrics ensures DetectiveQA is an evaluation that satisfies our needs for long-context reasoning ability for large language models. We then apply it to evaluate mainstream models and study their capabilities.

5.2 Evaluation Results

Models. We evaluate both closed-source and open-source LLMs featuring long-context capability. For closed-source models, we include GPT4-1106-preview-128k, OpenAI-O1-mini-128k (OpenAI, 2024), Claude3-opus-20240229-200k (Anthropic, 2024b), and KimiChat-200k. For open-source models, we evaluate LLaMA3.1-8B-Instruct-128k (Dubey et al., 2024), ChatGLM3-6B-128k (Zeng et al., 2023a), GLM4-9B-chat-1M (Zeng et al., 2023a), IntermLM2-7B-chat-200k (Cai et al., 2024), InternLM2.5-7B-chat-1M (InternLM, 2024) and Qwen2.5-7B-Instruct-128k (Team, 2024).

Main results. We first report the multiple-choice accuracy and reasoning metric under the *Question+Context* setting in Table 6.

Comparing the metrics, we find the multiple-choice accuracy and reasoning metrics show consistent trends across models, while gaps in numbers exist. This indicates a large number of questions are answered without a perfect reasoning chain, further highlighting the necessity to perform fine-grained evaluation on reference steps to study long-context reasoning capabilities.

Comparing across models, results indicate obvious discrepancies, which we summarize as follows.

- **Open-source models still lag behind esteem closed-source models.** While open-source models have made significant progress in recent development, claiming to approach state-of-the-art closed-source models (Dubey et al., 2024), the performance gaps still exist in long-context reasoning capabilities.
- **Distinction also exists among closed-source models.** Despite being known as a strong reasoner, OpenaAI-O1-mini-128k does not show distinctively superior performance. Instead, Claude3 performed the best among the closed-source models.
- **Most open-source models perform on par,** while Llama3.1 lags behind others. Our subsequent analysis (Section 5.3) attributes this to its failure on over 100k-long contexts.

Analysis on data contamination. The use of detective novels may raise concerns about data contamination issues, which we investigate the potential impact through the question-only setting. The main idea is that the model is able to answer the question correctly even without the context if data contamination occurs. Therefore, we perform a question-wise comparison between the *Question-*

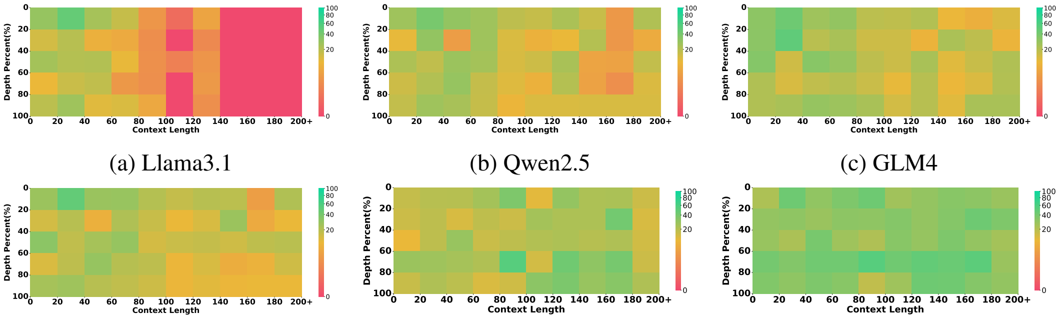
Table 5: Correlation coefficient table. We randomly selected 100 reasoning processes generated by the model and compared them with human-labeled reference steps. For each process, we computed Rouge-1, Rouge-2, and Rouge-L scores, along with human and GPT ratings. The scores for the 100 reasoning processes, based on these four evaluation metrics, were then correlated with the human ratings to determine the correlation coefficients.

	Ours	**Rouge-1**	**Rouge-2**	**Rouge-L**
Corr.	0.91	0.52	0.61	0.58

Table 6: Win rate was calculated for model responses based on the Question Only setting and the Question+Context setting, and G.M. is the geometric mean of the answer accuracy and reasoning scores.

Models	**Question+Context**			**Question-Only**			**Win Rate**
	Answer	Reasoning	G.M.	Answer	Reasoning	G.M.	
GPT-4-1106-preview-128k	73.99	27.43	45.05	43.16	10.99	21.77	84.34
OpenAI-O1-mini-128k	60.83	23.80	38.05	41.67	11.64	**22.03**	70.65
KimiChat-200k	64.13	27.79	42.21	**45.07**	9.64	20.84	67.27
Claude3-Opus-200k	**81.95**	**37.33**	**55.30**	23.43	**16.22**	19.49	**94.61**
LLaMA3.1-8B-Instruct-128k	28.17	21.15	24.41	39.42	8.08	17.84	69.44
GLM3-6B-128k	40.58	22.08	33.63	33.63	7.16	15.51	63.47
GLM4-9B-chat-1M	59.00	**24.07**	**37.68**	40.33	8.06	18.03	74.02
Qwen2.5-7B-Instruct-128k	**61.75**	21.16	36.15	**40.58**	9.09	19.21	76.86
InternLM2-7B-chat-200k	57.95	23.94	37.24	36.97	**12.65**	**21.62**	**81.69**
InternLM2.5-7B-chat-1M	60.92	22.45	36.98	39.17	7.76	17.44	77.99

only and *Question+Context* settings. We summarize the results in terms of win rate[2] in Figure 6a. The results suggest the data contamination issues are mild. The question-only setting merely wins on a small proportion of questions, and the proportions are similar for different models. Besides, as shown in Table 6, our main results on *question+context* settings correlate with the win rate. Therefore, we consider the data contamination is not severe and does not affect the validity of our evaluation with DetectiveQA.



(a) Llama3.1

(b) Qwen2.5

(c) GLM4

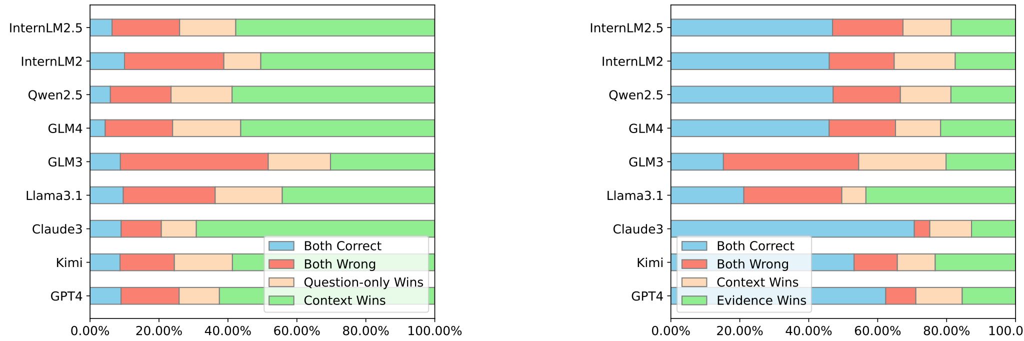
(d) InternLM2.5

(e) GPT4

(f) Claude3

Figure 5: Multi-needle-in-a-haystack test results for different models. We treat each clue in the reference steps found in the article as a "needle" and determine whether the needle is detected by checking if it is included in the model's reasoning process. We define "depth" as the percentage of the problem's total character count where the evidence appears far from the beginning of the document. Our analysis focuses on recall based on varying context lengths and clue depths.

[2] Win rate is defined as the score comparison of each question under two different settings, with scores comprising accuracy (0 or 1) and step-wise reasoning metric (0-1), where the higher score wins.



(a) Stacked bar charts for analyzing data contamination. The figure contrasts model perf. in Q-only vs. C+Q modes using win rate (Sec. 5.2). If both settings yield incorrect answers, they're not compared, categorized as "both lose" for calc.

(b) Stacked bar charts for reasoning analysis. The figure shows model perf. in evidence+Q and context+Q settings, focusing only on answer accuracy, as reasoning score comparison is unfair to context+Q.

Figure 6: Figures of two different analyses.

5.3 Identifying Performance Bottlenecks by Ablating Long Context and Reasoning

DetectiveQA evaluates the intersection of long-context processing and reasoning, two crucial abilities for large language models, simultaneously. To help identify the performance bottlenecks and provide insight to help improve model capability, we disentangle the effect of the two capabilities.

5.3.1 Analysis on evidence retrieval over long context.

To understand whether the evaluated models are proficient in long-context processing, we study the models' success rate in retrieving evidence in different positions over the context of different lengths, inspired by the multi-needle in a haystack task (Li et al., 2024).

Results are in figure 5, which unveils that some models underperform others due to their limitations in long-context processing. For instance, the overall performance of Llama3.1 in our main setting lags behind others. Correspondingly, Figure 5 shows that Llama3.1 almost fails to retrieve any clues when the context length exceeds 100k tokens, while the others still show a descent recall rate. This showcases our annotation on evidence and their position being helpful to ablate the long-context abilities in long-context reasoning.

5.3.2 Analysis on reasoning capability.

For how reasoning affects model performance on DetectiveQA, we study the comparison between *Question+Context* and *Evidence+Context* settings. The latter provides golden evidence instead of the long context, thus eliminating the influence of the long-context capability of the models. Similar to the comparison between *question-only* and *Question+Context* settings, we again conduct question-wise comparisons and summarize the win rates in the two settings, shown in figure 6b.

We summarize our findings from different types of answers, respectively. The **”evidence win”** takes up a number of answers for all the models, indicating the need to enhance long-context capability for all models. Notably, the answers from Llama 3.1 take up a large proportion in this part, aligning with our previous findings on its limitation in long-context processing (Section 5.3.1). Additionally, comparing the proportions of being **”both wrong”**, we find GLM3 stands out to have more answers belonging to this type, implying its less advanced reasoning ability. We also notice there exist a non-neglectable number of cases in **”context win”**, possibly due to that dropping most of the context may harm the understanding of the story thus harming performance[3].

[3] This may relate to the comparison between retrieval-based methods (Xu et al., 2023) and context-based methods for handling massive information. We leave this for future study.

6 Conclusion

We introduced DetectiveQA to test the models' ability to reason narratively over long contexts, the first benchmark for narrative reasoning with an average context length of 100k. We challenged the models' ability to reason over long texts as well as narrative reasoning using detective novels, the real-world texts. For each model, our test gives two scores (answer accuracy and reasoning score) in three settings. With a rich experimental setup, we can deeply analyze the performance of the model and find that the current model still faces challenges in long text comprehension, information extraction and narrative reasoning. We hope that our dataset will facilitate future improvements in model reasoning ability, leading to more robust AI applications and the highest machine intelligence.

Limitations

Our dataset only serves as an evaluation benchmark on long-context reasoning ability, while how to improve the model capability remains an open question. Meanwhile, our benchmark contains only data from detective novels and mainly serves narrative reasoning. More diverse scenarios can be included in the future.

Ethics Statement

We are committed to ensuring that DetectiveQA is used only for academic and scientific purposes, and therefore we have rigorously copyright-checked all of the reasoning novels used in Detective's annotations to ensure that the individual novels are not designed to create copyright problems in non-commercial areas. Through these screening tools, we aim to respect the principle of 'fair use' under copyright protection and ensure that our project navigates within legal and ethical boundaries in a responsible manner.

References

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *CoRR*, abs/2307.11088, 2023. doi: 10.48550/ARXIV.2307.11088. URL https://doi.org/10.48550/arXiv.2307.11088.

Anthropic. Model card and evaluations for claude models, 2024a. URL https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf.

Anthropic. Introducing the next generation of claude, 2024b. URL https://www.anthropic.com/news/claude-3-family.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023a. doi: 10.48550/ARXIV.2309.16609. URL https://doi.org/10.48550/arXiv.2309.16609.

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. Longbench: A bilingual, multitask benchmark for long context understanding. *CoRR*, abs/2308.14508, 2023b. doi: 10.48550/ARXIV.2308.14508. URL https://doi.org/10.48550/arXiv.2308.14508.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting

Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, and et al. Internlm2 technical report. *CoRR*, abs/2403.17297, 2024. doi: 10.48550/ARXIV.2403.17297. URL https://doi.org/10.48550/arXiv.2403.17297.

Maksym Del and Mark Fishel. True detective: A deep abductive reasoning benchmark undoable for gpt-3 and challenging for gpt-4. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pp. 314–322, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zhouhong Gu, Lin Zhang, Xiaoxuan Zhu, Jiangjie Chen, Wenhao Huang, Yikai Zhang, Shusen Wang, Zheyu Ye, Yan Gao, Hongwei Feng, and Yanghua Xiao. Detectbench: Can large language model detect and piece together implicit evidence?, 2024. URL https://arxiv.org/abs/2406.12641.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html.

InternLM. Internlm2.5-7b, July 2024. URL https://huggingface.co/internlm/internlm2_5-7b.

Greg Kamradt. Needle in a haystack - pressure testing llms, 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A" novel" challenge for long-context language models. *arXiv preprint arXiv:2406.16264*, 2024.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328, 2018. doi: 10.1162/TACL_A_00023. URL https://doi.org/10.1162/tacl_a_00023.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*, 2024.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *CoRR*, abs/2311.04939, 2023. doi: 10.48550/ARXIV.2311.04939. URL https://doi.org/10.48550/arXiv.2311.04939.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*, 2024.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

OpenAI. Introducing openai o1, 2024. URL https://openai.com/o1/.

Satomi Saito. Culture and authenticity : the discursive space of japanese detective fiction and the formation of the national imaginary. 2007. URL https://api.semanticscholar.org/CorpusID:190048951.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning, 2024. URL https://arxiv.org/abs/2310.16049.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. Moss: An open conversational large language model. *Machine Intelligence Research*, 2024. ISSN 2731-5398. doi: 10.1007/s11633-024-1502-8. URL https://doi.org/10.1007/s11633-024-1502-8.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. Novelqa: A benchmark for long-range novel question answering. *CoRR*, abs/2403.12766, 2024. doi: 10.48550/ARXIV.2403.12766. URL https://doi.org/10.48550/arXiv.2403.12766.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul F. Christiano. Recursively summarizing books with human feedback. *CoRR*, abs/2109.10862, 2021. URL https://arxiv.org/abs/2109.10862.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*, 2024.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. *CoRR*, abs/2310.03025, 2023. doi: 10.48550/ARXIV.2310.03025. URL https://doi.org/10.48550/arXiv.2310.03025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL `https://doi.org/10.18653/v1/d18-1259`.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL `https://openreview.net/pdf?id=-Aw0rrrPUF`.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL `https://openreview.net/pdf?id=-Aw0rrrPUF`.

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. ∞bench: Extending long context evaluation beyond 100k tokens. *CoRR*, abs/2402.13718, 2024. doi: 10.48550/ARXIV.2402.13718. URL `https://doi.org/10.48550/arXiv.2402.13718`.

{
”question”:”Which of the following is the reason for the disappearance of Sainsbury Seale?”,

”options”:
”A”: ”left voluntarily.”,
”B”: ”met an untimely end.”,
”C”: ”eloped with someone.”,
”D”: ”Sudden memory loss.”

”answer”:”B”,

”reasoning”: [
”Ms Sainsbury-Seal did not take her luggage with her when she disappeared.”,
”This does not appear to be a voluntary departure.”,
”Ms Seale had a dinner date with a friend to play solitaire.”,
”Normally at the appointed time she would have been back at the hotel.”,
”Therefore, based on the above evidences, it is surmised that it was Sainsbury Seale who met an untimely end.”
],

”evidence_position”:[*740*,-1,*734*,-1,-1],

”answer_position”: 1202
}

Figure 7: An example of a multiple-choice annotation in DetectiveQA. We highlight the evidences of reasoning in *blue italics*, and inference in green plain typeface . the ”reasoning” part includes evidences and inferences, while in the ”evidence_position” field, the part corresponding to the evidence will be the paragraph in which the evidence occurs in the article, while the part corresponding to the reference will be -1.

A Annotation Guideline

Data format. For each novel, we require the annotators to annotate (1) several multiple-choice questions involving reasoning with (2) the answers to the question and (3) the multi-step reasoning chains. For each step in reasoning chains, we annotate (4) tags indicating whether the step is a piece of evidence in the original novel and the corresponding position. An example of our annotated data is in Figure 7.

Annotation procedure. We visualize the annotation process in Figure 8. With the collected novels, we first apply our agents to summarize the novels, locating rationales from the detective in the novel and proposing candidate questions for each rationale. The prompts we use are in Fig. 7

With the outputs from the agent, we require the annotators to refine the data with the following procedure.

(i) **On questions.** The annotator should determine whether a question requires reasoning to answer. This involves assessing if answering the question necessitates extracting and synthesizing clues from the original text. Questions that can be answered through simple retrieval of a specific answer from the text, without requiring any additional contextual information, should be filtered out. Additionally, the validity of the question should be evaluated, with manual modifications applied to the original question as needed.

(ii) **On options.** The question must be accompanied by four options, including the correct answer and a minimum of one distractor.

(iii) **On evidence.** The annotators evaluate whether the model-provided evidence is relevant to the question, complete, and accurate. They should eliminate incorrect or irrelevant evidence and supplement missing ones.

(iv) **On answers and rationales.** The rationales should be separated into steps with each step being an evidence or inferences.

(v) **On positions.** The annotators should ensure the answer positions and the evidence positions are correct.

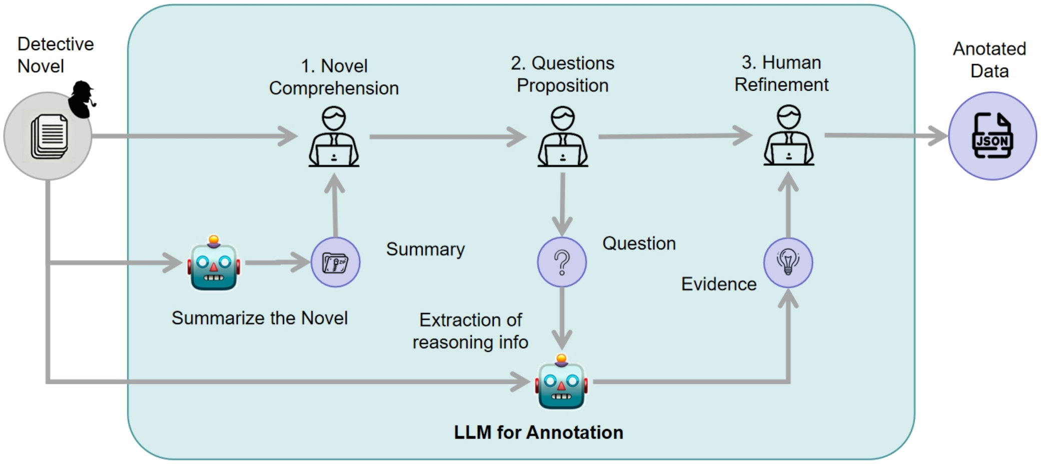


Figure 8: Annotation process

Agent Workflow Figure 8 illustrates the specific process of accelerated annotation through agent workflow. LLM extracts a wide variety of key information, which a human then integrates and constructs into a complete annotation file for the reasoning problem.

B Details in Reasoning Metric

In our step-wise reasoning metric, we mentioned the use of GPT4 for the number of contained leads, and we used the following prompt template 9.We use LLMs such as Kimi, Claude3, and others.

This template asks enough questions to get usable responses without adding additional samples for a few shots to help answer.

C More Validation Results

Table 7 lists the statistics of the novel parts that were manual annotation and accelerated through agent workflow-assisted annotation. While there is an increase in the length of questions for agent workflow-assisted accelerated annotation, the evidence and reasoning sections of manual annotation are more detailed and have slightly less coverage than agent workflow-assisted accelerated annotation. Overall, the difference in quality between manual and agent workflow-assisted accelerated annotation was minimal, suggesting that the use of agent workflow-assisted accelerated annotation is feasible.

Depth Distribution The exact distribution can be seen in Figure 10, where there will be more clues at 100% depth since the longer dependent questions will also have valid clues in the closer locations.

In order to evaluate a question-answering (QA) system's reasoning process regarding a particular inference question, specifically whether the reasoning process correctly includes certain reference steps, multiple reference steps will be provided.

Due to the nature of the inference question being based on a detective novel, the reasoning process may involve some sensitive content. However, for the purpose of this evaluation, please focus solely on determining whether the reasoning process explicitly or implicitly includes the provided reference steps.

The QA system's output for the reasoning process may not explicitly mention the provided reference steps, but it might implicitly incorporate them. In such cases, it should still be considered as correct including the provided reference steps.

The reasoning process output by the QA system and the reference reference steps to be considered are presented below. Please objectively assess whether the QA system's reasoning process explicitly or implicitly includes the provided reference steps, and clearly state which reference steps are included.

Reasoning Process:
[Reasoning Process]

reference steps:
[reference steps]

Provide an initial sentence explaining whether the reasoning process explicitly or implicitly includes each reasoning step. Then, in the second line, specify the indices of the included reference steps in a list format, such as [0, 1, 2, 3, ...].

Your response should maintain this format:

Explanation: ¡One-sentence explanation¿
Included Reference Steps: [Indices of the included reference steps]

Figure 9: GPT-4 Questioning Template: Replace **bolded font** with evidence and Model's Inference Process in Query.

Statistic	**w/ agent workflow (Max/Min/Avg)**	**w/o agent workflow (Max/Min/Avg)**
context length	148k/4k/81k	167k/4k/94k
reference steps	26 / 4 / 7.65	21 / 3 / 7.80
reasoning length	606 / 78 / 251.14	1073 / 92 / 301.94
total questions	308	338

Table 7: Annotation w/ agent workflow vs. Annotation w/o agent workflow Statistics Comparison where context length refers to the length of the problem and the meaning of the remaining metrics is detailed in Section 5.1.

D More Evaluation Results

We have done quite a lot of experiments under the 32K input length model, and we can find that the experimental results of the 32K model are all relatively unsatisfactory. The experimental results are shown in Table 8

However, a larger number of parameters would allow the model to make full use of the information in the 32K text and its own reasoning power to mitigate the problem.

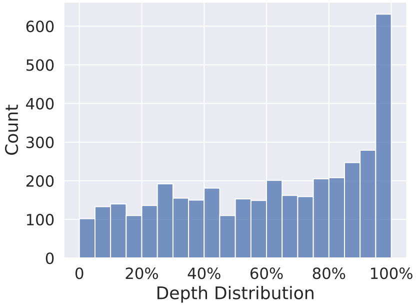


Figure 10: Distribution of different evidence depths. For each piece of evidence, we take as depth the percentage of the total number of words from the beginning of the word in which it occurs

Table 8: Performance of models supporting context lengths of 32K or less

Models	**Question+Context**			**Question-Only**			**Win Rate**
	Answer	Reasoning	G.M.	Answer	Reasoning	G.M.	
LongChat-v1.5-7B-32k	29.33	11.07	18.01	27.83	5.72	12.59	33.58
Vicuna-v1.5-7B-16k	30.33	12.63	19.57	27.67	6.69	13.60	32.57
Qwen1.5-7B-8k	49.50	10.09	22.34	35.33	7.74	16.53	60.71
Qwen1.5-72B-32K	**70.67**	19.69	**37.30**	**44.67**	10.55	**21.70**	76.51