FactGuard: Detecting Unanswerable Questions in Long-Context Texts for Reliable LLM Responses

Anonymous ACL submission

Abstract

Extractive reading comprehension systems are designed to locate the correct answer to a question within a given text. However, a persistent challenge lies in ensuring these models maintain high accuracy in answering questions while reliably recognizing unanswerable queries. Despite significant advances in large language models (LLMs) for reading comprehension, this issue remains critical, particularly as the length of supported contexts continues to expand. To address this challenge, we propose an innovative data augmentation methodology grounded in a multi-agent collaborative framework. Unlike traditional methods, such as the costly human annotation process required for datasets like SQuAD 2.0, our method autonomously generates evidence-based question-answer pairs and systematically constructs unanswerable questions. Using this methodology, we developed the FactGuard-Bench dataset, which comprises 25,220 examples of both answerable and unanswerable question scenarios, with context lengths ranging from 8K to 128K. Experimental evaluations conducted on eight popular LLMs reveal that even the most advanced models achieve only 61.79% overall accuracy. We emphasize the importance of a model's ability to reason about unanswerable questions to avoid generating plausible but incorrect answers. This capability provides valuable insights for the training and optimization of LLMs.

1 Introduction

Comprehending text and answering questions are foundational capabilities in the field of Natural Language Processing (NLP). Over the years, machine reading comprehension has garnered significant attention from both academia and industry (Hermann et al., 2015; Liu et al., 2019). With the rapid advancements of large language models (LLMs)

Paragraph: ...Apple launched the iPhone XS in 2018, and we have a full review of it, including its looks, performance, camera, charging, waterproofing, display, sound, and iOS 12 features and improvements...
Answerable Question: Which Apple 2018 phone is fully reviewed in the article? **Answer:** iPhone XS
Unanswerable Question: Which Apple 2017 phone is fully reviewed in the article? **Plausible Answer:** iPhone XS **Unanswerable Question Detection:** The answer is unknown. **Reasoning Response Generation:** The article does not review any Apple 2017 phone. It reviews the iPhone XS, which was released in 2018.

Table 1: Comparison of Responses to Answerable and Unanswerable Questions.

(Zhao et al., 2023; Liu et al., 2023), retrieval-augmented generation (RAG) has emerged as a promising framework for tackling reading comprehension tasks across diverse specialized domains (Zhao et al., 2024; Lewis et al., 2020). Nevertheless, even state-of-the-art RAG frameworks are susceptible to retrieval accuracy limitations (Hu et al., 2019; Wang et al., 2024), which emphasizes the critical importance of facticity (Jacovi et al., 2025; Bi et al., 2024), i.e., the ability of a model to generate factually consistent and verifiable responses in information-seeking scenarios.

Extracting answers to answerable questions or providing justifications for why certain questions are unanswerable is essential for enhancing the practicality of LLMs. Answerable questions are those that can be resolved using the information present within the given context, whereas unanswerable questions arise when the context lacks sufficient factual support to provide a definitive response. For unanswerable questions, the ideal challenging cases should simultaneously satisfy two critical criteria: high semantic relevance to the contextual topic and the presence of plausible answer candidates that match the expected answer

type. In such cases, generating an appropriate response requires the model to decline to answer, thereby demonstrating its ability to recognize and respect the limitations of the available information. As shown in Table 1, the question "Which Apple 2018 phone is fully reviewed in the article?" can be answered based on factual evidence provided in the passage. In contrast, the question "Which Apple 2017 phone is fully reviewed in the article?" is grounded in an incorrect assumption. An optimal response to the latter question would involve generating a reasoning-based explanation rather than outright refusing to provide an answer. The so-called "Plausible Answer" presented, however, is even more problematic, as it demonstrates a misunderstanding of the context and inadvertently reinforces the misinformation.

Given the recent advancements in LLMs, which have introduced long-context models capable of processing inputs ranging from 32K to 200K tokens (Li et al., 2024a,b), the efficacy of these models in long-context scenarios remains inadequately assessed due to the absence of reliable evaluation benchmarks. The FACTS Grounding leaderboard (Jacovi et al., 2025) offers a manually curated context dataset of up to 32K tokens, primarily assessing models' information-seeking abilities. However, it barely mentions of unanswerable questions. While SQuAD 2.0 (Rajpurkar et al., 2018) specifically focuses on unanswerable questions, like the aforementioned benchmarks, it relies on costly human annotation. Most of the existing datasets have encountered scalability challenges that prevent their cost-effective mass production.

To overcome the above limitations, we propose a novel method that leverages a multi-agent collaboration framework for automated data augmentation. We introduce **FactGuard-Bench**, a dataset of 25,220 questions (8,829 answerable and 16,391 unanswerable), developed through our framework. Even the best-performing model achieves an overall accuracy of 61.79% and performs significantly worse on unanswerable questions compared to answerable ones. Through further training, we achieved an accuracy of 82.39% on an 8B-parameter model. We highlight our contributions as follows:

1. **Innovative Multi-Agent Framework for Data Augmentation**: We introduce **FactGuard**, a multi-agent framework for dynamically generating answerable and unanswerable questions through collaborative multi-step processes, resulting in contextually difficult examples.

2. **Development of Benchmark for Long-Context Evaluation**: We curate **FactGuard-Bench**, a benchmark specifically tailored to assess the ability of LLMs to handle answerable and unanswerable questions within extended contexts.

3. **Limitations of LLMs on Unanswerable Questions**: Experiments with state-of-the-art LLMs show the importance of avoiding hallucinations and generating well-reasoned answers when solving unanswerable questions.

2 Related Work

2.1 Machine Reading Comprehension

Machine reading comprehension (MRC) is a hot research topic in the field of NLP, which focuses on reading documents and answering related questions (Liu et al., 2019; Baradaran et al., 2022).A significant milestone was the introduction of the SQuAD 2.0 dataset by Rajpurkar et al. (2018), which utilized crowdsourcing to annotate unanswerable questions. Datasets like Natural Questions (Kwiatkowski et al., 2019) and TyDi QA (Clark et al., 2020) also provide naturally occurring unanswerable queries, broadening the scope of evaluation. More recently, Kim et al. (2023) explored prompting large language models in the chain-of-thought style to identify unanswerable questions. Deng et al. (2024a,b) proposed self-alignment approach enabling large language models to identify and explain unanswerable questions. In this work, we emphasize scalable and robust evaluation of unanswerable question processing, especially in open-domain scenarios.

2.2 Long Context LLMs and Benchmarks

Recent studies have emphasized the importance of extending positional embeddings to improve the ability of LLMs to handle long contexts effectively (Su et al., 2021; Press et al., 2021; Chi et al., 2022). Closed-source LLMs, in particular, have emerged as leaders in long-context modeling, benefiting from progressively larger context windows. For instance, models such as GPT-4 (Achiam et al., 2023) and Gemini Pro 1.5-1000k (Team et al., 2024) are capable of processing increasingly longer documents, with context lengths

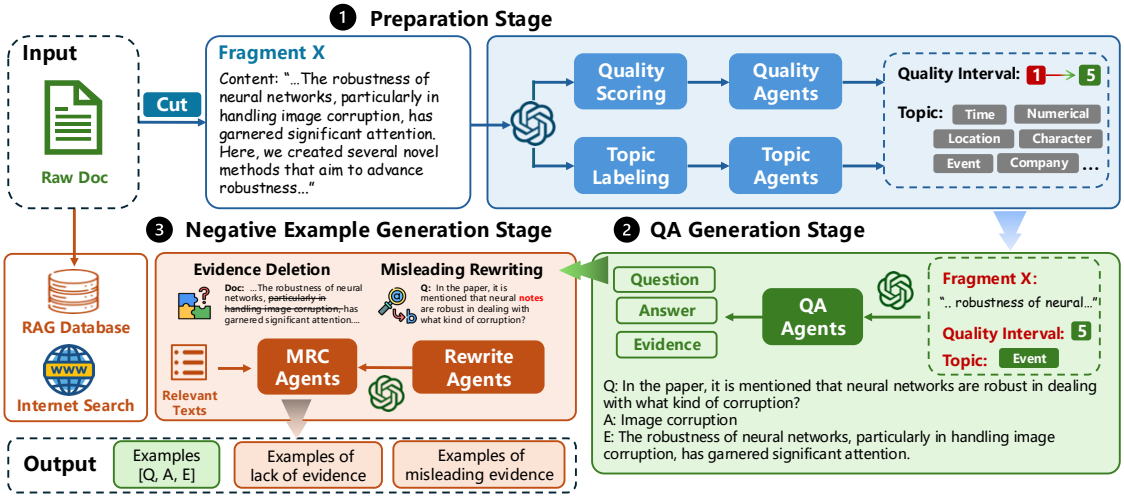

Figure 1: Illustration of FactGuard for data synthesis in a multi-agent collaboration framework.

ranging from 128k to 1000k tokens. Similarly, open-source LLMs, including Qwen 2.5 (Yang et al., 2024a) and DeepSeek (DeepSeek-AI, 2024), also support context lengths of at least 128k tokens. Key benchmarks for assessing long-context capabilities include NIAH (gkamradt, 2023; Yu et al., 2025), Longbench Series (Bai et al., 2023, 2024b), LooGLE (Li et al., 2023), and L-Eval (An et al., 2024), among others. In FactGuard-Bench, we utilize a wider range of context lengths to evaluate the LLM's ability to understand, learn, and reason about information in text.

2.3 Multi-agent Collaboration Frameworks and Data Synthesis

Multi-agent collaboration frameworks are pivotal in enabling cooperative problem-solving among autonomous agents, as demonstrated by the works of Russell and Norvig (2016) and Bai et al. (2024a). The interplay between collaboration and competition in interactive environments has been further explored by Bakhtin et al. (2022), and Hong et al. (2023) investigate the integration of human practices with multi-agent systems. From a data synthesis perspective, methodological innovations have emerged to address scalability and fidelity challenges. Mitra et al. (2024) propose AgentInstruct, a framework that operationalizes agent-generated interaction streams for synthetic data creation, emphasizing iterative self-refinement. Similarly, Long et al. (2024) formalize a generalized workflow for large-scale synthetic data generation using LLMs, identifying faithfulness and diversity as critical challenges in the process. Recently, Moradisani et al. (2024) examines automatic synthesis of unanswerable questions, although the input relies on the triples unit of the MRC dataset. The consensus is that efficient synthesis of data remains worth exploring.

3 FactGuard Methodology

As shown in Figure 1, our method consists of three primary stages: preparation, QA generation, and negative example generation. The agent console is responsible for aggregating the opinions of each agent and making decisions about next steps.

3.1 Preparation Stage

The preparation stage involves the selection of short text segments from extensive documents. The window size is kept at [500, 1000] and splicing is done on a paragraph by paragraph basis. We randomly select Fragment X for the following sub-steps:

- **Quality Scoring:** Using quality agents, we evaluate Fragment X in terms of fluency, coherence, and logicality, assigning a quality score on a 5-point scale ($score_i \in [1, 5]$).

- **Topic Labeling:** Topic agents are employed to extract structured information (e.g., temporal expressions, numerical values, locations, organizations, and events) from Fragment X. This process facilitates downstream tasks such as question generation and entity replacement.

3.2 QA Generation Stage

Leveraging QA generation agents, we generate questions, answers, evidence based on fragments and topics. Note that evidence consists of specific

Reasoning	Description	Example
Lack of Evidence	The question is related to the article, but the factual basis is deleted.	**Fragment:** ...There had been a lack of confidence in Murray since Romani, and the two failed Gaza battles increased his unpopularity among both the infantry and the mounted troops. ~~After the war Allenby acknowledged Murray's achievements in a June 1919 despatch in which he summed up his campaigns~~... **Question:** According to this article, in what year did Allenby recognize Murray's accomplishments in his circular? **Answer:** The question cannot be answered. The article mentions Murray's performance in the battle, but does not mention what year Allenby recognized his accomplishments.
Misleading Evidence	The key information of the question is misaligned against the facts of the article.	**Fragment:** Global and Local Mixture Consistency Cumulative Learning (GLMC) for Long-Tailed Visual Recognition...The paper introduces GLMC, a one-stage training strategy designed to improve long-tailed visual recognition by enhancing the robustness of the feature extractor and reducing the bias of the classifier towards head classes. GLMC uses a global and local mixture consistency loss and a cumulative head-tail soft label reweighted loss... **Raw Question:** What are the core ideas behind the Global and Local Mixture Consistency cumulative learning (GLMC) framework and how does it improve long-tailed visual recognition? **New Question:** What are the core ideas behind the Global and Local Augmentation Consistency Learning (GLACL) framework and how does it improve long-tailed visual recognition? **Answer:** The article focuses on GLMC and does not mention GLACL. The core ideas of GLACL cannot be answered, but about GLMC...

Table 2: Categorization of Negative Examples in FactGuard-Bench: A detailed overview of reasoning errors, including *Lack of Evidence*, where factual bases are missing, and *Misleading Evidence*, where key information is misaligned with the article's content.

text segments that substantiate the answer. This design ensures that each question is firmly grounded in the provided context. Since there are low-quality results for LLM generation, such as questions that are not fluent, we filter them for quality judgment after QA generation.

3.3 Negative Example Generation Stage

We synthesize the data mimicking the real-world **Negative Rejection** scenario. This involves two distinct approaches:

- **Contextually Missing Negative Example Generation:** We simply remove the evidence from the text, thus making the question unanswerable due to lack of information.
- **Misleading Negative Example Generation:** To create misleading questions, question rewriting agents perform entity substitutions, impossible condition insertions, and other types of false assumptions.

We have streamlined the review process for the generated data by employing Retrieval Augmented Generation (RAG) techniques. This approach allows us to extract the first N relevant passages from a lengthy article for short-reading comprehension and to filter out data that contain conflicting answers. By using the RAG mechanism, we enhance the likelihood of early detection of conflicting questions, thereby improving efficiency. Furthermore, we employ the World Wide Web to filter common-sense knowledge, effectively circumventing the inherent conflict between context-faithfulness and common-sense accuracy.

Remark These agents, inspired by multi-agent systems in distributed AI (Ferber and Weiss, 1999), function as independent decision-makers, assessing and processing inputs in parallel to optimize the preparation pipeline. The modularity of this approach ensures that updates or improvements to one agent's algorithms do not disrupt the system's overall functionality, thereby providing robustness and adaptability. FactGuard ensures the generation of high-quality, contextually relevant answerable and unanswerable questions. The multi-agent collaboration framework not only enhances the efficiency of the data augmentation process but also significantly improves the diversity and complexity of the generated datasets.

4 Benchmark Constructions

FactGuard dynamically generates answerable and unanswerable questions by leveraging a multi-agent collaboration process. We collect raw, lengthy texts from the open-source community as the initial input for our process. These texts

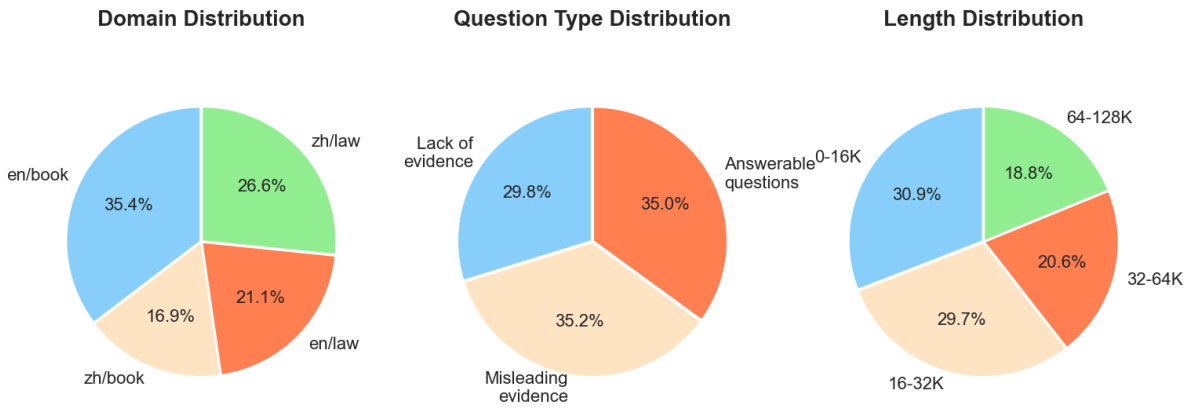

Figure 2: Distributions of FactGuard-Bench in terms of domain, question type and length.

	FactGuard-Bench		
	En	**Zh**	**Total**
Train			
Total examples	10,699	8,401	19,100
Total articles	5,730	5,649	11,379
Development			
Total examples	1,140	780	1,920
Total articles	1,056	729	1,785
Test			
Total examples	2,400	1,800	4,200
Total articles	2,072	1,506	3,578

Table 3: Dataset statistics of FactGuard-Bench.

cover both Chinese and English languages and span domains such as law and books. Specifically, the datasets include legal datasets such as Pile of Law (Henderson et al., 2022), Tiger Law (Chen et al., 2023), the book dataset Gutenberg [1], open-copyright Chinese books, and so on.

The model underlying the whole process is Qwen2.5-72B-Instruct (Yang et al., 2024b). By incorporating a variety of syntactic and semantic modifications to the original context, FactGuard ensures that the negative examples remain linguistically plausible but ultimately unanswerable. As shown in Table 2, for examples lacking evidence, we remove the evidence from the original Fragment. For examples with misleading evidence, the Fragment remains unchanged, but we rewrite the questions to include false assumptions. We further analyzed the generation process for misleading negative examples in one iteration, finding that FactGuard's data processing pipeline discarded approximately 28% of the candidate examples due to quality control measures.

[1] www.gutenberg.org

4.1 Characteristics

FactGuard-Bench includes 25,220 data examples generated from 16,742 texts. Detailed information regarding FactGuard-Bench is presented in Table 3 and illustrated in Figure 2. The dataset includes English (en) and Chinese (zh) across two domains, law and books, and features two types of questions: answerable and unanswerable. Unanswerable questions are either due to a lack of evidence (Contextually Missing Negative Examples) or misleading evidence (Misleading Negative Examples). Example lengths range from 8K to 128k tokens.

4.2 Manual Review

To verify the quality of the synthetic data, we randomly sampled 480 examples for manual review. We hired three people on a crowdsourcing platform to perform the annotation and used a voting method to select the final answer. We asked each annotator to spend a maximum of 10 minutes reading the text and evaluating each example.The results are shown in Table 4. The lower quality in the misleading evidence category was due to the omission of clarifications during the synthesis of answers, as the relevant instructions were not followed. However, the overall quality of 93.96% indicates the high value of our method.

QA class	Answerable	Unanswerable	
		Lack of evidence	Misleading evidence
Number	120	120	240
Quality(%)	94.17	93.89	
		96.67	92.50
Overall quality(%)	93.96		

Table 4: Manual review of synthetic data quality.

Model	FactGuard-Bench Test						
	Overall	En			Zh		
		Answerable questions	Lack of evidence	Misleading evidence	Answerable questions	Lack of evidence	Misleading evidence
GPT-4o (20240806)	49.68	86.72	48.90	49.43	**87.33**	39.53	37.14
DeepSeek-V3-0324	45.23	**89.71**	34.09	40.04	85.06	38.09	33.50
Gemini1.5-Pro (202409)	58.20	86.25	54.60	59.61	83.05	45.45	50.81
Mistral-Large-Instruct-2411	47.07	87.25	57.17	51.61	83.33	30.43	22.38
Qwen2.5-72B-Instruct	61.79	86.25	63.34	63.16	85.00	50.12	50.76
Qwen2.5-7B-Instruct	50.60	80.50	57.45	53.43	78.33	40.93	32.10
Llama-3.3-70B-Instruct	44.04	85.50	49.42	48.00	84.33	27.45	21.43
Llama-3.1-8B-Instruct	41.21	82.00	58.35	41.20	82.67	31.28	13.14
+ with sft	77.91	83.25	72.08	83.32	69.67	**86.31**	74.19
+ with sft&dpo	**82.39**	82.50	**79.93**	**88.84**	77.00	77.54	**82.08**

Table 5: Prediction accuracy on the test set of FactGuard-Bench. Note that unanswerable questions include lack of evidence and misleading evidence.

5 Experiments

5.1 Implementation Details

To evaluate the ability of LLMs on FactGuard-Bench, our experiments included several open-source models that have been instruction-tuned using Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) and Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020; Bai et al., 2022). Specifically, we utilized the following open-source models: Mistral-Large-Instruct-2411 (123B) (Jiang et al., 2024), DeepSeek-V3-0324 (685B) (Liu et al., 2024), Llama3.1-8B-Instruct and Llama3.3-70B-Instruct (Dubey et al., 2024), Qwen2.5-7B-instruct and Qwen2.5-72B-instruct (Yang et al., 2024a). We also obtained evaluation results through API calls for several proprietary models. These included GPT-4o[2] from OpenAI (Achiam et al., 2023), Gemini1.5 Pro (GeminiTeam, 2024). Please note that we provide the operational URL addresses of these proprietary models and document the version numbers used in our experiments to ensure reproducibility.

We utilize full-parameter SFT and DPO (Rafailov et al., 2024) training on Llama3.1-8B-Instruct to enhance the model's ability to verify the validity of the dataset. We utilized the AdamW optimizer, setting the learning rate to 2×10^{-5} with 1 epoch and 5×10^{-7} for full-paramenter SFT and DPO respectively. We set the warm-up ratio to 0.1 and the weight decay to 0.1. Additionally, the low-quality responses used in the DPO experiments were selected from the generated results of the baseline models.

[2]https://openai.com/index/gpt-4o-system-card/

5.2 Evaluation Settings and Metrics

We consider two evaluation tasks aimed at assessing different aspects of the model's capabilities: (1) the consistency of the predicted answers with the ground truth, and (2) the reasoning ability of the model when handling unanswerable questions.

Task 1: Answer Consistency Evaluation We adopt accuracy (ACC) as the evaluation metric, instead of metrics such as Exact Match (EM) and F1 (Rajpurkar et al., 2018), which require threshold tuning. Leveraging the discriminative capabilities of LLMs (Chan et al., 2023), our evaluation differentiates between answerable and unanswerable questions. For answerable questions, a prediction is assigned a score of 1 if it contains the correct information fragments from the ground truth; otherwise, it is scored 0. For unanswerable questions, responses are assigned a score of 1 if they appropriately recognize the unanswerable nature of the question (e.g., through rejection), and a score of 0 if they generate hallucinatory content.

Task 2: Reasoning Ability for Unanswerable Questions We evaluate the model's ability to refuse to answer unanswerable questions and to avoid generating misleading content. Specifically, we investigate whether the model outright rejects the question or provides supplementary reasoning, such as error correction or clarification, which serves as an indicator of its reasoning proficiency. We employ LLMs to categorize responses into three distinct types: *incorrect answers*, *correct answers - direct refusals*, and *correct answers - reasoned answers*. The evaluation metric for this task is the proportional distribution of each response type. When the model provides explanations for why a question is unanswerable (Liao et al., 2022;

Model	Answerable questions				Lack of evidence				Misleading evidence			
	0-16K	16-32K	32-64K	64-128K	0-16K	16-32K	32-64K	64-128K	0-16K	16-32K	32-64K	64-128K
GPT-4o (20240806)	90.86	85.43	85.06	**85.91**	55.12	42.80	38.20	37.99	45.85	45.60	44.05	40.19
DeepSeek-V3-0324	**93.74**	**87.50**	**85.44**	81.75	47.92	33.77	26.06	28.17	44.18	41.01	32.55	28.55
Gemini1.5-Pro (202409)	86.78	83.33	83.77	86.57	58.18	45.21	46.81	57.53	60.20	55.06	53.03	53.31
Mistral-Large-Instruct-2411	91.37	85.00	81.82	82.52	56.12	44.88	36.02	42.69	44.75	41.48	38.00	29.75
Qwen2.5-72B-Instruct	88.32	85.00	85.06	83.89	62.10	56.16	53.14	58.13	60.73	58.66	55.77	55.09
Qwen2.5-7B-Instruct	86.80	76.50	75.97	77.85	58.94	46.64	47.45	45.31	44.40	44.46	44.42	43.77
Llama-3.3-70B-Instruct	88.32	84.50	83.77	82.55	53.92	39.68	31.03	27.59	45.63	38.64	34.22	24.57
Llama-3.1-8B-Instruct	85.79	82.50	82.47	77.18	55.78	45.74	40.93	41.58	32.70	28.41	28.73	26.04
+ with sft	80.20	80.50	77.27	69.80	85.53	75.46	75.71	73.33	80.61	76.99	78.98	81.47
+ with sft&dpo	83.76	82.00	80.52	72.48	**84.08**	**75.74**	**77.73**	**76.56**	**86.90**	**84.37**	**87.88**	**84.85**

Table 6: Prediction accuracy of different length intervals on the test set of FactGuard-Bench.

Lee et al., 2020) —such as insufficient context, inherent ambiguity, etc.—it not only demonstrates a more sophisticated understanding of its own limitations but also enhances user trust by transparently communicating the boundaries of its capabilities. This reasoned refusal is superior to a simple direct rejection, as it helps users refine their queries or adjust their expectations.

Remark We selected Qwen2.5-72B-Instruct (Yang et al., 2024b) as the discriminant model for our experiments. The accuracy of LLM-based evaluation is about 94% after manual evaluation, and more details will be discussed in the Appendix.

5.3 Experimental Results

5.3.1 Answer Consistency Evaluation

The evaluation of answer consistency on the FactGuard-Bench test set is presented in Table 5. The analysis distinguishes between answerable and unanswerable questions, with the latter further divided into lack of evidence and misleading evidence categories. The highest overall accuracy observed is 82.39%, achieved by the model augmented with both SFT and DPO. It is evident from the results that while baseline models perform well on answerable questions, their performance on unanswerable questions is suboptimal. For instance, Qwen2.5-72B achieves an 86.25% accuracy on answerable questions but only manages 63.34% and 63.16% on lack of and misleading evidence, respectively. This highlights a significant performance gap and accentuates the limitations of LLMs in handling unanswerable queries, thereby justifying the necessity of the FactGuard-Bench.

Notably, the implementation of SFT significantly improved the model's performance on unanswerable questions, achieving a 61.05% improvement in handling misleading evidence in Chinese. However, this improvement came at the cost of the model's performance on answerable Chinese questions, which dropped from 82.67% to 69.67% after applying SFT. The subsequent application of DPO alleviated this issue, bringing the performance on answerable Chinese questions back up to 77.00%, thereby achieving an overall performance enhancement. Specifically, the overall performance improved from 41.21% before SFT to 77.91% after SFT, and further to 82.39% after DPO. This demonstrates the efficacy of combining SFT and DPO. Moreover, although Table 4 indicates the presence of noise in our synthesized dataset, it still holds learning value, particularly when contrasted with the relatively poor performance metrics observed on unanswerable questions.

Table 6 presents a detailed analysis of model performance across varying text length intervals. The results indicate that the majority of models achieve optimal performance on shorter texts (0–16K tokens), with a discernible decline in efficacy as text length increases. Notably, models fine-tuned via SFT using FactGuard-Bench's training data exhibit substantial improvements in addressing unanswerable questions, consistently outperforming baseline systems across all length categories. These findings underscore the utility of FactGuard-Bench in enhancing model robustness and highlight its effectiveness as a benchmark for advancing evaluation and development in this domain.

5.3.2 Reasoning Ability Evaluation

Table 7 presents the evaluation of reasoning capabilities, with a specific focus on model performance for unanswerable questions. The results reveal a consistent pattern among baseline models: a predominant tendency to generate incorrect answers rather than employing refusal mechanisms or providing reasoned responses. However, the application of SFT and DPO yields significant improvements, not only enhancing response accuracy but also substantially increasing the rates of reasoned answers. Here, the optimal model variant achieves

Model	Incorrect ↓ answers	Correct answers ↑ direct refusals	Correct answers ↑ reasoned answers
GPT-4o (20240806)	57.77	11.31	30.91
DeepSeek-V3-0324	63.07	7.70	29.22
Gemini1.5-Pro (202409)	47.11	11.96	40.93
Mistral-Large-Instruct-2411	60.60	12.02	27.39
Qwen2.5-72B-Instruct	43.00	16.52	40.48
Qwen2.5-72B-Instruct	55.20	13.80	31.00
Llama-3.3-70B-Instruct	64.15	10.23	25.61
Llama-3.1-8B-Instruct	67.01	12.58	20.41
+ with sft	21.99	22.45	55.56
+ with sft&dpo	**17.16**	**22.71**	**60.14**

Table 7: Percentage breakdown of unanswerable question types in the FactGuard-Bench test set. The three categories sum to 100%, with lower incorrect proportions and higher correct proportions indicating better performance.

a 60.14% rate of reasoned answers, demonstrating a more sophisticated ability to recognize unanswerable questions.

These findings indicate that our method: (1) improves relevance discrimination - enabling models to better identify when questions, while topically related, lack substantiating evidence in context; and (2) strengthens type-matching awareness - preventing false positives when no elements that satisfy the required answer type. Elements here include numerical values, locations, organizations, etc.

6 Discussion

FactGuard enables flexible generation of answerable and unanswerable questions. With the goal of detecting and processing unanswerable questions, FactGuard introduces a paradigm shift in the evaluation and enhancement of long-context machine reading comprehension. Similar to how SQuAD 2.0 compels models to determine whether a question can be answered given a contextual passage (Rajpurkar et al., 2018), FactGuard-Bench extends this challenge to significantly longer contexts, pushing the boundaries of current LLMs. As shown in Table 8, models trained with FactGuard-Bench were predicted on the dev set of SQuAD 2.0 and showed significant improvements in their overall metrics, especially in handling unanswerable questions. However, we observe a performance dip in answerable questions, suggesting a trade-off between optimizing for answerable and unanswerable scenarios. This phenomenon highlights the inherent challenge of balancing the dual objectives of question-answering systems: maintaining high accuracy for answerable questions while simultaneously improving detection of unanswerable ones. Future work may explore adaptive weighting mechanisms or auxiliary training objectives to better harmonize these competing demands.

Model	Overall	answerable	unanswerable
Llama-3.1-8B-Instruct	73.22	**93.55**	52.95
+ with sft	80.34	80.83	**79.85**
+ with sft&dpo	**80.91**	86.45	75.37

Table 8: Prediction accuracy on the dev set of SQuAD 2.0.

7 Conclusion and future work

In this paper, we presented FactGuard, a multi-agent framework for dynamically generating both answerable and realistic unanswerable questions, alongside FactGuard-Bench, a meticulously curated benchmark designed to evaluate LLMs' performance in long-context information extraction. Our key contributions include: (1) the development of an innovative multi-agent data synthesis pipeline that simulates real-world unanswerability by modeling ambiguity and contextual irrelevance; (2) the creation of a long-context evaluation benchmark to assess LLM robustness; and (3) empirical evidence highlighting the persistent limitations of state-of-the-art LLMs in handling unanswerable questions, underscoring unresolved challenges in machine reading comprehension.

Future work will focus on two key directions to advance this research frontier. First, we will enhance FactGuard by integrating adaptive multi-agent collaboration strategies and advanced data augmentation techniques to generate even more nuanced and adversarial unanswerable questions. Second, a crucial long-term research objective involves improving the model's capability to accurately identify unanswerable questions while preserving the performance on answerable questions.

8 Limitations

First, limited by the automated process, all synthetic datasets still have a certain percentage of noise. Second, due to the limitations of available resources, we have to admit that we cannot scale our experiments to larger models. For example, claude3.5 (Anthropic, 2024) is limited by the security policy of the API. Our training experiments are only performed on a widely adopted 8B open-source LLM (i.e., Llama-3.1-8B).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Tianyi Bai, Ling Yang, Zhen Hao Wong, Jiahui Peng, Xinlin Zhuang, Chi Zhang, Lijun Wu, Jiantao Qiu, Wentao Zhang, Binhang Yuan, and 1 others. 2024a. Multi-agent collaborative data selection for efficient llm pretraining. *arXiv preprint arXiv:2410.08102*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2024. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. *arXiv preprint arXiv:2404.00216*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023. Tigerbot: An open multilingual multitask llm. *Preprint*, arXiv:2312.08688.

Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. 2022. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024a. Don't just say "i don't know"! self-aligning large language models for responding to unknown questions with explanations. Association for Computational Linguistics.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024b. Gotcha! don't trick me with unanswerable questions! self-aligning large language models for responding to unknown questions. *arXiv preprint arXiv:2402.15062*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jacques Ferber and Gerhard Weiss. 1999. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-wesley Reading.

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

gkamradt. 2023. Needle in a haystack - pressure testing llms.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *arXiv preprint*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read+ verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6529–6537.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, and 7 others. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *Preprint*, arXiv:2501.03200.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Najoung Kim, Phu Mon Htut, Samuel R Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. Squad2-cr: Semi-supervised annotation for cause and rationales for unanswerability in squad 2.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5425–5432.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 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. 2023. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. 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)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

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

Jinzhi Liao, Xiang Zhao, Jianming Zheng, Xinyi Li, Fei Cai, and Jiuyang Tang. 2022. Ptau: Prompt tuning for attributing unanswerable questions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1219–1229.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, and 1 others. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei-ge Chen, Olga Vrousgos, Corby Rosset, and 1 others. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

Hadiseh Moradisani, Fattane Zarrinkalam, Julien Serbanescu, and Zeinab Noorian. 2024. Unanswgen: A systematic approach for generating unanswerable questions in machine reading comprehension. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 280–286.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ofir Press, NoahA. Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv: Computation and Language,arXiv: Computation and Language*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Pearson.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *Cornell University - arXiv,Cornell University - arXiv*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Key Prompts

Example of topic labeling:

```
## Text begins
{example.document}
## Text ends
Please output the text of the main one entity, the entity type can be time, place, event, weather, items, etc. Return in json format:
{
  "original_entity": str,
  "entity_type": str
}
```

Example of Q generation:

```
## Text begins
{example.document}
## Text ends
Please ask a question based on the original text, with "{original_entity}" in the question text. And cite the original text as the evidence for your question.
Return the result in JSON format:
{
  "question": str,
  "evidence": str
}
```

Example of rewriting entity discovery:

```
## Text begins
{example.document}
## Text ends
Replace the "{original_entity}" in the text with a similar entity, ensuring the following:
1. The replacement entity does not already appear in the text.
2. The replacement entity is of the same type and semantically similar to the original.
3. The replacement entity should be as lexically (word-form) close as possible to the original.
4. The replacement entity should clearly not be the original entity.
Return the result in JSON format:
{
  "entity": str
}
```

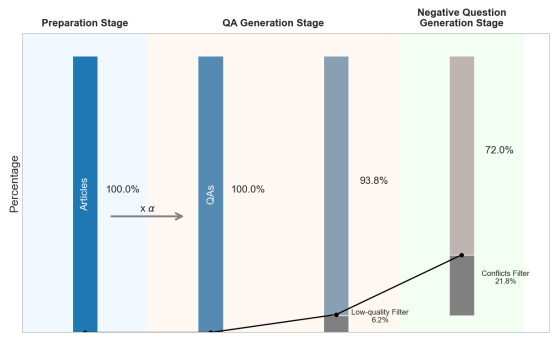

Figure 3: For misleading negative example generation, the percentage of attrition in FactGuard's data processing program.

Task 1: Answer Consistency Evaluation.			
QA class	Answerable question	Lack of evidence	Misleading evidence
Number	80	60	160
Quality(%)	95.00	93.33	93.75
Overall quality(%)	94.00		
Task 2: Reasoning Ability for Unanswerable Questions.			
Answer class	Incorrect answers	Direct refusals	Reasoned answers
Number	60	111	49
Quality(%)	93.33	91.89	97.96
Overall quality(%)	93.64		

Table 9: Manual review results of judgment quality by the discriminative model on Qwen2.5-72B response answers.

B Data Synthesis Efficiency

As an example, the efficiency of each stage of the data synthesis process for misleading data is shown in Figure 3. During the preparation stage, the ratio between the amount of raw textual data and the number of selected segments is defined by a configurable parameter α. In this experiment, α=1, meaning one segment is extracted from each article to generate a single QA pair. By adjusting α, multiple segments can be selected to generate multiple QA pairs. In the subsequent QA generation stage, the total number of generated QA pairs after filtering was decreases by about 6.2% due to noise in the generation process, such as poorly organized statements and incomplete answers. During the stage of generating negative examples, a post-processing review procedure is applied following the initial agent's processing. This review process removes questions that fail to meet the requirements, including those related to questions with conflicting answers in different locations and context-independent common sense, resulting in a reduction of approximately 21.8% in the number of examples.

C LLM-Based Evaluation

To ascertain the reliability of the discriminative model employed in our evaluation, we randomly selected 300 samples for manual review based on the discriminant model's results of discriminating Qwen2.5-72B answers from standardized answers. Consistent with our approach to validating synthetic data quality, we employed a three-person voting mechanism. The outcome of this manual review is detailed in Table 9.

In **Task 1: Answer Consistency Evaluation**, human annotators evaluated whether the discriminative model accurately identified the consistency between its predictions and the ground truth for answerable and unanswerable questions. The results demonstrate that the discriminative model achieved a commendable accuracy of **94.00%** in Task 1.

In **Task 2: Reasoning Ability for Unanswerable Questions**, the manual review focused on whether the discriminative model could accurately classify responses into three distinct categories: *incorrect answers*, *direct refusals*, and *reasoned answers*. The evaluation revealed that the model achieved an overall classification accuracy of **93.64%**. However, due to subtle or ambiguous rejection/clarification phrasing, the model produced more false negatives than false positives. Although slightly outperformed by human benchmarks, the automated system excels in efficiency, consistency, and scalability, enabling robust iterative refinement.

D Case Study

To facilitate a clear and intuitive comparison of various models for generating reasoning-based answers to both answerable and unanswerable questions, we present three distinct scenarios in Figure 4. In the answerable scenario, all models exhibit a high degree of accuracy in identifying the correct answers. However, the answers generated by Mistral-Large, GPT4o and Llama3.1-8B, among others, often appear superficial and lack supporting evidence derived from the original text.

In contrast, Qwen2.5-72B and the fine-tuned versions of Llama3.1-8B produce more comprehensive and satisfactory answers. In the second scenario, GPT4o and Llama3.1-8B display significant hallucination in their responses, frequently generating factually incorrect answers. Qwen2.5-72B had both rejection tendencies and reasoning, making it a highly desirable response. In the third scenario, all baseline models are misled by the question, re-

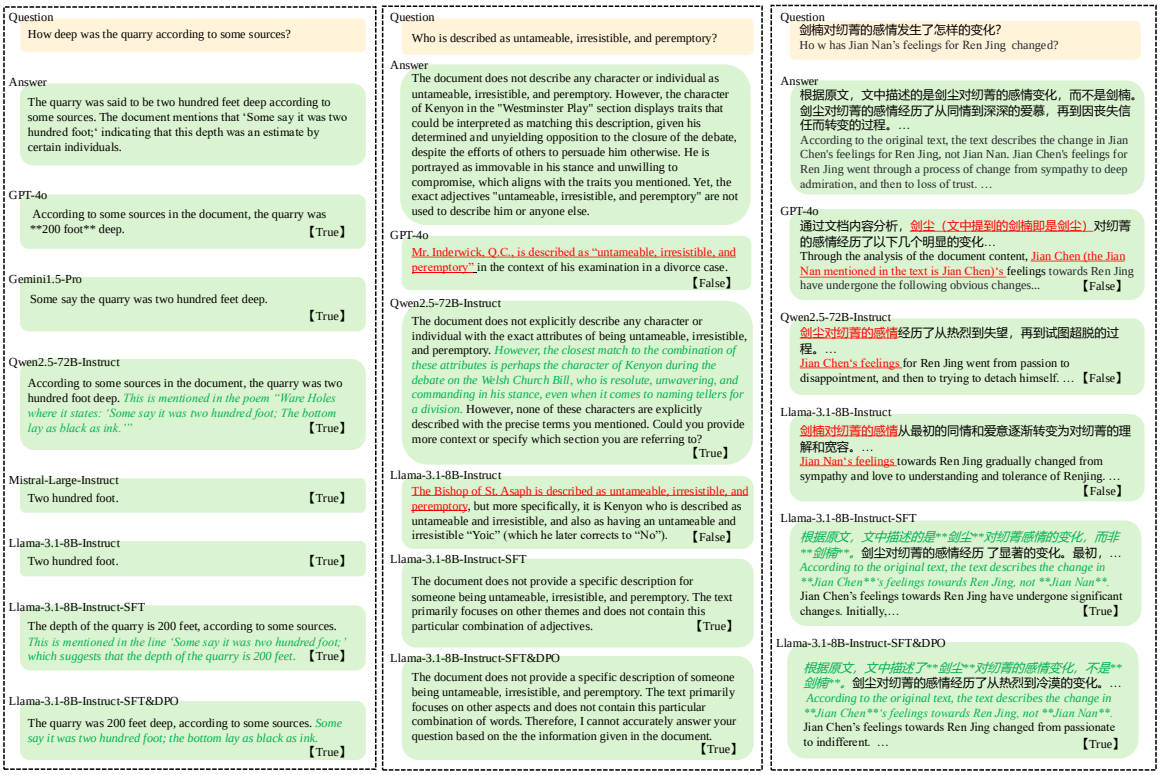

Figure 4: Case study. An examples of answerable questions in English on the left, an example of lack of evidence in English in the center, and an example of misleading evidence in Chinese on the right (translated below). Red underlined text indicates hallucinatory content and green italicized text indicates useful explanations.

sulting in incorrect answers. However, after fine-tuning with SFT and DPO, this issue is mitigated, enabling the models to provide accurate responses that align with the given text.