Abg-SciQA: Benchmarking and Enhancing Ambiguity Detection and Clarification in Language Models

Anonymous ACL submission

Abstract

Asking ambiguous questions is a natural aspect of human communication, making it essential for Large Language Models (LLMs) to effectively recognize and address ambiguities. However, there is a lack of a comprehensive analysis of how well LLMs detect and solve ambiguities. Besides, though there exist several datasets on ambiguity, the absence of explicit explanations of ambiguity and annotations of ambiguity types limits the comprehensive evaluation. To address this issue, we introduce Abg-SciQA, a dataset designed to evaluate and help LLMs detect ambiguities and generate appropriate clarification questions using challenge questions in the area of social and nature science. Abg-SciQA encompasses four tasks: Ambiguity Detection, Ambiguity Type Classification, Clarification Question Generation, and Clarification-Based Question Answering, where each task has corresponding annotations. We evaluate the dataset using both closed-source and open-source LLMs and fine-tune it on open-source LLMs. Our experiments show that the most state-of-the-art LLMs still encounter difficulties in resolving ambiguity in natural questions, and fine-tuning on Abg-SciQA can significantly enhance their capabilities to understand and address ambiguities. Notably, in the Ambiguity Detection task, the F1 score of Llama2-7b improves significantly from 16.6% to 79.1%. On the other hand, Abg-SciQA remains a challenging benchmark for LLMs, revealing ample room for model improvement. Our dataset can be found here [1].

1 Introduction

Large Language Models (LLMs) have become widely used in various applications, including conversational systems (Achiam et al., 2023), code generation (Du et al., 2024), and optimization (Yu et al., 2023). However, LLMs often face challenges when dealing with ambiguous questions—questions that can have multiple interpretations or unclear meanings. For the first example in Fig. 1, the question "What?" could refer to "What did he do for the children?", or simply asking for a repeat for the previous answer, depending on the context. Such ambiguity makes it difficult for LLMs to provide accurate answers, as they may exhibit overconfidence in their responses (Xiong et al., 2023). Given that ambiguous questions are common in natural human communication (Clark and Brennan, 1991), addressing this issue is crucial for improving LLM performance and reliability.

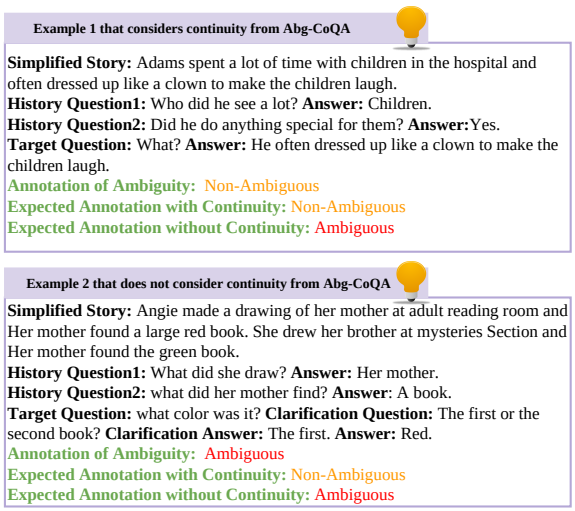

Figure 1: An example of quality issues in previous datasets. Abg-CoQA (Guo et al., 2021) follows a conversational format. The first question considers dialogue continuity, resulting in an unambiguous query, while the second question lacks this consideration, leading to ambiguity. This inconsistency may confuse both humans and LLMs regarding the dataset's standards.

To address ambiguity in natural language processing, researchers have focused on generating clarification questions as a key strategy. Language models are often employed to automatically generate these questions to resolve ambiguities (Zamani et al., 2020; Deng et al., 2023), while other

[1] https://anonymous.4open.science/r/Abg-Sci-DF10/README.md

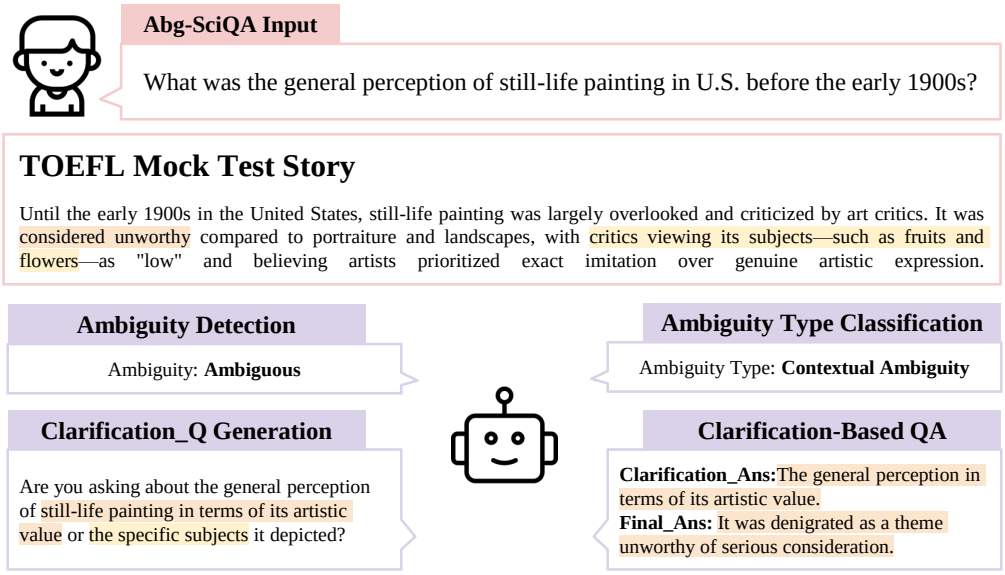

Figure 2: An example of a data sample in Abg-SciQA. Each sample in Abg-SciQA includes a story and a corresponding question, covering four tasks: 1) Ambiguity Detection, 2) Ambiguity Type Classification, 3) Clarification Question Generation, and 4) Clarification-Based Question Answering. Unlike previous datasets, Abg-SciQA features an additional task for classifying types of ambiguity, enabling a more comprehensive analysis.

approaches rely on pre-defined clarification questions (Eberhart and McMillan, 2022; Aliannejadi et al., 2019). The success of these methods is heavily dependent on the quality of the datasets used. High-quality datasets are crucial not only for producing accurate clarification questions but also for enhancing the overall ability of LLMs to manage ambiguous queries. Several datasets have been developed with this goal in mind. For instance, Abg-CoQA (Guo et al., 2021) is an extension of CoQA (Reddy et al., 2019) that includes ambiguous questions and related clarifications. Similarly, AmbigQA (Min et al., 2020) is built on NQ-Open (Lee et al., 2019). Other datasets, like the one proposed by Rao and Daumé III (2018), use StackExchange as source data.

However, existing benchmarks have the following limitations. First, question-answers in these source datasets, such as CoQA (Reddy et al., 2019), are publicly available and thus may be part of the LLM per-training data mixture. As a result, evaluating on those benchmarks may not reveal the models' real capabilities in addressing the ambiguity. Second, ambiguity annotations in dialogue-based datasets are sometimes questionable. Given the characteristics of continuity in dialogues, a question is usually considered non-ambiguous even if the explicit reference is missing. Taking the first example in Fig. 1, it is obvious that "What" refers to "what did he do for the children?" considering the dialogue continuity, thus non-ambiguous. Similarly, the second example should also be labeled as non-ambiguous since the book right after "drawing of her mother" is in red. Third, many of these datasets lack detailed annotations for different types of ambiguity, limiting their effectiveness in broader evaluations.

To address these limitations, we introduce a new dataset on ambiguity, Abg-SciQA, which leverages the capabilities of LLMs and incorporates articles from various natural and social science domains. Resolving ambiguity ensures scientific precision, enabling clear cross-disciplinary communication, ethical decision-making, and reliable knowledge accumulation, driving progress in both natural and social sciences. To avoid overlapping with pretraining data, ambiguous questions in Abg-SciQA are automatically generated by LLM. Then, an auxiliary LLM is then employed to assess the quality of the generated samples, with human evaluators also involved in the evaluation process. Finally, Abg-SciQA classifies each ambiguity question into four distinct types of ambiguity, which is new to this area and enables a more compressive analysis. In addition, each instance in Abg-SciQA consists of a unique question for the story, which avoids confusion brought by dialogues. We compare Abg-SciQA with other datasets in the ambiguity area in Table 1. Specifically, Abg-SciQA includes more than 13,000 instances and we evaluate Abg-SciQA on both closed-source LLMs and open-source LLMs. Fig. 2 shows a data instance in

Dataset	Data Size	# of Entries	Abg Rate	Ambiguity Detection	Type Classification	Clarification Generation	Clarification-Based QA
AmbigQA (Min et al., 2020)	64.0M	14,042	51.1%	✓	✗	✗	✗
Abg-CoQA (Guo et al., 2021)	21.1M	8,615	11.5%	✓	✗	✓	✓
ASQA (Stelmakh et al., 2022)	14.0M	6,316	45%	✓	✗	✗	✗
CAMBIGNQ (Lee et al., 2023)	27.8M	5,653	100%	✓	✗	✓	✓
Abg-SciQA (Ours)	52.3M	13,729	73.3%	✓	✓	✓	✓

Table 1: Comparisons of different datasets on ambiguity. The "# of Entries" means the total number of samples in the whole dataset. Abg-SciQA comprises four different tasks. Compared to other datasets on ambiguity, our dataset covers the widest range of tasks with decent numbers of entries and Abg Rate.

Abg-SciQA. We outline our contributions:

- We introduce Abg-SciQA, a dataset that includes challenging ambiguous questions from diverse scientific fields, complete with annotations for different types of ambiguity.
- To the best of our knowledge, we are the first to establish benchmarks for solving ambiguous questions using both closed-source LLMs, such as GPT-o1, and open-source LLMs, such as Llama2 (Touvron et al., 2023).
- Our comprehensive experiments demonstrate that fine-tuning LLMs with Abg-SciQA can significantly improve their ability to handle ambiguous questions.

2 Related Work

Many datasets address ambiguity in conversation and question answering. To our knowledge, Braslavski et al. (2017) introduces the first ambiguous dataset using community question-answering websites. Rao and Daumé III (2018) utilize data from StackExchange, while Saeidi et al. (2018) focus on rules and laws. Wu et al. (2023) creates an ambiguous dataset by extracting conversations from Wikipedia using web searches. Other ambiguous datasets are based on well-known public datasets. For example, Guo et al. (2021) propose Abg-CoQA, which clarifies ambiguities based on CoQA (Reddy et al., 2019). Min et al. (2020) generate AmbigQA for open-domain question answering based on NQ-Open (Lee et al., 2019). Stelmakh et al. (2022) uses AmbigQA to enhance long-form QA in the context of ambiguity. Lee et al. (2023) further refine AmbigQA with the assistance of InstructGPT (Ouyang et al., 2022).

Though there are many ambiguous datasets, these datasets do not provide annotations of different types of ambiguity. Besides, most datasets consider using data from simple areas like community conversations or public datasets which may used to train the language models. On the other hand, none of the previous works consider evaluating the close-source commercial LLMs such as GPT-o1. To this end, Our dataset contains not only annotations of ambiguous types but also high-quality passages and questions from various science areas. We include the evaluation of our datasets on the commercial LLMs as well, which distinguishes our work from previous.

3 Dataset Collection

We build Abg-SciQA based on various questions and different areas in both natural and social science. Abg-SciQA is composed of 1,353 stories and 13,729 questions, where 10,202 questions are annotated as ambiguous. The comparison in Table 1 shows that our dataset is one of the largest datasets in the ambiguity area. We present the distributions of source domains for Abg-SciQA in Fig. 4.

3.1 Task Definition

Given a story S and a question Q, the ultimate task is to resolve any ambiguity in the question Q if it is ambiguous. We consider:

Ambiguity Detection: Determining whether the question Q is ambiguous based on the story S.

Ambiguity Type Classification: Classifying the ambiguity type of question Q based on predefined ambiguity definitions.

Clarification Question Generation: Generating a clarification question CQ to resolve the ambiguity in Q if ambiguity is detected.

Clarification-Based Question Answering: Producing an unambiguous answer to Q using story S, the generated clarification question CQ, and a possible response R to CQ.

3.2 Material Collection

The previous datasets on ambiguity are mainly based on the public Natural Language Processing (NLP) dataset such as Abg-CoQA (Guo et al., 2021), which is generated based on CoQA (Reddy

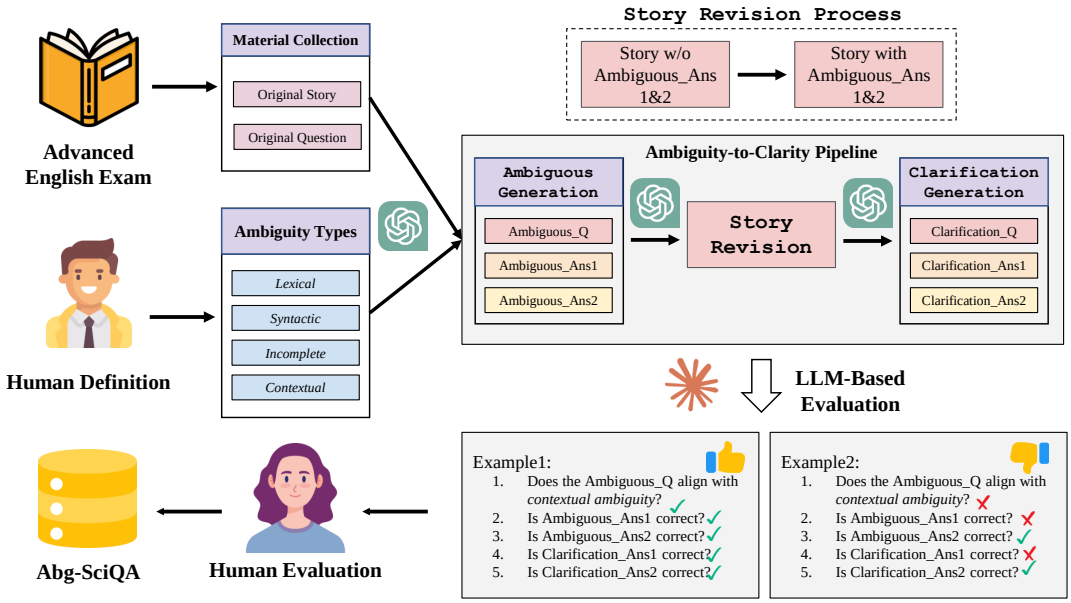

Figure 3: The Abg-SciQA pipeline for generating ambiguous questions and evaluation. It starts with an advanced English exam's original stories and original questions, along with a predefined ambiguity type, to generate an ambiguous question and two answers using GPT-4o. The story is then revised to align with the ambiguous answers, followed by generating clarification questions. For quality control, Claude-3.5-Sonnet and humans evaluate ambiguity and answer consistency. Only valid entries are added to the dataset, with a subset undergoing human evaluation before finalizing Abg-SciQA.

et al., 2019). Most of these public datasets are based on some simple tests and CoQA is based on the children's stories from MCTest (Richardson et al., 2013) and middle and high school English exams from RACE (Lai et al., 2017). These exams are less challenging compared with more advanced science questions and thus may be easier for LLMs to understand the contexts. Therefore, we collect stories, questions, and corresponding answers from TOEFL, IELTS, GRE, and GMAT reading comprehensive Mock Tests from the Internet. However, most of these questions are not ambiguous, which means we need to generate ambiguous questions.

3.3 Ambiguity Types

Before introducing how we generate ambiguous questions, we introduce the ambiguity types used in our paper. In Appendix, Table 8 and Table 9 show a breakdown and examples of the ambiguity type in Abg-SciQA. We use four categories:

Lexical Ambiguity occurs when a word has multiple meanings or multiple interpretations. For example, the *lexical ambiguity* example is: "Why did the medieval church need an alarm arrangement?". "The alarm arrangement" can be interpreted as waking people up or A signal warning of a threat. The answer will vary depending on the context.

Syntactic Ambiguity arises from the structure of a sentence, allowing multiple interpretations depending on how it is parsed. A *Syntactic Ambiguity* example is: "What aspect of creating new roles would most weaken the limited impact thesis criticized by women's rights activists?" The phrase "criticized by women's rights activists," can modify either "limited impact thesis" or "creating new roles.", leading to ambiguity.

Incomplete Ambiguity occurs when a question lacks essential information, such as location, time, event, or people, resulting in multiple possible interpretations. An example is: "When did he come back after the event". In this case, the question lacks specificity, making "he" ambiguous—it could refer to two different persons in the context.

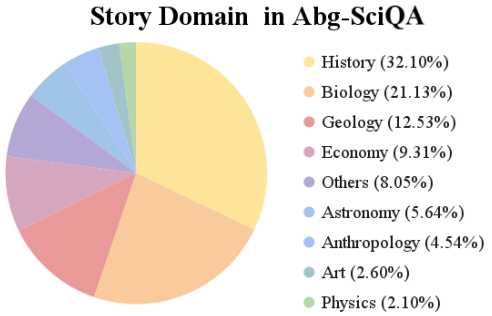

Figure 4: The distribution of story domains in Abg-SciQA. Among all domains, History questions account for the most in Abg-SciQA.

	Ambiguity Detection			Type Classification			Clarification Generation		QA
Model	Precision	Recall	F1	Precision	Recall	F1	BLEU	Rouge-L	F1
GPT-o1-mini	0.6707	0.1375	0.2282	0.2039	0.2425	0.1764	0.1393	**0.3517**	**0.0062**
GPT-o1	0.7985	0.5450	**0.6478**	0.2548	0.2550	0.2194	**0.1490**	0.3516	0.0025
Gemini-1.0	0.9259	0.0312	0.1605	0.3292	0.3325	0.3013	0.0716	0.2433	0.0012
Gemini-1.5	**0.9685**	0.2312	0.3733	0.4004	0.3112	0.2106	0.0822	0.2585	0.0050
Claude-haiku	0.8703	0.1762	0.2931	0.6263	0.3737	0.3416	0.0752	0.2591	0.0025
Claude-sonnet	0.3675	**0.8212**	0.5077	**0.6818**	**0.4587**	**0.3992**	0.1054	0.2933	0.0037

Table 2: Performance of different closed-source LLMs for all tasks provided by Abg-SciQA. We highlight the **best** performance and the second best. The results show that Claude-sonnet performs the best. However, even a powerful model like Claude-3.5-sonnet and GPT-o1 still perform not very well in all tasks. The full model name is o1-mini-2024-09-12, o1-preview-2024-09-12, Gemini-1.0-pro, Gemini-1.5-pro, Claude-3-5-haiku-20241022, and Claude-3-5-sonnet-20241022.

Contextual Ambiguity occurs when a question is unambiguous but contains two possible answers supported by the story. For example, the *Contextual Ambiguity* question is: "What insights were gathered from the research on how deer use their foreheads when they rub against trees?" There are two possible insights based on the context: one is the communication function of forehead rubbing. Another is the seasonality and physical traits associated with the behavior.

Each of these types of ambiguity represents distinct challenges in understanding ambiguity, which ensures the diversity of Abg-SciQA.

3.4 Ambiguity-to-Clarity Pipeline

In contrast to the Abg-CoQA dataset (Guo et al., 2021), which generates ambiguous questions using partial question histories, we employ GPT-4o to generate them automatically based on the Ambiguity-to-Clarity Pipeline. This pipeline consists of three components: `Ambiguous Generation`, `Story Revision`, and `Clarification Generation` and we provide the generation pipeline in Fig. 3.

`Ambiguous Generation` In this part, we employ GPT-4o to automatically generate ambiguous questions and corresponding answers from unambiguous questions, following established definitions of ambiguity. To facilitate this process, we design a prompt for LLMs in Appendix Fig. 5 that includes definitions of four ambiguity types, each accompanied by examples. This helps LLMs accurately interpret ambiguity. By providing clear definitions, we reduce the risk of misunderstandings. For instance, samples in Abg-CoQA (Guo et al., 2021) may be inconsistently interpreted regarding continuity in Fig. 1, causing similar cases to be classified under different ambiguity types, potentially leading to unfair evaluations. To maintain consistency, we prompt GPT-4o to generate only one type of ambiguity at a time.

`Story Revision` Though GPT4o could generate ambiguous questions and corresponding answers, GPT4o might generate answers that are not true according to the story. To ensure that the story supports the ambiguous answers, we continue using GPT4o to revise the original story to align with the generated answers, which ensures the generated questions are indeed ambiguous. The detailed prompt for `Story Revision` is in Appendix Fig. 6.

`Clarification Generation` In addition to generating ambiguous questions, we also employ GPT-4o for `Clarification Generation`. We prompt the model to generate clarification questions that resolve ambiguity in the target question. With the direct information of why there is an ambiguity, it is possible for GPT4o to generate the clarification question. We provide the corresponding prompt provided in Appendix Fig. 7.

3.5 Quality Control

To ensure the quality of our results, we employ a two-stage evaluation process that consists of both LLM-Based and Human Evaluation.

3.5.1 LLM-Based Evaluation

In the first stage, we use Claude-3.5-Sonnet to assess the quality of the generated dataset. Specifically, we provide the LLMs with the generated samples, a corresponding ambiguity type definition, and the revised story. We then ask whether the proposed question meets the ambiguity requirements and aligns with the given ambiguity definition. Additionally, we ask the LLMs whether the answers to the ambiguous and clarification questions can be found in the revised story to ensure their accuracy. In total, we generate 25,000 samples using GPT-

4o, and Claude-3.5-Sonnet filters out 8,975 entries during this stage of evaluation.

3.5.2 Human Evaluation

In the second stage, we conduct human evaluation. To ensure the quality of human evaluation, evaluators must pass an exam consisting of five entries labeled by a human expert before proceeding with the assessment. We begin by conducting a human evaluation to identify and remove poorly constructed ambiguous questions from the dataset and filtering out around 2,296 entries. After this refinement, We also randomly select 50 entries from the dataset and have four human evaluators answer the same questions as Claude to assess its reliability. We use Cohen's Kappa statistic (ML. et al., 2012) to measure the agreement between Claude's evaluations and human evaluations, yielding a final result of 0.6535, which indicates moderate agreement between Claude's assessments and human evaluations. This suggests that Claude's assessments are generally reliable.

4 Evaluation on Abg-SciQA

In this paper, we evaluate Abg-SciQA in both closed-source commercial LLMs and open-source public LLMs to show how different language models deal with our dataset comprehensively. We also include the results of fine-tuned LLMs with Abg-SciQA to show Abg-SciQA can guide the improvement of LLMs. All of our experiments are done on one single NVIDIA A100-80G GPU. In detail, we randomly sample 80% of Abg-SciQA as the training set and use the rest as the evaluation set. We prompt the LLMs using few-shot examples.

4.1 Evaluation Metrics

Abg-SciQA contains four tasks, one more than previous work (Guo et al., 2021). For Ambiguity Detection, since we treat this task as a binary classification, we report precision, recall, and F1 as the evaluation metrics. For Ambiguity Type Classification, since this task can be treated as a multi-class classification, we compute macro-precision, macro-recall, and macro-F1. For Clarification Question Generation, we use BLEU and Rouge-L as metrics with the labeled clarification question as the gold standard. In addition to directly measuring the quality through automatic metrics, we also manually evaluate whether the generated question is reasonable and helpful for clarifying the existing ambiguity for a small subset. Finally, for Clarification-Based Question Answering, we follow the common practice to compute the macro-average F1 score of word overlap (Reddy et al., 2019).

4.2 Evaluation on Closed-source LLMs

We evaluate Abg-SciQA in 3 closed-source LLMs: 1) GPT-o1 (Achiam et al., 2023) 2) Gemini (Team et al., 2023) 3) Claude (Anthropic, 2024). We do not evaluate Abg-SciQA on GPT-4o (Achiam et al., 2023) because we use GPT-4o (Achiam et al., 2023) to generate our dataset. We provide the results in Table 2 and detailed results for each type of ambiguity in the Table 10. We have the following observations given the results:

1)Different closed-source LLMs exhibit varying performance across different tasks. For example, GPT-o1 performs the best in Ambiguity Detection, while Claude-sonnet shows a stronger performance in tasks like Ambiguity Type Classification.

2) However, as we can see, even GPT-o1 and Claude-sonnet do not solve the problem very well. For example, the F1 score for Ambiguity Type Classification in GPT-o1 is only 0.2194, and the F1 score for Ambiguity Detection in Claude-sonnet is 0.5077. What's more, all the model performances are really bad on Clarification-Based QA. This indicates that our dataset is highly challenging, and even state-of-the-art models such as GPT-o1 struggle to handle it well.

3) Compared to all tasks, Ambiguity Detection appears to be the easiest with the highest F1 scores. This indicates current models could understand the basic concept of ambiguity. However, the poor performances on other tasks show that resolving ambiguity is still a challenging task.

4) Even though Ambiguity Type Classification is a comparatively easier problem among all tasks provided in Abg-SciQA, the results in Table 10 show that it is hard for LLMs to understand all types of ambiguity. Claude-sonnet with the highest overall performance on Ambiguity Type Classification shows a very good understanding of *Lexical ambiguity* and *Contextual Ambiguity*. However, Claude-sonnet can hardly understand the rest two types, especially for *Syntactic ambiguity*.

4.3 Evaluation on Open-source LLM

Now we evaluate Abg-SciQA in 5 open-source LLM: 1) Llama2 (Touvron et al., 2023) 2) Llama3 3) Gemma (Team et al., 2024) 4) Phi3.5 (Abdin et al., 2024) 5) Mistral (Albert et al., 2023) and their variants. We present our overall results in

	Ambiguity Detection			Type Classification			Clarification Generation		QA
Model	Precision	Recall	F1	Precision	Recall	F1	BLEU	Rouge-L	F1
Llama2-7B	0.7528	0.2219	0.2731	0.2958	0.1200	0.1655	0.0303	0.1714	0.0096
Llama2-13B	0.7491	0.3421	0.2527	0.2430	0.1737	0.1666	0.0050	0.0912	0.0083
Llama3.1-8B	0.8837	0.3612	0.5128	0.3453	0.0662	0.0969	0.0035	0.0741	0.0175
Llama3.2-3B	**0.8850**	0.0962	0.1736	0.2460	0.2062	0.1338	0.0027	0.0658	**0.0187**
Gemma-2B	0.7689	0.5825	0.6628	0.2839	**0.2714**	0.1668	0.0069	0.0821	0.0054
Phi3.5	0.8089	0.5187	0.6321	0.3071	0.2525	0.1415	0.0326	0.1627	0.0154
Mistral-0.1	0.7940	**0.7348**	**0.7632**	0.2615	0.2450	**0.2240**	0.0482	0.1984	0.0112
Mistral-0.2	0.8235	0.1925	0.3120	**0.3821**	0.2587	0.1564	**0.0697**	**0.2300**	0.0087

Table 3: Performance of different open-source LLMs for all tasks provided by Abg-SciQA. we highlight the **best** performance and the second best. The results show that open-source LLMs are not bad at Ambiguity Detection. However, most of them fall short on other tasks, compared with more powerful closed-source LLMs.

	Ambiguity Detection			Type Classification			Clarification Generation		QA
Model	Precision	Recall	F1	Precision	Recall	F1	BLEU	Rouge-L	F1
Llama2-7B	0.9989	**0.9989**	**0.9989**	0.4741	0.5055	0.6000	**0.1847**	**0.4773**	0.0114
Llama2-13B	0.9982	0.9948	0.9969	**0.7812**	**0.7894**	**0.7917**	0.1224	0.4256	0.0167
Llama3.1-8B	0.9683	0.9739	0.9829	0.6650	0.7285	0.5235	0.1749	0.2316	**0.0195**
Llama3.2-3B	**1.0000**	0.9692	0.9843	0.5459	0.6977	0.6451	0.0920	0.3736	0.0187
Gemma-2B	**1.0000**	0.9794	0.9896	0.4990	0.4269	0.6287	0.0313	0.1928	0.0092
Phi3.5	0.9979	0.9984	0.9984	0.5196	0.5502	0.5948	0.0738	0.2018	0.0179
Mistral-0.1	0.9858	0.9986	0.9923	0.5107	0.6155	0.6235	0.0759	0.2935	0.0141
Mistral-0.2	0.9979	0.9983	0.9984	0.5069	0.5218	0.6297	0.0775	0.2821	0.0093

Table 4: Performance of different open-source LLMs for all tasks after fine-tuning on Abg-SciQA. we highlight the **best** performance and the second best. The results show that fine-tuning on Abg-SciQA can significantly increase the performance of all LLMs and make smaller-size models even better than closed-source LLMs.

Table 3 and results for different ambiguity types in Table 6. We have the following observations:
1) Similar to previous results with closed-source LLMs, all open-source models struggle to solve ambiguous problems effectively. Additionally, compared to closed-source LLMs, some open-source models perform worse, likely due to the significant difference in the number of parameters between closed-source and open-source models.
2) Even though Mistral-0.1 achieves the best performance in Ambiguity Detection and Ambiguity Type Classification, it can only understand three types of ambiguity well, as shown.
3)Aside from Mistral-0.1 and Gemma-2B, the other open-source LLMs perform poorly across most tasks. While some models, such as Llama3-1.8B, excel in Ambiguity Detection, their performance in Ambiguity Type Classification, Clarification Generation, and QA remains subpar. The variation in performance across tasks suggests that these models struggle to consistently address ambiguity, highlighting the need for further improvement.

Finally, we tend to fine-tune the open-source LLMs with Abg-SciQA to see if our dataset can guide the further training of LLMs. In detail, we randomly sample 80% of Abg-SciQA as the training set and use the rest as the evaluation set. We choose the same open-source models in the previous section as training models. We use LoRA to train all the models with AdamW (Loshchilov and Hutter, 2017), Lora rank 8, learning rate 5e-6, and training epochs 3. And we present the results of the evaluation set in Table 4 and Table 7. We can have the following observations:
1) Fine-tuning can significantly improve the performance of chosen open-sourced LLMs. We can see that the best model Llama2-13B can even beat the closed-source LLMs in Ambiguity Detection and Ambiguity Type Classification, demonstrating the effectiveness of training on Abg-SciQA to solve the problem.
2) Though fine-tuning significantly improved performance, similar to open-source LLMs, there still remain a lot of problems for LLMs in understanding the meaning of all types of ambiguity. For example, Though Llama2-13B has a good performance on Ambiguity Type Classification after fine-tuning, Llama2-13B can only have a good understanding of two types of ambiguity. These results indicate that further improvement is needed and Abg-SciQA can help to guide development.

	Ambiguity Detection			Clarification Generation		QA
Model	Δ Precision	Δ Recall	Δ F1	Δ BLEU	Δ Rouge-L	Δ F1
Llama2-7B	-2.7%	<u>388.2%</u>	<u>100.2%</u>	**865.3%**	432%	1.8%
Llama2-13B	**53.9%**	-35.3%	-6%	<u>276.9%</u>	**615.1%**	30.3%
Llama3.1-8B	-15.1%	358.4%	14%	100.5%	<u>245.9%</u>	1.2%
Llama3.2-3B	-1%	**586.7%**	**114.9%**	53.9%	157.7%	7.3%
Gemma-2B	2.1%	5.1%	1.8%	27.9%	40.3%	**31.5%**
Phi3.5	<u>4.9%</u>	-0.8%	10%	60.1%	33.8%	<u>30.4%</u>
Mistral-0.1	2.5%	30.5%	-0.6%	205.5%	214.6%	15.2%
Mistral-0.2	-2.7%	-12.6%	-24.1%	116.7%	174.7%	21.2%

Table 5: Performance increasing of different open-source LLMs for all tasks on Abg-CoQA after fine-tuning on Abg-SciQA. we highlight the **best** performance and the <u>second best</u>.The results indicate that, in general, fine-tuning on Abg-SciQA enhances performance on Abg-CoQA, demonstrating Abg-SciQA' strong generalization ability in addressing ambiguity.

Model	*Lexical*	*Syntactic*	*Incomplete*	*Contextual*
Llama2-7B	0.2202	0.0667	0.1479	0.2276
Llama2-13B	0.2395	0.0000	0.3983	0.0287
Llama3.1-8B	0.0857	0.0000	0.1233	0.1787
Llama3.2-3B	0.3611	0.0747	0.0538	0.0457
Gemma-2B	0.1812	0.0291	0.0095	0.1573
Phi3.5	0.0294	0.0000	0.1379	0.3987
Mistral-0.1	0.2717	0.0611	0.3088	0.2544
Mistral-0.2	0.0100	0.0000	0.2242	0.3916

Table 6: Detailed F1 score of Ambiguity Type Classification for each type on open-source LLMs without fine-tuning. The results show that most LLMs cannot understand understand ambiguity well before fine-tuning.

Model	*Lexical*	*Syntactic*	*Incomplete*	*Contextual*
Llama2-7B	0.8704	0.0098	0.0000	0.6881
Llama2-13B	0.9684	0.7232	0.5350	0.8121
Llama3.1-8B	0.9386	0.6792	0.0613	0.7529
Llama3.2-3B	0.9635	0.0744	0.1921	0.7079
Gemma-2B	0.9804	0.0000	0.0000	0.6982
Phi3.5	0.9150	0.0952	0.1373	0.6801
Mistral-0.1	0.9798	0.0288	0.0479	0.6939
Mistral-0.2	0.9925	0.0286	0.0000	0.6971

Table 7: Detailed F1 score of Ambiguity Type Classification for each type on open-source LLMs after fine-tuning. The results show that most LLMs can only understand two types well even after fine-tuning.

4.4 Transfer Ability for Abg-SciQA

We evaluated open-source LLMs that are fine-tuned on Abg-SciQA by Abg-CoQA (Guo et al., 2021) to see whether Abg-SciQA can help to increase the ability to solve other general ambiguity questions. We focused on Abg-CoQA because Abg-CoQA has more comprehensive tasks (three tasks in total). To better analyze transfer ability, we consider the performance increasing between models without fine-tuning and models with fine-tuning on Abg-SciQA. Our results are presented in Table 5. Based on these results, we have the following observations:

1) Compared to evaluating open-source LLMs, Abg-CoQA performed better on our fine-tuned model across tasks. For instance, in Ambiguity Detection with Llama2-7b, the F1 score has an improvement of approximately two times compared to the model without fine-tuning.

2) When evaluating on Abg-SciQA, Llama3.2-3B shows minimal performance improvement after fine-tuning. However, when evaluated on Abg-CoQA, its performance improves significantly in recall and F1, particularly in Ambiguity Detection. While its overall improvement is not the highest among all models, these results further validate that fine-tuning on Abg-Sci enhances models' ability to handle a wide range of ambiguous questions, demonstrating its effectiveness as a training set for ambiguity resolution.

5 Conclusion

In this paper, we introduce Abg-SciQA, a dataset aiming at evaluating LLMs on detecting and solving ambiguity comprehensively. Derived from advanced science questions and enhanced with generated ambiguous questions, Abg-SciQA encompasses four key tasks to analyze ambiguity better. Our extensive experiments on both closed-source and open-source LLMs reveal that even state-of-the-art models struggle with these tasks, highlighting areas for improvement. Notably, fine-tuning open-source LLMs on Abg-SciQA leads to substantial performance gains, demonstrating its potential to guide LLM development in ambiguity handling. Additionally, we evaluated Abg-CoQA using Abg-SciQA fine-tuned models, which also showed significant improvement. This demonstrates the flexibility of Abg-SciQA fine-tuned models and its potential to perform well on other datasets. Abg-SciQA thus serves as a valuable benchmark for advancing language understanding and interaction on ambiguous questions.

Limitations

While Abg-SciQA significantly advances ambiguity detection and resolution, it has several limitations. The dataset primarily focuses on social and natural sciences, limiting its applicability to other domains like law and medicine. Expanding its coverage could improve generalizability. The evaluation metrics used, such as BLEU, Rouge-L, and F1 scores, may not fully capture the effectiveness of clarifications, necessitating more precise assessment frameworks.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774, 2023*.

Jiang Albert et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

A.I. Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 345–348.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *American Psychological Association*.

Yang Deng, Shuaiyi Li, and Wai Lam. 2023. Learning to ask clarification questions with spatial reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2113–2117.

Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. 2024. Evaluating large language models in class-level code generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Zachary Eberhart and Collin McMillan. 2022. Generating clarifying questions for query refinement in source code search. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 140–151. IEEE.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. *In 3rd Conference on Automated Knowledge Base Construction*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking clarification questions to handle ambiguity in open-domain qa. *arXiv preprint arXiv:2305.13808*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

McHugh ML. et al. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*.

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

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203,

Seattle, Washington, USA. Association for Computational Linguistics.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805, 2023*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295, 2024*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288, 2023*.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Inscit: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Samuel Yu, Shihong Liu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. 2023. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pages 418–428.

A Dataset and Code

As mentioned in the abstract, our code and dataset can be found at https://anonymous.4open.science/r/Abg-Sci-DF10/README.md.

B Ambiguity Type Definition

In this section, we provide the detailed definition of each ambiguity type in Table 8 and the examples for each ambiguity type in Table 9 .

C Dataset Structure

In this section, we provide the structure of Abg-SciQA. In detail, Abg-SciQA is stored in a JSON file, and the following example shows the detailed structure of the JSON file.


```
{
 "id": 464,
 "story": "Petroleum, consisting of
     crude oil and natural gas, seems to
      originate from organic matter in
     marine sediment. Microscopic
     organisms settle to the seafloor
     and accumulate in marine mud. The
     organic matter may partially
     decompose, using up the dissolved
     oxygen in the sediment. As soon as
     the oxygen is gone, decay stops and
      the remaining organic matter is
     preserved. Continued sedimentation-
     the process of deposits' settling
     on the sea bottom-buries the
     organic matter and subjects it to
     higher temperatures and pressures."
     ,
 "target_turn": {
  "question": "What happens to the
      organic matter in marine sediment
      over time?",
 },
 "ambiguity_turn": {
  "ambiguity": "ambiguous",
  "ambiguity_type": "Literal vs. Implied
       Interpretation"
 },
 "clarification_turn": {
  "question": "Are you asking about the
      initial decomposition process of
      the organic matter or its
      transformation into oil and gas?",
  "answers": [
   {
    "clr_ans": "The initial
        decomposition process of the
        organic matter in marine
        sediment.",
    "org_ans": "The organic matter may
        partially decompose, using up
        the dissolved oxygen in the
        sediment and accumulate in
        marine mud."
   },
   {
    "clr_ans": "The transformation of
        the organic matter into oil and
        gas.",
    "org_ans": "The organic matter is
        subjected to higher temperatures
         and pressures, which convert it
         to oil and gas."
   }
  ]
 }
}
```


D Prompt for Generating Ambiguous Question

We present our prompt for generating ambiguous questions in Fig. 5. This prompt includes the story, original question, and original answer. It then defines and provides examples of three types of ambiguity before instructing GPT-4o to generate the required output using Chain of Thought reasoning. Any samples that do not conform to the expected format are adjusted by human annotators.

In Fig. 6, we show the prompt used for story revision, ensuring that the ambiguous question aligns more closely with the story. Fig. 7 details the prompt for generating clarification questions. Finally, Fig. 8 presents the prompts given to Claude-sonnet-3.5 to assess the alignment of ambiguous questions with ambiguous definitions and verify the correctness of all answers based on the story.

E Extra Experiment

In this section, we present an additional experiment based on the Ambiguity Type Classification of each category using a closed-source LLM, as shown in Table 10.

F Comparison of Fine-Tuned and Non-Fine-Tuned Models on Ambiguity Detection and Ambiguity Type Classificationn

In this section, we compare the performance of the model without fine-tuning and the fine-tuned model on Ambiguity Detection and Ambiguity Type Classification. In Fig. 9, we perform Ambiguity Detection using Mistral_0.1, both before and after fine-tuning with Abg-SciQA. The results indicate that the fine-tuned model successfully detects ambiguity. The target question was initially considered unambiguous by LLMs because its wording appeared clear and straightforward. However, after fine-tuning the LLMs on Abg-SciQA, they successfully detected that the question was ambiguous due

Ambiguity Type	Definition
Lexical	This occurs when a word has multiple meanings or multiple interpretations.
Syntactic	This arises from the structure of a sentence. A sentence can have multiple interpretations depending on how it's parsed.
Incomplete	This occurs when a statement or question lacks essential contextual information—such as location, time, event, or people—resulting in multiple possible interpretations.
Contextual	This type of ambiguity occurs when a question is clear and unambiguous in its wording, but it contains two possible answers due to differing contexts, interpretations, or sources.

Table 8: The definition of four types of ambiguity: *Lexical*, *Syntactic*, *Incomplete*, and *Contextual Ambiguity*.

Ambiguity Type	Example
Lexical (19.65%)	**Q:** Why did the medieval church need an alarm arrangement? **C:** Are you asking if the medieval church needed an alarm arrangement to wake people up or to signal a threat? **C_A:** To wake people up / **Q_A:** The medieval church used an alarm arrangement to wake people up. **C_A:** To signal a threat / **Q_A:** The medieval church used an alarm arrangement to signal a threat to the community.
Syntactic (15.67%)	**Q:** What aspect of creating new roles would most weaken the limited impact thesis criticized by women's rights activists? **C:** Do you mean new roles in high-tech and service sectors or the broader societal and economic transitions? **C_A:** New roles in high-tech and service. / **Q_A:** The aspect where new roles in the high-technology and service sectors were being created. **C_A:** Societal and economic transitions. / **Q_A:** Critics argue the transition was painful but temporary for broader societal impact.
Incomplete (30.46%)	**Q:** When did he come back after the event? **C:** Are you asking about Tom or Jason? **C_A:** Tom / **Q_A:** 6 p.m. **C_A:** Jason / **Q_A:** 9 p.m.
Contextual (34.22%)	**Q:** What insights were gained about deer using their foreheads to rub trees? **C:** Are you asking about the communication function or the seasonality and traits of forehead rubbing? **C_A:** The communication function / **Q_A:** Studies show forehead rubbing in deer communicates identity, sex, and dominance through scent. **C_A:** The seasonality and physical traits / **Q_A:** Studies link forehead rubbing in male deer to the rutting season, high glandular activity, and darker pelage.

Table 9: Four types of ambiguity: *Lexical*, *Syntactic*, *Incomplete*, and *Contextual* Ambiguity. In the examples, 'Q' denotes an ambiguous question, 'C' represents a clarification question, 'C_A' stands for the clarification answer, and 'Q_A' signifies the answer to the ambiguous question after clarification.

Model	*Lexical*	*Syntactic*	*Incomplete*	*Contextual*
GPT-o1-mini	0.2461	0.0417	0.0515	0.3666
GPT-o1	0.2152	0.1107	0.1905	0.3613
Gemini-1.0	0.5298	0.0000	0.3521	0.3236
Gemini-1.5	0.3740	0.0000	0.4222	0.0465
Claude-haiku	0.7003	0.0392	0.2121	0.4150
Claude-sonnet	0.8177	0.0583	0.2472	0.4737

Table 10: Detailed F1 score of Ambiguity Type Classification for each type on closed-source LLMs. The results show that LLMs often cannot understand ambiguity well even though the powerful model like GPT-o1 and Claude-sonnet.

to the presence of two possible answers. As highlighted in yellow, these two sentences illustrate the multiple valid interpretations of the target question. Thus, this is a good case for detect ambiguity.

Conversely, in Fig. 10, despite fine-tuning, the model fails to detect ambiguity.The target question was initially considered unambiguous by LLMs because its wording appeared clear and unambiguous. Even after fine-tuning, the model still classified it as unambiguous. However, the correct classification should be ambiguous since the target question has two possible answers. Therefore, this is a failure case.

In Fig. 11, we perform Ambiguity Type Classification on both the original Llama3.1-8B (without fine-tuning) and the fine-tuned version with Abg-SciQA. The results show that after fine-tuning, the model successfully predicts the ambiguity type. The model without fine-tuning classifies the target question as lexically ambiguous because the keyword "population" can have different meanings depending on the context. However, in the given story, "population" has only one meaning. After fine-tuning on Abg-SciQA, the model classifies the target question as contextually ambiguous because there are multiple instances of the word "affect" in the story, leading to two possible answers. Thus, this is a good case for Ambiguity Type Classification, as the model successfully identifies the correct ambiguity type after fine-tuning.

Similarly, in Fig. 12, we conduct the same Ambiguity Type Classification task, but in this case, the fine-tuned model fails to predict the ambiguity type. In this case, the model without fine-tuning classifies the target question as lexically ambiguous because the word "substantial" has multiple meanings. However, this classification is incorrect since the keyword does not appear in the story. After fine-tuning, the model classifies the question as incomplete ambiguity, reasoning that it lacks essential contextual information. However, this classification is also incorrect because the question is clearly stated but allows for two possible answers, as highlighted in yellow in the story.

Ambiguous Generation

Story: {**story**}, Original Question: {**original question**}, Original Answer: {**original answer**}.
The ambiguous question has four types: Lexical, Syntactic, Incomplete, and Contextual Ambiguity
Example1: Lexical Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**},
Example 2: Syntactic Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**},
Example3: Incomplete Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**}.
Example4: Contextual Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**}.
Instructions:
Please use the story, original question, and original answer to generate an ambiguous question based on <Lexical, Syntactic, Incomplete, or Contextual> Ambiguity. Please think step by step and tell me the reason why the question you generated is ambiguous and give me two possible answers based on the story. Please generate the ambiguous question based on the story, original question, and original answer rather than the examples. The answers of ambiguous questions must be clearly found in the story and please give me two ambiguous answers.
Output Format:
Ambiguous Question: <**your generated question**>
Ambiguous Answer 1: <**first possible answer**>
Ambiguous Answer 2: <**second possible answer**>
Explanation: <**explanation of why the question is ambiguous**>

Figure 5: The prompt for `Ambiguous Generation` in Abg-SciQA.

Story Revision

Story: {**story**}, Ambiguous Question: {**generated question**}, Ambiguous Answer 1: { **answer for ambiguous question**}, Ambiguous Answer 2: {**answer for ambiguous question**}, Explanation: {**reasons why the question is ambiguous**}
Instructions:
Please revise the story based on the ambiguous question, ambiguous answers, and explanation, and make the ambiguous answer true. Please give me the full story after revised. You should make sure the ambiguous answer is followed the revised story. The ambiguous question and answers must be the same as the input. The Ambiguous answer cannot be Partially Correct. It should be fully correct based on the revsied story.
Output Format:
Revised Story:<**revised story**>
Ambiguous Question:<**ambiguous question**>
Ambiguous answer1:<**answer for ambiguous question**>
Ambiguous answer2:<**answer for ambiguous question**>

Figure 6: The prompt for `Story Revision` in Abg-SciQA.

Clarification Generation

Revised Story:{**revised story**}, Ambiguous Question:{**ambiguous question**}, Ambiguous answer1:{**answer for ambiguous question**}, Ambiguous answer2:{**answer for ambiguous question**}
Instructions:
Please ask a clarification question to clarify the ambiguous question based on the revised story, ambiguous question, and answers. For the c_answer, please don't start with 'If you are referring to'. If the Clarification Question start with 'Are you asking about A or B?' The c_answer1 should be I'm asking about A. The c_answer2 should be I'm asking about B. The Ambiguous Answers should not be the same as Clarification Answers Please provide me with the clarification answer in format: "
Output Format:
Clarification Question:<**clarification question**>
c_answer1:<**clarification answer1**>
c_answer2:<**clarification answer2**>

Figure 7: The prompt for `Clarification Generation` in Abg-SciQA.

LLM-Based Evaluation

Revised Story:{**revised story**}, Ambiguous Question:{**ambiguous question**}, Ambiguous answer1:{**answer for ambiguous question**}, Ambiguous answer2:{**answer for ambiguous question**}, Clarification question:{**clarification question**}, c_answer1:{**clarification answer1**}, c_answer2:{**clarification answer2**}
The ambiguous question has four types: Lexical, Syntactic, Incomplete, and Contextual Ambiguity
Example1: Lexical Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**},
Example 2: Syntactic Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**},
Example3: Incomplete Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**}.
Example4: Contextual Ambiguity Question: {**definition and a question example**}
clarification question: {**clarification question**}.
Instructions:
Verify if the ambiguous question aligns with the definition of "<Lexical, Contextual, Syntactic, or Incomplete> Ambiguity". If the ambiguous question overlaps with other types of ambiguity, please directly output "False". If the ambiguity rely on one word and have multiple interpretations, it must be Lexcial Ambiguity. For each answer (`ambiguous_answer_1`, `ambiguous_answer_2`, `c_answer1`, `c_answer2`), assess its correctness and provide an explanation with a supporting sentence from the `revised_story`.
Output Format:
Match with <Lexical, Contextual, Syntactic, or Incomplete>Ambiguity Definition: <**True or False**>
Explanation: <**reasons why the question aligns or does not align with the Ambiguity**>
Correctness of ambiguous_answer_1: <**Correct or Incorrect**>
Explanation1: <**Reasons why ambiguous_answer_1 is correct or incorrect**>
Correctness of ambiguous_answer_2: <**Correct or Incorrect**>
Explanation2: < **Reasons why ambiguous_answer_2 is correct or incorrect**>
Correctness of c_answer1: <**Correct or Incorrect**>
Explanation3: <**Reasons why c_answer1 is correct or incorrect**>
Correctness of c_answer2: <**Correct or Incorrect**>
Explanation4: <**Reasons why c_answer2 is correct or incorrect**>

Figure 8: The prompt for `LLM-Based Evaluation` in Abg-SciQA.

Ambiguity Detection on Abg-SciQA Good Case

Story:

By 1850, the United States possessed roughly 9,000 miles of railroad track; ten years later, it had over 30,000 miles, more than the rest of the world combined. Much of the new construction during the 1850s occurred west of the Appalachian Mountains—over 2,000 miles in the states of Ohio and Illinois alone. The effect of the new railroad lines rippled outward through the economy, fundamentally transforming both trade routes and agricultural practices.\n\n\n\n The new railroad networks shifted trade dynamics by redirecting western trade from the south to the east, significantly impacting the economic relationships of the time. In 1840, most northwestern grain was shipped south down the Mississippi River to the bustling port of New Orleans. However, this route was fraught with difficulties: low water levels made steamboat travel hazardous in the summer, and ice shut down traffic in winter. Products such as lard, tallow, and cheese spoiled quickly in New Orleans' hot and humid warehouses. Increasingly, traffic from the Midwest flowed from west to east, over the new rail lines. This shift moved trade away from New Orleans and made Chicago a crucial trade hub, linking the farms of the upper Midwest to New York and other eastern cities by more than 2,000 miles of track in 1855. As a result, while the value of goods shipped by river to New Orleans continued to increase, the South's overall share of western trade dropped dramatically. Simultaneously, the new rail networks empowered farmers along the tracks to specialize in crops that they could market in distant locations, enhancing the commercial orientation of agriculture and shifting the overall economic landscape. For instance, before the railroad reached Tennessee, the state produced about 25,000 bushels (or 640 tons) of wheat, which sold for less than 50 cents a bushel. Once the railroad came, farmers in the same counties grew 400,000 bushels (over 10,000 tons) and sold their crop at a dollar a bushel. A sharp rise in demand for grain abroad also encouraged farmers in the Northeast and Midwest to become more commercially oriented. Wheat, which in 1845 commanded $1.08 a bushel in New York City, fetched $2.6 in 1855; similarly, the price of corn nearly doubled. Farmers responded by specializing in cash crops, borrowing to purchase more land, and investing in equipment to increase productivity….

Target Question:

How did the new rail networks transform the trade dynamics in the western United States during the 1850s?

Detected Ambiguity After Fine-Tuning:

Unambiguous

Explanation:

The question is clear and unambiguous in its wording. It specifies the subject (new rail networks), the geographical area (western United States), and the time period (1850s).

Detected Ambiguity After Fine-Tuning:

Ambiguous

Explanation:

The question is ambiguous because there are two possible answers for the target question.

Correct Ambiguity:

Ambiguous

Figure 9: Good Case for Ambiguity Detection on Abg-SciQA: This is a strong example because, before fine-tuning, the LLM classified this case as unambiguous. However, after fine-tuning, it successfully detected the ambiguity in the task and provided a reasonable explanation for why it is ambiguous.

Ambiguity Detection on Abg-SciQA Bad Case

Story:
Because of industrialization, but also because of a vast increase in agricultural output without which industrialization would have been impossible, Western Europeans by the latter half of the nineteenth century enjoyed higher standards of living and longer, healthier lives than most of the world`s peoples. In Europe as a whole, the population rose from 188 million in 1800 to 400 million in 1900. By 1900, virtually every area of Europe had contributed to the tremendous surge of population, but each major region was at a different stage of demographic change.Improvements in the food supply continued trends that had started in the late seventeenth century. New lands were put under cultivation, while the use of crops of American origin, particularly the potato, continued to expand. Setbacks did occur. Regional agricultural failures were the most common cause of economic recessions until 1850, and they could lead to localized famine as well. A major potato blight (disease) in 1846-1847 led to the deaths of at least one million persons in Ireland and the emigration of another million, and Ireland never recovered the population levels the potato had sustained to that point. Bad grain harvests at the same time led to increased hardship throughout much of Europe.After 1850, however, the expansion of foods more regularly kept pace with population growth, though the poorer classes remained malnourished. Two developments were crucial. First, the application of science and new technology to agriculture increased. Led by German universities, increasing research was devoted to improving seeds, developing chemical fertilizers, and advancing livestock. After 1861, with the development of land-grant universities in the United States that had huge agricultural programs, American crop-production research added to this mix. Mechanization included the use of horse-drawn harvesters and seed drills, many developed initially in the United States. It also included mechanical cream separators and other food-processing devices that improved supply. The second development involved industrially based transportation. With trains and steam shipping, it became possible to move foods to needy regions within Western Europe quickly. Famine (as opposed to malnutrition) became a thing of the past. Many Western European countries, headed by Britain, began also to import increasing amounts of food, not only from Eastern Europe, a traditional source, but also from the Americas, Australia, and New Zealand. Steam shipping, which improved speed and capacity , as well as new procedures for canning and refrigerating foods (particularly after 1870), was fundamental to these developments. …
Target Question:
What caused the food supply to increase in most of Western Europe during the nineteenth century?
Detected Ambiguity After Fine-Tuning:
Unambiguous
Explanation:
The question is clear and unambiguous in its wording.
Detected Ambiguity After Fine-Tuning:
Unambiguous
Explanation:
The question is clear and not ambiguous.
Correct Ambiguity:
Ambiguous(There are two possible answers)

Figure 10: Bad Case for Ambiguity Detection on Abg-SciQA: This is a poor example because the target question exhibits *contextual ambiguity*. The definition and wording of the sentence are clear, but there are two possible answers, making it *contextually ambiguity*. However, despite fine-tuning, the model still classifies it as unambiguous, highlighting a failure in ambiguity detection.

Ambiguity Type Classification on Abg-SciQA Good Case

Story:

Phytoplankton are minute, free-floating aquatic plants. In addition to the marked changes in abundance observed in phytoplankton over the course of a year, there is also a notable change in species composition. This seasonal shift in the dominant species is known as seasonal succession and occurs in diverse locations. Under seasonal succession, certain species dominate for a period before being replaced by others, a cycle that repeats annually. ... These organic metabolites likely include various classes of organic compounds. Some of these are toxins, such as those released by dinoflagellates (a type of plankton) during red tides, which inhibit the growth of other photosynthetic organisms. During red tides, dinoflagellate populations grow so extensively that the water takes on a reddish-brown hue due to the sheer number of cells. While each cell secretes only a small amount of toxin, the high population density results in toxin concentrations that can be lethal to many marine creatures. This toxin can also accumulate in filter-feeding organisms like clams and mussels, making them toxic to humans. Importantly, no special mechanism is required for this population explosion; the sheer abundance of dinoflagellates alone is sufficient to alter water coloration. Another class of metabolite is vitamins. Certain phytoplankton species require specific vitamins for survival, with considerable variation in these requirements across species. The B vitamins—especially B12, thiamine, and biotin—are among the most essential. Some species cannot thrive unless specific vitamins are available in the water, which are produced by other species. This dependency can lead to a succession pattern where a vitamin-producing species appears first, followed by vitamin-dependent species. Other organic compounds, such as amino acids, carbohydrates, and fatty acids, may also influence phytoplankton succession. Although laboratory studies show that different phytoplankton species vary in their ability to produce and require essential vitamins, the extent of this influence in natural marine ecosystems remains unclear. Additionally, selective grazing by herbivorous crustaceans, such as copepods and invertebrate larvae, can affect species composition by selectively feeding on certain phytoplankton. Increasing evidence suggests that multiple factors interact simultaneously to drive species succession, with the relative importance of each factor varying based on species and environmental conditions.

Target Question:

How does the population size of dinoflagellates during red tides affect the environment according to the passage?

Ambiguity Type Classification Before Fine-Tuning:

Lexical Ambiguity

Explanation:

The question is ambiguous because the term ”population” can refer to either the number of dinoflagellates or their environmental impact, leading to multiple possible interpretations.

Ambiguity Type Classification After Fine-Tuning:

Contextual Ambiguity

Explanation:

The ambiguity arises because the question has several “affect” and leads to several answers, so it’s contextual ambiguity.

Correct Ambiguity Type:

Contextual Ambiguity

Figure 11: Good Case for Ambiguity Type Classification on Abg-SciQA:Before fine-tuning, the LLM misclassified the case as *lexical ambiguity*, assuming "population" had multiple meanings. However, in context, it has only one meaning. After fine-tuning, the LLM correctly identified it as *contextual ambiguity*, as the question is clear but has multiple possible interpretations. This shows the fine-tuned LLM can successfully classify ambiguity types.

Ambiguity Type Classification on Abg-SciQA Bad Case

Story:

Earth has abundant water in its oceans but very little carbon dioxide in its relatively thin atmosphere. By contrast, Venus is very dry and its thick atmosphere is mostly carbon dioxide. The original atmospheres of both Venus and Earth were derived at least in part from gases spewed forth, or outgassed, by volcanoes. The gases that emanate from present-day volcanoes on Earth, such as Mount Saint Helens, are predominantly water vapor, carbon dioxide, and sulfur dioxide. These gases suggest the possible original atmospheric compositions of both Venus and Earth, emphasizing why Venus now has a thick carbon dioxide atmosphere while Earth does not. Much of the water on both planets is also thought to have come from impacts from comets, icy bodies formed in the outer solar system.\n\n\n\nIn fact, water probably once dominated the Venusian atmosphere. Venus and Earth are similar in size and mass, so Venusian volcanoes may well have outgassed as much water vapor as on Earth, and both planets would have had about the same number of comets strike their surfaces. Studies of how stars evolve suggest that the early Sun was only about 70 percent as luminous as it is now, so the temperature in Venus' early atmosphere must have been quite a bit lower. Thus water vapor would have been able to liquefy and form oceans on Venus. But if water vapor and carbon dioxide were once so common in the atmospheres of both Earth and Venus, what became of Earth's carbon dioxide? And what happened to the water on Venus?... But Venus being closer to the Sun than Earth is, enough of the liquid water on Venus would have vaporized to create a thick cover of water vapor clouds. Since water vapor is a greenhouse gas, this humid atmosphere, perhaps denser than Earth's present-day atmosphere, would have efficiently trapped heat from the Sun. At first, this would have had little effect on the oceans of Venus… Over time, the rising temperatures would have leveled off, solar ultraviolet radiation having broken down atmospheric water vapor molecules into hydrogen and oxygen. With all the water vapor gone, the greenhouse effect would no longer have accelerated.\n\n

Target Question:

What evidence from the passage suggests that Venus may have once had substantial water?

Ambiguity Type Classification Before Fine-Tuning:

Lexical Ambiguity

Explanation:

The target question is ambiguous because the word "substantial" has multiple meanings. However, there is no keyword 'substantial' in the context.

Ambiguity Type Classification After Fine-Tuning:

Incomplete Ambiguity

Explanation:

The target question lacks essential contextual information, so it's incomplete ambiguity.

Correct Ambiguity Type:

Contextual Ambiguity

Figure 12: Bad Case for Ambiguity Type Classification on Abg-SciQA:Before fine-tuning, the LLM misclassified the question as *lexical ambiguity*, assuming "substantial" had multiple meanings, though it only appeared in the target question, not the context. After fine-tuning, it incorrectly labeled it as *incomplete ambiguity*, claiming missing information, whereas the question is clear but has multiple valid interpretations, making it *contextual ambiguity*. Since both models failed to classify it correctly, this is a bad case.