CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models

Tong Zhang♠♡#, **Peixin Qin**♠♡#, **Yang Deng**♣, **Chen Huang**♠♡,
Hongru Liang♠♡, **Junhong Liu**♢, **Dingnan Jin**♢,
Wenqiang Lei♠♡†, **Tat-Seng Chua**♣

♠ College of Computer Science, Sichuan University, China
♡ Engineering Research Center of Machine Learning and Industry Intelligence,
Ministry of Education, China
♣ National University of Singapore
♢ University of Electronic Science and Technology, China
wenqianglei@scu.edu.cn

Abstract

Large language models (LLMs) are increasingly used to meet user information needs, but their effectiveness in dealing with user queries that contain various types of ambiguity remains unknown, ultimately risking user trust and satisfaction. To this end, we introduce CLAMBER, a benchmark for evaluating LLMs using a well-organized taxonomy. Building upon the taxonomy, we construct $\sim 12K$ high-quality data to assess the strengths, weaknesses, and potential risks of various off-the-shelf LLMs. Our findings indicate the limited practical utility of current LLMs in identifying and clarifying ambiguous user queries, even enhanced by chain-of-thought (CoT) and few-shot prompting. These techniques may result in overconfidence in LLMs and yield only marginal enhancements in identifying ambiguity. Furthermore, current LLMs fall short in generating high-quality clarifying questions due to a lack of conflict resolution and inaccurate utilization of inherent knowledge. In this paper, CLAMBER presents a guidance and promotes further research on proactive and trustworthy LLMs. Our dataset is available at `https://github.com/SCUNLP/CLAMBER`.

1 Introduction

Given well-defined user queries, large language models (LLMs) have demonstrated remarkable proficiency in facilitating the information search process (Pan et al., 2023; Kamalloo et al., 2023; Zhang and Choi, 2023; Huang et al., 2023). They provide more precise search results with the help of the inherent knowledge stored within LLMs. Nonetheless, as evidenced by previous studies (Kuhn et al., 2023; Deng et al., 2023a), the practical utility of LLMs is hindered by unclear and ambiguous user queries in real-world scenarios. For instance, in a query like "*what are the strategies for saving?*", the term "*saving*" can have multiple interpretations, such as "*saving money*" or "*saving from sins*", depending on the user's actual need. This necessitates LLMs proactively identifying (i.e., determine if the query is ambiguous or not) and clarifying the ambiguities rather than providing potentially incorrect answers that may not align with the user's true needs, ultimately risking user trust and satisfaction (Liao et al., 2023).

Driven by this concern, recent works have explored LLMs' capacity to address ambiguous queries (Deng et al., 2023b; Kuhn et al., 2023). However, these investigations have been somewhat fragmented, lacking a comprehensive taxonomy, leading to incomplete and inconsistent handling of ambiguity distributions (Keyvan and Huang, 2022; Rahmani et al., 2023). As a notable example, they are often limited to contextual ambiguity, where the given context is insufficient for producing a definitive answer. In the era of LLMs, there should be more emphasis on the LLM-oriented ambiguity that may occur when inherent knowledge stored within LLMs have conflict understanding about the query. Consequently, it still remains unclear which ambiguities LLMs can effectively identify and clarify, along with the challenges that LLMs persistently encounter in this regard.

To this end, we introduce CLAMBER (**Cl**arifying **Amb**iguous Qu**er**y), a novel benchmark for comprehensively evaluating LLMs in identifying and clarifying various ambiguities using a well-organized taxonomy. Drawing inspiration from the input-process-output framework

[†] Corresponding author.
[#] Both authors contributed equally to this study.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10746–10766
August 11-16, 2024 ©2024 Association for Computational Linguistics

for evaluating collaborative systems (Pinsonneault and Kraemer, 1989), we establish a taxonomy that consolidates both input understanding and task completion perspectives into three primary dimensions, as illustrated in Table 1. These dimensions are further conceptualized into eight fine-grained categories to facilitate in-depth evaluation. Building upon this taxonomy, we construct $\sim 12K$ data for analyzing the pros and cons of LLMs when identifying and clarifying ambiguities.

With CLAMBER, we comprehensively evaluate strengths, weaknesses, and potential risks of various LLMs. Our findings indicate that ChatGPT (OpenAI, 2022) outperforms other small-scale LLMs, especially excelling in identifying and clarifying ambiguities in multifaceted queries (Clarke et al., 2009), such as "*What is the largest manufacturer in China?*", which does not specify the type of "*manufacturer*". However, they still encounter numerous challenges: 1) **current LLMs, despite leveraging chain-of-thought (CoT) and few-shot prompting, face challenges in identifying ambiguities.** Our results suggest that CoT and few-shot prompting may lead to the overconfidence issue in small-scale LLMs, impacting ambiguity identification negatively. Even with a large number of shots and CoT support, LLMs only achieve a marginal improvement. Moreover, current LLMs struggle to leverage contextual cues to disambiguate pronouns, highlighting the inadequacy in deducing underlying ambiguities. 2) **Current LLMs fail to ask high-quality clarifying questions, due to the inability of knowing their knowledge gap.** Despite LLMs recognize a query containing ambiguities, their lack of conflict resolution and inaccurate use of inherent knowledge results in uncertainty about which ambiguity to clarify. This prompts the need of developing effective methods for LLMs to resolve conflicts and accurately utilize their inherent knowledge.

In this paper, CLAMBER stands as a valuable resource to provide guidance and insight into evaluating LLMs and addressing ambiguous information needs for future improvements. In conclusion, our contributions are threefold:

- We introduce a taxonomy for categorizing various query ambiguities. This taxonomy combines three primary dimensions, detailed as eight categories for facilitating fine-grained evaluations.
- We present a novel benchmark called CLAMBER, tailored to the characteristics of LLMs. It contains $\sim 12K$ data featuring ambiguous user queries across diverse categories.
- With CLAMBER, we evaluate the off-the-shelf LLMs in an inclusive manner. Our findings shed light on why current LLMs struggle to identify and clarify ambiguities. These insights will guide future research in this field.

2 Related Works

Our research is closely tied to the taxonomy and resolution of ambiguities in LLMs. We provide a literature review and highlight our differences.

Ambiguity Taxonomy. As evidenced by a recent survey (Rahmani et al., 2023), there is a lack of a well-organized taxonomy for ambiguity in information retrieval. While previous research attempts to integrate ambiguity taxonomies, their taxonomies are fragmented and underdeveloped (Ginzburg, 1996; Song et al., 2007), failing to facilitate comprehensive evaluations. Recent taxonomies Min et al. (2020); Guo et al. (2021) are formulated based on a limited set of factual questions and lack precise definitions for each category. Moreover, existing taxonomies were established before the era of LLMs, disregarding the ambiguity specific to LLMs that may arise from conflicting interpretations of queries by the inherent knowledge stored within LLMs. This is evident when LLMs encounter unfamiliar entities (Yin et al., 2023) or potential inconsistencies within queries (Tamkin et al., 2022). For the first time, we introduce a well-organized taxonomy for categorizing various query ambiguities. Our taxonomy draws inspiration from the input-process-output view to evaluate collaborative systems. It combines three primary dimensions that capture potential ambiguities during input understanding and task completion of LLMs. Using this taxonomy, we construct $\sim 12K$ data for analyzing the pros and cons of LLMs in resolving different ambiguities.

Resolving ambiguity in LLMs. Recent efforts resort to CoT and few-shot prompting to enhance LLMs' capacity in identifying and clarifying ambiguous queries (Deng et al., 2023b; Kuhn et al., 2023; Cole et al., 2023). While these efforts have shown some improvements in performance, they are confined to tasks involving specific types of ambiguities, such as lexical ambiguity. In this paper, we incorporate CoT and few-shot prompting as baselines to evaluate their efficacy and inadequacy

Dimension	Category	Explanation	Example
Epistemic Misalignment	**UNFAMILIAR**	Query contains unfamiliar entities or facts	Find the price of Samsung Chromecast.
	CONTRADICTION	Query contains self-contradictions	Output 'X' if the sentence contains [category withhold] and 'Y' otherwise. The critic is in the restaurant.>X. The butterfly is in the river.>Y. The boar is in the theatre.>?
Linguistic Ambiguity	**LEXICAL**	Query contains terms with multiple meanings	Tell me about the source of Nile.
	SEMANTIC	Query lacks of context leading multiple interpretations	When did he land on the moon?
Aleatoric Output	**WHO**	Query output contains confusion due to missing personal elements	Suggest me some gifts for my mother.
	WHEN	Query output contains confusion due to missing temporal elements	How many goals did Argentina score in the World Cup?
	WHERE	Query output contains confusion due to missing spatial elements	Tell me how to reach New York.
	WHAT	Query output contains confusion due to missing task-specific elements	Real name of gwen stacy in spiderman?

Table 1: The proposed taxonomy of ambiguous queries and examples. The clarifying questions of each example are provided in Table 8.

across a broader range of ambiguity types using CLAMBER. Other related works try to examine which of the two queries exhibits more ambiguity (Zhang and Choi, 2023), unable to determine if a query is ambiguous.

3 CLAMBER Benchmark

To evaluate LLMs in an inclusive manner, we present CLAMBER, which introduces a taxonomy encompassing three key dimensions (i.e., *Epistemic Misalignment*, *Linguistic Ambiguity*, *Aleatoric Output*) that capture potential ambiguities during input understanding and task completion. These three dimensions are further divided into eight specific categories. We delve into the taxonomy and data collection process in following sections. Each data comprises a user query, a binary ambiguity label, and a clarifying question for ambiguous queries. See details of data collection in Appendix C.

3.1 Epistemic Misalignment (EM)

Building upon the input understanding perspective, EM occurs when inherent knowledge stored within LLMs have conflict understanding about the query (Cole et al., 2023; Zhang and Choi, 2023). This ambiguity is a distinctive feature of LLMs, as they respond to queries using their inherent knowledge. We categorize EM into two categories based on the source of conflicting:

- **Unfamiliar**. It refers to situations where LLMs encounter entities or facts that are unfamiliar to them, either because they are not within the LLMs' inherent knowledge or because they contradict it. Given a query "*Find the price of Samsung Chromecast*", if LLMs only have inherent knowledge on "*Google Chromecast*" or "*Samsung Chromebook*" and are unfamiliar with "*Samsung Chromecast*", LLMs should proactively ask for clarification about "*Samsung Chromecast*" rather than provide answers regarding "*Google Chromecast*" or "*Samsung Chromebook*", ultimately risking user satisfaction.
- **Contradiction**. It refers to situations where LLMs infers contradictions within queries based on their inherent knowledge. For example, given a query "*Output 'X' if the sentence contains [category withhold] and 'Y' otherwise. Examples: The critic is in the restaurant.>X. The butterfly is in the river.>Y. Sentence: The boar is in the theatre.>?*", LLMs may infer two different categories (i.e., human and indoor location) from provided examples. This contradiction could lead to confusion for LLMs. Consequently, LLMs should seek clarification by asking: "*Does this category a human or an indoor location?*"

Data Collection. To evaluate the *Unfamiliar* category, it is important to determine exactly what LLMs are unfamiliar with (Wang et al., 2023). To mitigate bias stemming by training data, CLAMBER opts to utilize entirely new, fabricated knowledge that are unfamiliar to all LLMs. To achieve this, we resort to the ALCUNA dataset (Yin et al., 2023) as our data resource, which includes queries that contain new entities fabricated by modifying existing ones. We classify the queries containing new entities as ambiguous, while the rest are unambiguous. Subsequently, we instruct GPT-4 to generate a clarifying question for each ambiguous query, focusing on the ambiguity of new entities. As for the *Contradiction* category, the contradiction in CLAMBER occurs when the query and the given examples fail to match within a single interpretation. To achieve this, we directly utilize the AmbiTask dataset (Tamkin et al., 2022) to provide ambiguous queries, which encodes contradiction among queries and provided examples. Additionally, we create clarifying questions for ambiguous queries by rule-based templates and manually transform ambiguous queries into unambiguous ones by resolving contradictions.

3.2 Linguistic Ambiguity (LA)

Building upon the input understanding perspective, LA arises when a word, phrase, or statement can be interpreted in multiple ways due to its imprecise or unclear meaning (Berry and Kamsties, 2004; Ortega-Martín et al., 2023). We categorize LA into the the lexical and semantic ambiguities[1], which encapsulate the main challenges in information retrieval (Coden et al., 2015; Xu et al., 2019).

- **Lexical Ambiguity**. It concerns individual terms with multiple meanings. For example, given a query "*Tell me about the source of Nile*", the term "*source of Nile*" can be interpreted in two meanings: the origin of the Nile river or the board game named "*source of Nile*". In this case, LLMs should ask for clarification: "*Are you referring to the Nile river or the board game?*"
- **Semantic Ambiguity**. It involves the lack of context leading to more than one interpretation of a sentence (Ortega-Martín et al., 2023). For example, given a query "*When did he land on the moon?*", it is unclear who "*he*" may refer to without context. In this case, LLMs should ask for clarification: "*Who is 'he' referring to?*"

Data Collection. Lexical Ambiguity pertains to individual terms with multiple meanings, often found in entity names and polysemy words (Keyvan and Huang, 2022). In this paper, we resort to the AmbER (Chen et al., 2021) and AmbiPun dataset (Mittal et al., 2022), which contain ambiguous entity names and ambiguous polysemy words, respectively. We extract these terms along with their various meanings from the datasets and then create ambiguous queries, clarifying questions and unambiguous queries using GPT-4. As for Semantic Ambiguity, CLAMBER pay special focus investigating referent ambiguity following (Kuhn et al., 2022; Ortega-Martín et al., 2023). This type of ambiguity occurs in queries containing pronouns that lack contextual clues for clarification. Specifically, we employ the AmbiCoref dataset (Yuan et al., 2023), which consists of minimal pairs featuring ambiguous and unambiguous referents. In this regard, an ambiguous query can be achieved by reducing context sizes to a single sentence and creating sentences where the verbs involved limit the interpretation of their arguments. Additionally, we obtain the unambiguous queries by instructing GPT-4 and obtain clarifying questions by rule-based templates.

3.3 Aleatoric Output (AO)

Building upon the task completion perspective, AO occurs when the input is well-formed but the output contains potential confusion due to the lack of essential elements. It is prevalent across various types of queries in information retrieval, including faceted queries (Clarke et al., 2009), queries missing details (Trienes and Balog, 2019), board queries (Song et al., 2007) and under-specific queries (Aliannejadi et al., 2021). Previous studies have focused on specific aspects of this ambiguity, but there is a need for a more comprehensive understanding of this ambiguity in order to advance research. Inspired by (Zamani et al., 2020), we categorize AO into four specific categories based on the type of missing elements:

- **Whom** denotes the absence of personal details, such as expertise. Given a query "*Suggest me some gifts for my mother*", the response may vary due to missing the personal preferences of his mother. In this case, a clarifying question like: "*What specific preferences does your mother have?*" would be preferred.
- **Where** pertains to the lack of spatial information, such as departure place. For example, given a query "*Tell me how to reach New York*", the response may vary due to missing the specific departure information. In this case, LLMs should ask for clarification "*Where do you start from?*"
- **When** refers to the absence of temporal elements, such as specific dates. Given a query "*How many goals did Argentina score in the World Cup?*",

[1] We omit the syntactic and pragmatic ambiguities as they are not commonly used in information retrieval.

the response may vary due to missing the specific World Cup year. This ambiguity requires LLMs to seek further details by asking clarifying questions "*Which year of the World Cup are you referring to?*"

- **What** refers to the remaining types. For example, when a query is "*Who played Thanos in Guardians of the Galaxy?*", the response may vary due to missing the specific version of Guardians of the Galaxy. Clarifying question should arise: "*Which version are you referring to: TV series, 2014 film, or Telltale Series?*"

Data Collection. We construct four categories of ambiguities by recognizing the specific missing elements in well-structured queries. To accomplish this, we resort to the the AmbigQA dataset (Min et al., 2020) and the Dolly-16K dataset (Conover et al., 2023)) containing factual and instrumental user search intent (Alexander et al., 2022). As for AmbigQA dataset, queries with multiple answers are deemed ambiguous, while those with a single answer are considered unambiguous. Ambiguous queries are manually categorized into the four categories. In the Dolly dataset, each query is automatically labeled as ambiguous or unambiguous by GPT-4, then manually verified and classified into the four categories if marked as ambiguous. Due to the difficulty in crafting category-specific unambiguous queries, all four categories share the same set of unambiguous queries.

3.4 Validation and Revision

To ensure the quality of our dataset, we engage five linguistic experts for validation and revision. Initially, each data is validated by four experts, and subsequently consolidated by the remaining expert. The validation process includes verifying the accuracy of ambiguity labels and assessing the effectiveness of clarifying questions. If there are discrepancies between the four experts' validation, the final expert examines their feedback and implements necessary data revisions. For further details on the validation and revision procedures, please refer to Appendix F. Finally, our data statistics are presented in Table 2.

4 Experimental Design

We consider two tasks to evaluate off-the-shelf LLMs, including identifying ambiguities (cf. Section 5) and asking clarifying questions (cf. Section 6). Each task utilizes different evaluation metrics, outlined in the corresponding sections.

Category	Sources	Distribution		
		Ambig.	Non-Ambig.	ALL
Unfamiliar	ALCUNA	684	547	1231
Contradiction	AmbiTask	600	600	1200
Lexical	AmbER,AmbiPun	815	921	1,736
Semantic	AmbiCoref	400	400	800
What	AmbigQA, Dolly	1255	3884 in total	7167 in total
Whom	AmbigQA, Dolly	762		
When	AmbigQA, Dolly	779		
Where	AmbigQA, Dolly	487		

Table 2: CLAMBER Dataset Sources and Statistics.

Test Dataset. Our experiments are conducted on sub-sample of 3600 instances randomly selected, preserving the same number of data samples per category. There are 200 positive and negative examples for each category. Particularly, the negative examples of each category within Aleatoric Output is 800 since its uniform nature.

Usage of LLMs. As our set of LLMs, we evaluate Llama2-13B-Chat (i.e., Llama2-13B), Llama2-13B-Instruct (i.e., Llama2-13B-I), Vicuna-13B, Llama2-70B-Chat (i.e., Llama2-70B), and the GPT-3.5-Turbo-16k (i.e., ChatGPT). These LLMs are widely used in recent studies of information search (Deng et al., 2023b; Zhang and Choi, 2023).

Prompting Schemes. Following (Deng et al., 2023b), we devise four prompting schemes for evaluation: 1) Zero-shot w/o CoT, where the LLM is evaluated directly on the test dataset, 2) Zero-shot w/ CoT (Wei et al., 2022), where the LLM starts with ambiguity analysis before making predictions, 3) Few-shot w/o CoT (Dong et al., 2022), where the LLM is evaluated by providing examples, 4) Few-shot w/ CoT, where the LLM is evaluated by providing examples with their corresponding ambiguity analysis. In the few-shot setting, we provide two randomly selected examples, one is ambiguous and the other is unambiguous. Importantly, we carefully selected 3 prompts and test all LLMs on these prompts. We present the average performance across various prompts to guarantee the statistical significance of the experimental findings. Details on prompts are presented in Appendix A.

5 Task 1: Identifying Ambiguity

This section aims to evaluate the ability of LLMs to identify different categories of ambiguous user queries, focusing on both the overall performance (cf. Section 5.1) and performance specific to each category (cf. Section 5.2). Following (Hu et al.,

Methods	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Average Performance	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vicuna-13B	50.62	39.97	54.75	51.13	50.25	36.62	53.50	**52.23**	52.28	44.99
Llama2-13B-I	45.66	43.64	45.57	45.29	47.13	47.04	45.97	42.26	46.08	44.56
Llama2-13B	**55.47**	50.99	50.97	36.80	46.56	35.08	52.19	45.15	51.30	42.01
Llama2-70B	50.37	34.27	53.06	40.29	46.66	39.64	**54.93**	45.42	51.26	39.91
ChatGPT	54.34	**53.45**	**57.38**	**56.91**	**51.66**	**49.28**	53.60	51.42	**54.25**	**52.77**

Table 3: Overall ambiguity identification evaluation of LLMs with varying prompting schemes. ChatGPT emerges as the superior model, yet there is still considerable room for improvement, even enhanced by the CoT and Few-Shot.

2023; Deng et al., 2023b), we adopt the Accuracy and F1 score as metrics.

5.1 Overall Evaluation

As shown in Table 3, our findings suggest that current LLMs, despite leveraging CoT and few-shot prompting, face challenges in identifying ambiguities. Our detailed observations are as follows.

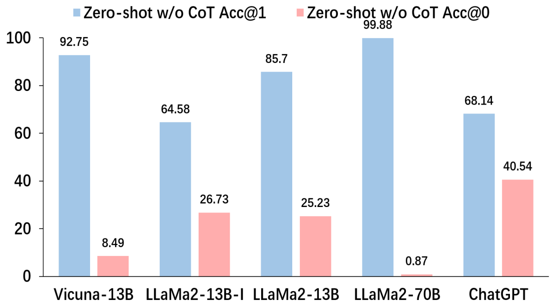

Figure 1: Investigation on the identification accuracy when handling ambiguous (i.e, Acc@1) versus unambiguous queries (i.e, Acc@0). We report the results under Zero-shot w/o CoT setting. Small-scale LLMs tend to classify most queries as ambiguous.

In general, current LLMs are struggle to identify ambiguities. We observe that small-scale LLMs are unable to differentiate between ambiguous and unambiguous queries. In particular, they not only show significantly low performance but also demonstrate a substantial discrepancy between accuracy and F1 score. For instance, the accuracy of Llama2-70B with Zero-shot w/o CoT is 50.37, while its F1 score is notably lower at 34.27. This implies a notable variation in their performance when handling ambiguous versus unambiguous queries. As depicted in Figure 1, these models tend to classify most queries as ambiguous, even those that are actually unambiguous. Compared to small-scale LLMs, ChatGPT stands out as the superior model. However, it only reaches an accuracy of 54.25% and an F1 score of 52.77%. There remains large room for improvement.

Metric	Model	Zero-shot w/o CoT	Zero-shot w/ CoT	Difference
ECE ↓	Vicuna-13B	21.47	19.81	-1.66
	Llama2-13B-I	22.43	19.91	-2.52
	Llama2-13B	28.48	45.14	+16.66
	Llama2-70B	48.21	47.24	-0.97
	ChatGPT	29.74	16.30	-13.44
ROC ↑	Vicuna-13B	49.73	51.37	+1.64
	Llama2-13B-I	56.18	56.40	+0.22
	Llama2-13B	57.00	48.22	-8.78
	Llama2-70B	50.74	56.33	+5.59
	ChatGPT	54.35	57.35	+3.00

Table 4: Overconfidence evaluation on LLMs with and without CoT. Significant differences are marked in grey.

Metric	Model	Zero-shot w/o CoT	Few-shot w/o CoT	Difference
ECE ↓	Vicuna-13B	21.47	25.66	+4.19
	Llama2-13B-I	22.43	20.99	-1.44
	Llama2-13B	28.48	44.10	+15.62
	Llama2-70B	48.21	31.68	-16.53
	ChatGPT	29.74	13.40	-16.34
ROC ↑	Vicuna-13B	49.73	48.70	-1.03
	Llama2-13B-I	56.18	56.56	+0.38
	Llama2-13B	57.00	50.55	-6.45
	Llama2-70B	50.74	43.84	-6.9
	ChatGPT	54.35	51.57	-2.78

Table 5: Overconfidence evaluation on LLMs with and without few-shot prompting. Significant differences are marked in grey.

CoT and few-shot prompting hold promise for enhancing ambiguity identification, but their effectiveness is not guaranteed. They may lead to the overconfidence issue in small-scale LLMs, leading to negative outcomes. As shown in Table 3, the effectiveness of CoT and few-shot prompting doesn't consistently improve. To delve deeper, we follow Cole et al. (2023) and gauged LLMs' prediction confidence [2] using Expected Calibration Error (ECE) and Area Under the Receiver Operating Characteristic curve (AUROC). ECE assesses the alignment of confidence scores with actual accuracy, while AUROC measures the ability of confidence scores to distinguish between correct and incorrect predictions. Our in-depth analysis, pre-

[2] Self-consistency confidence with 4 candidate answers are used to obtain the LLM's uncertainty (Xiong et al., 2023).

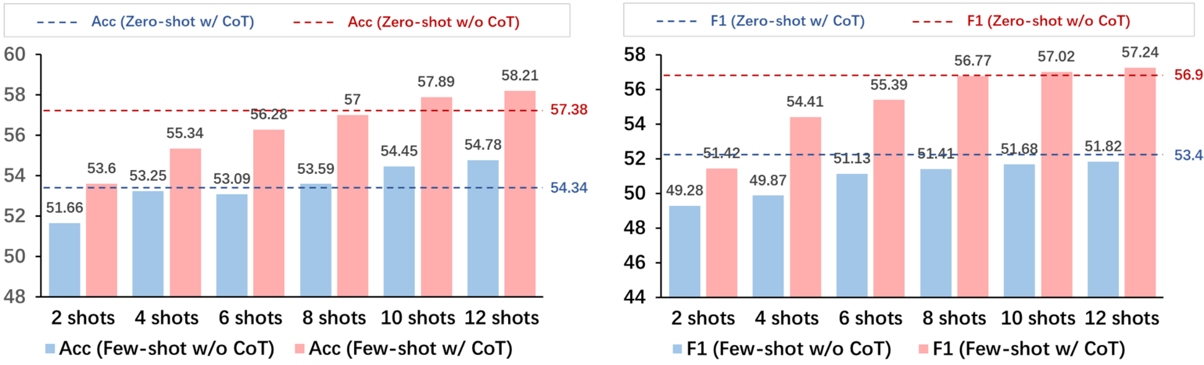

Figure 2: Performance of ChatGPT enhanced with multiple examples. We ensure a variety of categories in the examples and maintain an equal balance of ambiguous and unambiguous instances.

Methods	Epistemic Misalignment				Linguistic Ambiguity				Aleatoric Output							
	contradiction		unfamiliar		lexical		semantic		what		whom		when		where	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vicuna-13B	51.75	37.11	59.50	59.33	**72.00**	**71.52**	49.75	33.22	44.81	41.74	46.95	44.57	44.86	41.82	42.96	39.24
Llama2-13B-I	49.50	33.11	46.75	46.47	52.50	49.20	48.50	41.31	30.24	30.14	31.37	31.32	27.97	27.72	29.57	29.44
Llama2-13B	50.25	33.89	54.25	46.65	56.75	49.11	50.00	33.33	34.73	34.64	36.86	36.85	34.27	34.16	34.17	34.05
Llama2-70B	**63.25**	**58.83**	50.75	35.81	55.25	44.04	50.00	33.33	31.04	30.77	31.37	31.07	31.37	31.07	31.47	31.16
ChatGPT	38.00	28.17	**60.00**	**59.67**	58.75	58.06	**50.75**	**49.32**	**65.40**	**50.54**	**68.77**	**57.48**	**65.00**	**45.66**	**63.10**	**45.24**

Table 6: The fine-grained ambiguity identification evaluation results under Few-shot w/o CoT setting. ChatGPT demonstrates excellent performance across all categories of Aleatoric Output, but it does not effectively address the *semantic* and *contradiction* categories.

sented in Table 4 and Table 5, reveals that employing CoT and few-shot prompting leads small-scale LLMs (e.g., Llama2-13B) to exhibit over-confidence and less accurate ambiguity prediction, contrary to our intended outcome.

Even bolstered by numerous shots and CoT support, LLMs still struggles to accurately identify query ambiguity. Figure 2 illustrates the performance of ChatGPT when enhanced with multiple shots. The results indicate that the improvement seen with few-shot prompting is minimal and often inferior to the zero-shot counterpart. A considerable number of shots (e.g., 12 shots) are required for few-shot prompting to outperform the zero-shot method. However, this also entails longer input lengths, risking exceeding the length limit for most small-scale LLMs in our study. Providing examples alone to ChatGPT could result in the learning of superficial patterns that contradict its inherent knowledge, thereby diminishing its performance. Furthermore, ChatGPT's difficulty in fully grasping correct reasoning with limited examples could be another contributing factor.

5.2 Fine-Grained Evaluation

This section analyzes the challenges LLMs faced in comprehending different ambiguities, offering insights to guide future enhancements. Table 6 details the ambiguity identification performance of LLMs on each category. Here, we consider the Few-shot w/ CoT setting and leave more details in Appendix E. Our observations are as follows.

ChatGPT displays superior performance on Aleatoric Output compared to small-scale LLMs. Across all categories of Aleatoric Output, ChatGPT attains an average increase of 5% in accuracy and 8% in F1 score. This superior performance may stem from its vast world knowledge, enabling it to infer the absence of task-oriented elements in user queries. Additional results reveal ChatGPT performs exceptionally well in the "whom", while struggles more with the "when" and "where" categories. This suggests room for future improvement in handling queries lacking temporal and spatial elements.

The *semantic* category presents a significant challenge for all LLMs. As shown in Table 6, all LLMs exhibit subpar performance when dealing with ambiguous queries requiring semantic comprehension. This indicates that current LLMs struggle to use contextual cues to clarify pronouns, highlighting their inadequacy in robustly understanding context and inferring underlying ambiguity.

ChatGPT lags behind other small-scale LLMs

Methods	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Average Performance	
	BS	Help.	BS	Help.	BS	Help.	BS	Help.	BS	Help.
Vicuna-13B	21.63	36.64	24.42	40.97	20.45	33.97	17.65	31.87	21.04	35.86
Llama2-13B-I	19.65	33.07	14.29	26.34	8.01	14.16	7.96	17.36	12.48	22.73
Llama2-13B	21.46	35.38	23.46	40.40	10.79	19.05	23.68	40.75	19.85	33.89
Llama2-70B	22.05	37.35	19.17	32.72	19.71	32.80	22.49	39.99	20.86	35.71
ChatGPT	**27.47**	**46.45**	**30.22**	**50.47**	**31.16**	**51.58**	**33.48**	**53.29**	**31.33**	**50.45**

Table 7: Overall ambiguity clarification evaluation of LLMs with varying prompting schemes. ChatGPT emerges as the superior model to other open-sourced LLMs. We report BertScore (i.e., BS) and Help.

on the *contradiction* category. As shown in Table 6, ChatGPT only achieves limited accuracy (i.e., 38) and a low F1 score (i.e., 28.17). We observe that 81.97% of errors are false negatives, indicating that ChatGPT often misidentifies queries with self-contradictions as unambiguous. This limitation could be attributed to its training approach (i.e., SFT and RLHF (Ouyang et al., 2022)), which compels ChatGPT to generate responses for all user queries, irrespective of potential contradictions.

6 Task 2: Asking Clarifying Questions

This section investigates the ability of LLMs to produce effective clarifying questions for resolving ambiguities. Overall, current LLMs fail to ask high-quality clarifying questions, due to the inability of assessing their knowledge boundaries. Detailed observations are outlined below.

6.1 Overall Evaluation

We utilize *BertScore* for automated assessment, as lexical matching metrics can not adequately capture clarification abilities (Guo et al., 2021). Specifically, we compute the semantic similarity using BERT between the generated question and annotated clarifying questions. Additionally, we also conduct human evaluation[3] to score whether the generated question is helpful in resolving query ambiguity (denoted as **Help.**[4])

ChatGPT demonstrating its superior capabilities in generating clarifying questions compared to small-scale LLMs. Table 7 showcases the effectiveness of clarifying questions produced by different LLMs. It is evident that ChatGPT demonstrates an average performance improvement of 10.29 compared to Vicuna-13B, the top-performing small-scale LLM. This indicates that ChatGPT excels in generating natural and useful clarifying questions (i.e., what to ask).

6.2 Fine-Grained Evaluation

We provide an in-depth error analysis to reveal the inadequacies in asking clarifying questions. Since *ChatGPT + Few-shot w/ CoT* stands as the most effective model, our analysis focus on it. Specifically, we randomly sampled 50 error clarifying questions (whose *Help* scores are 0) from each category, 400 in total. Inspired by (Deng et al., 2023b), we categorize these failure cases into four groups:

- *Wrong Aspect*. It refers the case when the generated question is aimed to clarify an incorrect aspect of the user's query.
- *Under-specified*. The generated question is too unspecific, making it difficult for the user to provide useful feedback.
- *Over-specified*. The generated question is an overly detailed one when the needed information is already evident in the user's original query.
- *Generation error*. ChatGPT doesn't generate the output as the required format, such as no clarification question.

As illustrated in Figure 3, **inability of knowing their knowledge gap is the main reason for the inadequacies in asking effective clarifying questions**. Specifically, when dealing with the Epistemic Misalignment and Linguistic Ambiguity, most errors are concentrated on *Under-specified* and *Over-specified*, while *Wrong Aspect* is evident in Aleatoric Output, with an average of 52.25% error rate. This indicates that ChatGPT can not fully comprehend semantic nuances and lack of conflict resolution despite their large parameters. Moreover, ChatGPT use their inherent knowledge inaccurately to clarify the missing elements of ambiguous queries. These findings imply that there exists a gap between inherent knowledge within LLMs and the ambiguities contained in user queries.

[3] Refer to Appendix D for details on human evaluation.

[4] It entails assigning a binary score (0 or 1) to each generated question.

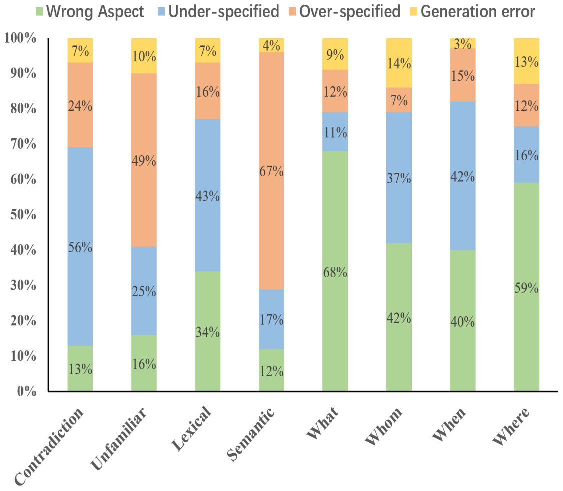

Figure 3: The statistics of error analysis. ChatGPT is unable to recognize their knowledge gap for the inadequacies in asking the effective clarifying questions.

7 Conclusion

In this work, we introduce CLAMBER, a benchmark for evaluating LLMs in identifying and clarifying ambiguous user queries through a well-organized taxonomy. Our CLAMBER comprises $\sim 12K$ high-quality data covering a wide range of ambiguity categories. With CLAMBER, we assess strengths, weaknesses, and potential risks of various off-the-shelf LLMs. Our results indicate that current LLMs still face difficulties in achieving optimal performance in ambiguity identification and clarification, limiting their practical utility in advanced information search applications. In this paper, CLAMBER acts as a foundation for enhancing the proactive capabilities of LLMs in addressing ambiguity. Moving forward, we plan to integrate more challenging and comprehensive datasets into our CLAMBER based on our taxonomy.

Limitations

In this section, we discuss the limitations of this work from the following perspectives:

Sensitivity of Prompts. Similar to other studies on prompting LLMs (Amayuelas et al., 2023; Deng et al., 2023b), the evaluation results are likely to be sensitive to the prompts. While we employ three different prompts and report the average results, it is challenging to assert that they are the most suitable ones for our specific issue. Indeed, the sensitivity of prompts and their optimality present significant research areas within LLMs, warranting further exploration in future studies.

Limited LLMs. We only use 5 Large Language Models (LLMs) in our CLAMBER benchmark due to computational constraints. If given additional resources and an improved experimental environment, it would be advantageous to evaluate the performance of other LLMs, such as PaLM540B, etc., in our CLAMBER benchmark.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62272330 and No. 62206191); in part by the Natural Science Foundation of Sichuan (No. 2023NSFSC0473), and in part by the Fundamental Research Funds for the Central Universities (No. 2023SCU12089 and No. YJ202219).

References

Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. 2022. Orcas-i: Queries annotated with intent using weak supervision. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3057–3066, New York, NY, USA. Association for Computing Machinery.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeffrey Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. *arXiv preprint arXiv:2109.05794*.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models.

Daniel M Berry and Erik Kamsties. 2004. Ambiguity in requirements specification. In *Perspectives on software requirements*, pages 7–44. Springer.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the trec 2009 web track. In *Trec*, volume 9, pages 20–29.

Anni Coden, Daniel Gruhl, Neal Lewis, and Pablo N Mendes. 2015. Did you mean a or b? supporting clarification dialog for entity disambiguation. In *Sumprehswi@ eswc*.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Yang Deng, Wenqiang Lei, Minlie Huang, and Tat-Seng Chua. 2023a. Rethinking conversational agents in the era of llms: Proactivity, non-collaborativity, and beyond. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 298–301.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023b. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Jonathan Ginzburg. 1996. Interrogatives: Questions, facts and dialogue. *The handbook of contemporary semantic theory*, 5(18):359–423.

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*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts?

Chen Huang, Peixin Qin, Wenqiang Lei, and Jiancheng Lv. 2023. Reduce human labor on evaluating conversational information retrieval system: A human-machine collaboration approach. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10876–10891, Singapore. Association for Computational Linguistics.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution.

Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *ACM Computing Surveys*, 55(6):1–40.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Clam: Selective clarification for ambiguous questions with generative language models.

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

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.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. Ambipun: Generating puns with ambiguous context. Association for Computational Linguistics (ACL).

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

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.

Haojie Pan, Zepeng Zhai, Hao Yuan, Yaojia Lv, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2023. Kwaiagents: Generalized information-seeking agent system with large language models. *arXiv preprint arXiv:2312.04889*.

Alain Pinsonneault and Kenneth L Kraemer. 1989. The impact of technological support on groups: An assessment of the empirical research. *Decision Support Systems*, 5(2):197–216.

Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A survey on asking clarification questions datasets in conversational systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2698–2716, Toronto, Canada. Association for Computational Linguistics.

Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying ambiguous queries in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 1169–1170.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2022. Task ambiguity in humans and language models.

Jan Trienes and Krisztian Balog. 2019. Identifying unclear questions in community question answering websites. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 276–289. Springer.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

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

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*.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1618–1629.

Xunjian Yin, Baizhou Huang, and Xiaojun Wan. 2023. Alcuna: Large language models meet new knowledge. *arXiv preprint arXiv:2310.14820*.

Yuewei Yuan, Chaitanya Malaviya, and Mark Yatskar. 2023. Ambicoref: Evaluating human and model sensitivity to ambiguous coreference.

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.

Michael J. Q. Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms.

A Prompt Design

Table 13 presents our four prompting schemes used for evaluation. In the case of few-shot prompting, we randomly choose two examples from our CLAMBER benchmark. The demonstration of chain-of-thoughts is written by human annotators, which represents their own ambiguity analysis.

B Implementation Details

The implementation of our LLMs is based on Pytorch and Transformers toolkit. In particular, for Llama2-13B-Chat[5] and Llama2-70B-Chat[6], we adopt the official version in Huggingface. For Llama2-13B-instruct, We adopt the version[7] that is fine-tuned on multiple instruction-following datasets. For Vicuna-13B, we choose the Vicuna-13B-delta-v1.5 version[8]. In particular, we set the temperature to 0 for ChatGPT and 0.5 for other open-sourced LLMs. In addition, we set the maximum number of new tokens to 128. During inference, the decoding strategy of open-sourced LLMs is top-p sampling with a top-p of 0.8. For the F1 score, we use the weighted F1 score as our metric, given the balanced nature of our test set. Our aim is to ensure the model's accuracy without ambiguity, minimizing the need for excessive clarification. All of our experiments are conducted on two NVIDIA A100 GPUs.

C Details of Data Collection

In this section, we describe the detailed data collection process of each category, including the data processing and the prompts used by GPT-4.

C.1 ALCUNA Dataset

The ALCUNA dataset (Yin et al., 2023) creates new entities by altering existing entity attributes and relationships, resulting in artificial entities that are distinct from real-world entities. It contains numerous question-answer pairs designed as a benchmark to evaluate the capabilities of LLMs, especially in handling new knowledge. Specifically, we classify questions containing new entities in this

[5] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[6] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[7] https://huggingface.co/Expert68/Llama2_13b_instructed_version2

[8] https://huggingface.co/lmsys/vicuna-13b-v1.5

dataset as ambiguous queries, and those involving existing entities as unambiguous queries. Furthermore, we randomly select 1500 ambiguous queries and employ GPT-4 to generate a clarifying question for each one, focusing on the ambiguity of the new entity. We provide the data examples and the prompt of generating clarifying question in Table 14.

C.2 AmbiTask Dataset

The AmbiTask Dataset (Tamkin et al., 2022) constructs multiple classification tasks, each accompanied by an instruction and two provided examples. The two examples can lead to multiple explanations of the instruction, resulting in contradictions. In particular, we select three classification tasks: "propn negation", "religious pronoun" and "subject location", totaling 1200 instances. we rephrase the ambiguous tasks using rule-based templates to enhance their clarity. These rephrased ambiguous tasks serve as our ambiguous queries. We generate clarifying questions based on the rule-based templates for these ambiguous queries. Additionally, we create unambiguous queries by manually modifying the instruction and make sure the two examples can lead to just one interpretation. We also rephrase these unambiguous queries to make them clear. The rule-based templates and data examples are provided in Table 16.

C.3 AmbER Dataset

The AmbER dataset (Chen et al., 2021) includes instances of entity ambiguity, where a single name can refer to multiple entities. Each ambiguous entity is annotated with its different meanings, and each meaning is associated with a factual question. Specifically, we have chosen the top-500 most frequent entities in the *non-human* category from AmbER as our data source. We feed the ambiguous entity and its questions related to each meaning into GPT-4, which then generates ambiguous queries along with corresponding clarifying questions. By providing these generated ambiguous queries and corresponding clarifying questions, we guide GPT-4 to produce a clear and unambiguous version. Further information about the prompts and data samples can be found in Table 17.

C.4 AmbiPun Dataset

The AmbiPun dataset (Mittal et al., 2022) comprises pun words that carry diverse meanings depending on the context. Each pun words is annotated with its various meanings. We randomly select 500 instances as our data resource, following the same data collection process as the AmbER dataset. Please refer to Table 18 for the prompts and data examples.

C.5 AmbiCoref Dataset

The AmbiCoref dataset (Yuan et al., 2023) consists of minimal pairs featuring ambiguous and unambiguous referents. This dataset extends the scope of psycholinguistic research on how individuals perceive ambiguity in specific verb structures and their arguments. We incorporate the ambiguous and unambiguous referent of this dataset as corresponding queries into our benchmark. For those ambiguous queries, we use a template to generate a clarifying question. The templates and examples are in Table 15.

C.6 AmbigQA Dataset

The AmbigQA dataset (Min et al., 2020) consists of ambiguous factoid questions sourced from Natural Questions (Kwiatkowski et al., 2019). We classify the questions with multiple answers as ambiguous while those those with a single answer are considered unambiguous. Furthermore, we rely on the clarifying question annotations in (Lee et al., 2023), we use the key word in their annotations and further categorize each ambiguous question manually into four categories. We adopt their annotated clarifying questions directly. The data examples are presented in Table 19.

C.7 Dolly Dataset

The Dolly dataset (Conover et al., 2023) is commonly used for instructional fine-tuning purposes. We specifically choose the instructions from the open-qa sub-category as they align with the task of information retrieval. Our approach involves instructing GPT-4 to differentiate between ambiguous and unambiguous queries, generating clarifying questions for the ambiguous ones, and then classifying them into our predefined categories. Please refer to Table 20 for examples of the prompts and data.

D Human Evaluation Details

To evaluate the effectiveness of clarifying questions produced by LLMs, we engage 3 annotators to conduct a human evaluation. Each annotator is tasked with evaluating each clarifying question alongside

Category	**Query**	**Clarifying Question**
UNFAMILIAR	Find the price of Samsung Chromecast.	Do you mean Google Chromecast or Samsung Chromebook?
CONTRADICTION	Output ’X’ if the sentence contains [category withhold] and ’Y’ otherwise. The critic is in the restaurant.>X. The butterfly is in the river.>Y. The boar is in the theatre.>?	Is the category either human or indoor location?
LEXICAL	Tell me about the source of Nile.	Are you referring to the Nile river or the board game?
SEMANTIC	When did he land on the moon?	Who do you mean by "he" in the question?
WHO	Suggest me some gifts for my mother.	What are your mother’s interests, hobbies, or something she has been wanting?
WHEN	How many goals did Argentina score in the World Cup?	Which World Cup are you referring to?
WHERE	Tell me how to reach New York.	Please tell me where your departure place is?
WHAT	Real name of gwen stacy in spiderman?	Which one are you referring to: the actress,or the character?

Table 8: The example clarifying questions associated with ambiguous queries in Table 1. There are no discerning patterns according to the ambiguity category.

the corresponding ambiguous query and its associated category of ambiguity. The annotators are instructed to adhere to a specific protocol for evaluating the quality of clarifying questions: Initially, they are to verify if the clarifying questions generated by LLMs adhere to the correct format. Subsequently, they are to determine whether the clarifying questions effectively aid in resolving ambiguity within user queries. In cases where a clarifying question is considered unhelpful, the annotator will categorize the failure into one of four error types as detailed in Deng et al. (2023b): wrong aspect, under-specified, over-specified, or generation error. Overall, we assess totally 400 queries and measure the inter-annotator agreement. We achieve an inter-annotator reliability of Krippen-dorff’s alpha of above 0.70 for all ambiguity categories in our taxonomy. In Table 12, we provide examples of generated clarifying questions for each error category.

E More Task Results

Table 9, 10, 11 present the results of all LLMs across different categories under three different settings: Zero-shot w/o CoT, Zero-shot w/ CoT, and Few-shot w/o CoT. We discover that while the exact values vary, the overall performance and analysis conclusions remain largely consistent with Sec 5.2.

F Human Validation and Revision

We initially engage 8 language experts via online platforms. Subsequently, they are assigned the task of reviewing 50 data samples according to provided instructions as part of a qualifying assessment. The 5 experts who successfully pass this assessment are then designated to validate and revise our dataset. For each query, they are given the respective ambiguity label and a corresponding clarifying question if the query is ambiguous. They are required to adhere to a specific protocol for validating and revising our dataset: Firstly, they need to verify if the query is ambiguous and if the ambiguity label assigned is accurate. Secondly, if the query is deemed ambiguous, they should evaluate whether the clarifying question effectively resolves any ambiguity. In instances of differing opinions during validation, discussions should be held to reach a consensus on the final data outcome. If significant disagreement persists even after discussion, the data will be discarded. We ensure the quality of our final data in two ways. The two authors of this paper acted as meta-reviewers, selecting 50 questions from each of the eight categories across the three dimensions in CLAMBER. The meta-reviewers assessed the correctness of ambiguity labels and the effectiveness of clarifying questions. For the 400 data samples, the average label accuracy was 92.4% and the average BLEU score was 73.2. Based on the results from the meta-reviewers, the data in CLAMBER is considered to be of high quality.

Methods	Epistemic Misalignment				Linguistic Ambiguity				Aleatoric Output							
	contradiction		unfamiliar		lexical		semantic		what		whom		when		where	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vicuna-13B	54.25	54.24	49.75	37.17	49.25	33.43	50.25	33.89	21.26	18.75	21.38	18.80	21.58	18.95	21.78	19.09
Llama2-13B-I	46.00	32.67	44.75	43.76	45.50	44.49	49.75	**48.55**	37.82	37.43	34.67	33.86	33.17	32.10	34.67	33.86
Llama2-13B	**64.25**	**59.01**	**50.75**	**44.42**	47.50	41.92	48.25	33.38	44.01	43.25	42.26	41.22	45.55	45.01	44.66	44.00
Llama2-70B	50.50	34.43	50.00	33.33	50.75	34.98	50.00	33.33	20.96	17.97	20.68	17.74	20.88	17.88	20.88	17.88
ChatGPT	39.50	30.10	**50.75**	36.59	**53.50**	**49.23**	**54.50**	44.75	**49.70**	**46.56**	**49.95**	**46.90**	**52.44**	**50.15**	**49.35**	**46.10**

Table 9: The fine-grained ambiguity identification evaluation results under Zero-shot w/o CoT setting.

Methods	Epistemic Misalignment				Linguistic Ambiguity				Aleatoric Output							
	contradiction		unfamiliar		lexical		semantic		what		whom		when		where	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vicuna-13B	**67.75**	**67.41**	52.25	44.79	56.50	51.54	52.25	44.56	37.23	37.00	36.16	35.85	36.86	36.61	38.16	38.01
Llama2-13B-I	39.50	28.32	48.50	48.37	48.50	48.03	53.75	**53.65**	38.32	36.52	37.86	35.96	37.96	36.08	38.26	36.46
Llama2-13B	51.25	36.46	50.50	34.43	49.50	33.11	50.00	33.33	24.45	22.86	24.18	22.61	24.08	22.53	24.88	23.17
Llama2-70B	67.50	63.66	50.75	36.20	53.00	39.67	50.00	33.33	22.16	19.59	21.78	19.27	21.78	19.27	21.88	19.35
ChatGPT	42.48	38.97	**55.25**	**55.24**	**74.00**	**72.79**	**54.00**	43.26	**65.70**	**53.44**	**64.35**	**50.61**	**64.00**	**49.93**	**63.30**	**48.42**

Table 10: The fine-grained ambiguity identification evaluation results under Zero-shot w/ CoT setting.

Methods	Epistemic Misalignment				Linguistic Ambiguity				Aleatoric Output							
	contradiction		unfamiliar		lexical		semantic		what		whom		when		where	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vicuna-13B	**50.00**	**33.33**	49.50	38.37	51.50	38.48	50.50	35.68	22.46	20.42	22.58	20.49	22.58	20.49	22.18	20.18
Llama2-13B-I	48.50	32.66	56.00	55.51	55.75	55.08	47.75	46.95	39.22	36.54	39.46	36.87	39.06	39.06	39.46	36.87
Llama2-13B	14.75	12.85	52.50	41.10	48.25	33.38	50.00	33.33	24.45	22.99	24.58	23.06	24.78	23.22	24.88	23.30
Llama2-70B	**50.00**	**33.33**	51.00	42.73	45.75	41.90	50.00	33.77	27.45	27.44	29.87	29.80	26.97	26.97	26.47	26.47
ChatGPT	39.00	28.05	**60.00**	**59.67**	**58.75**	**58.06**	**50.75**	**49.32**	**65.40**	**50.54**	**68.77**	**57.48**	**65.00**	**49.66**	**63.10**	**45.24**

Table 11: The fine-grained ambiguity identification evaluation results under Few-shot w/o CoT setting.

Wrong Aspect
Question: What is the dry mass of a seed of Dicent canadensis? **Ground Truth Clarifying Question**: What is the intended meaning of "Dicent canadensis"? Is it a specific plant species or a misspelling of a different plant name? **LLM-generated Clarifying Question**: Are you asking for the average dry mass of a seed of Dicentra canadensis or the specific dry mass of a particular seed?
Under-Specified
Question: What is the impact of being thrown? **Ground Truth Clarifying Question**: Are you referring to the physical act of being propelled through the air or the emotional state of being confused? **LLM-generated Clarifying Question**: Being thrown in what context or situation?
Over-Specified
Question: Give me the steps to purchase a new pair of shoes. **Ground Truth Clarifying Question**: Where do you want to buy new shoes, online of offline? **LLM-generated Clarifying Question**: Which specific steps are you referring to?
Generation Error
Question: What are some common symptoms of someone who is moped? **Ground Truth Clarifying Question**: Are you referring to someone who is feeling apathetic, gloomy, or dazed, or are you referring to a motorbike that can be pedaled or driven by a low-powered gasoline engine? **LLM-generated Clarifying Question**: #

Table 12: Examples of Error Analysis

Prompting Schemes
Zero-shot w/o CoT: Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is". Question: <Question>
Zero-shot w/ CoT: Given a query first identify whether the question is ambiguous or not. If it is ambiguous, ask a clarifying question. If it is not ambiguous, answer the question. The response should start with the ambiguity analysis of the question and then follow by "Therefore, the question is not ambiguous. The answer is" or "Therefore, the question is ambiguous. The clarifying question is". Question: <Question>
Few-shot w/o CoT: Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is". Question: Who played michael myers in rob zombie's movie? Output: The answer is Tyler Mane. Question: Give me some Mother's Day gift ideas. Output: The clarifying question is What are your mother's interests, hobbies, or something she has been wanting? Question: <Question>
Few-shot w/ CoT: Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is". Question: Who played michael myers in rob zombie's movie? Output: In Rob Zombie's "Halloween" films, the role of Michael Myers was primarily played by Tyler Mane. Mane portrayed the iconic horror character in both Rob Zombie's "Halloween" (2007) and its sequel, "Halloween II" (2009). Therefore, the question is not ambiguous. The answer is Tyler Mane. Question: Give me some Mother's Day gift ideas. Output: there could be underlying ambiguities depending on the interests of the specific mother in question, the budget, and the giver's relationship to the mother. Without this context, a broad range of ideas could be considered appropriate. Therefore, the question is ambiguous. The clarifying question is: What are the interests or hobbies of the mother, and is there a particular budget range for the gift? Question: <Question>

Table 13: Four prompting schemes for ambiguity identification and clarification.

ALCUNA dataset

The prompt of generating clarifying questions:
Given the user question: <question>.
Note that the <entity> is a non-existent entity fabricated by existing entities.
You need to generate a clarifying question about the <ENTITY> to better know its intended meaning.

Your Generated Clarifying Question:

Data Examples:
An ambiguous example
Query: What is the latitude of the habitat of inyidiidae?
Clarifying Question: Can you please provide more information about "inyidiidae"?

A unambiguous example
Query: Is Mozambique a geographic distribution of Mantodea?

Table 14: The prompt and data examples of the ALCUNA dataset

AmbiCoref dataset

The template of clarifying questions:
What does <PRONOUN> refer to? <A> or <B>?

Data Examples:
An ambiguous example
Query: Matthew bought Joshua a pizza after he asked for more food. Who asked for more food?
Clarifying Question: What does he refer to? Matthew or Joshua?

A unambiguous example
Query: Matthew made Joshua a square pizza before he submitted the order. Who submitted the order?

Table 15: The template and data examples of the AmbiCoref dataset

AmbiTask dataset

The template of rephrasing ambiguous queries:
The all possible word categories are either <category 1> or <category 2>.
The following two examples share a specific word category. You need to first infer the specific word category from the examples.
Please output "X" if the given sentence mentions the specific word category. Please output "Y" if the given sentence does not mention the word category.

Examples:
The photographer is not in the restaurant.
Bernie Sanders has been in the theatre.

The Given Sentence:
Paul Atreides may not be in the hotel lobby.

The template of rephrasing unambiguous queries:
Please output "X" if the given sentence contains a word of <category>. Please output "Y" if the given sentence does not contain any word of <category>.

Examples:
The fugitive has not been in the museum.(Output: X)
Noam Chomsky was in the film studio.(Output: Y)

The Given Sentence:
The hiker was in the laboratory.

The template of clarifying questions:
Is the category either <category 1> or <category 2>?

Data Examples:
An ambiguous example
Query: The all possible word categories are either "does not contain a negation" or "proper noun".
The following two examples share a specific word category. You need to first infer the specific word category from the examples.
Please output "X" if the given sentence mentions the specific word category. Please output "Y" if the given sentence does not mention the word category.

Examples:
The photographer is not in the restaurant.
Bernie Sanders has been in the theatre.

The Given Sentence: Paul Atreides may not be in the hotel lobby.
Clarifying Question: Is the category either does not contain a negation or proper noun?

A unambiguous example
Query: Please output "X" if the given sentence contains a word of "common noun". Please output "Y" if the given sentence does not contain any word of "common noun".

Examples:
The fugitive has not been in the museum.(Output: X)
Noam Chomsky was in the film studio.(Output: Y)

The Given Sentence: The hiker was in the laboratory.

Table 16: The prompt, clarifying question template and data examples of the AmbiTask dataset

AmbER dataset
The prompt of generating ambiguous queries and clarifying questions: ############### <QUESTION 1> <QUESTION 2> ############### According to the above example questions, Note that <ENTITY> is an ambiguous entity and has multiple meanings. You should generate a new question using the <ENTITY> and random context. You need to make sure the generated question is ambiguous and answering the generated question requires further clarification. FORMAT: {"question": <STRING>, "clarifying_question": <STRING>}
The prompt of generating unambiguous queries: Given an ambiguous query and its clarifying question, you need to generate a unambiguous query based on them. FORMAT: {"unambiguous query": <STRING>}
Data Examples: An ambiguous example Query: What is the history of Alcatraz? Clarifying Question: Are you referring to the history of the Alcatraz Island or the history of the band Alcatraz? A unambiguous example Query: What are the tracks in the album or soundtrack called Birds?

Table 17: The prompt and data examples of the AmbER dataset

AmbiPun dataset

The prompt of generating ambiguous queries and clarifying questions:
//1. Generate ambiguous queries Given a polysemy word <WORD>, it has two senses, including of <SENSE1> and <SENSE2>.
You need to generate an information-seeking question based on the word <WORD>.
You need to make the generated question be ambiguous due to the polysemy of word <WORD>.
Note the question needs to contain the word <WORD>.
Answering the generated requires a clarifying question to better understand the word <WORD>.
generated question:

//2. Generate clarifying question
Given a question: <QUESTION>
Note the polysemy word <WORD> has two senses, including of <SENSE1> and <SENSE2>.
The given question has ambiguity due to the polysemy word <WORD>.
You need to generate a clarifying question based on the word <WORD> to better clarify the ambiguity of the given question.
clarifying question:

The prompt of generating unambiguous queries:
Given an ambiguous query and its clarifying question, you need to generate a unambiguous query based on them.
FORMAT: {"unambiguous query": <STRING>}

Data Examples:
An ambiguous example
Query: What is the meaning of Smart?
Clarifying Question: Are you referring to the adjective ’smart’ or a specific brand called ’Smart’?

A unambiguous example
Query: What are the common strategies for saving money?

Table 18: The prompt and data examples of the AmbiPun dataset

AmbigQA dataset

Data Examples:
An ambiguous example
Query: Who played kelly on the drew carey show?
Clarifying Question: Which role: Kellie Newmark, Marlo Kelly, Grace Kelly, or Kelly Walker?

A unambiguous example
Query: Where did they film ash vs evil dead?

Table 19: The prompt and data examples of the AmbigQA dataset

Dolly dataset

The prompt of generating clarifying questions and category classification:
Give you an instruction, you first need to judge whether the instruction is ambiguous or not.
If you think the instruction is ambiguous and falls into one of the following ambiguous types,
you need to output its ambiguous type and the corresponding clarifying questions to help answer the ambiguous instruction.
If you think the instruction is not ambiguous and does not miss any specific information,
you need to rewrite it and make sure it falls into one of the following ambiguous types.
Ambiguous types:
1. Missing personal information.
For example, the instruction "Suggest me some good movies" misses the information of the user personal preference.
2. Missing spatial information.
For example, the instruction "How to reach a destination" misses the spatial information of the departure location.
3. Missing temporal information.
For example, the instruction "Make a restaurant reservation" misses the temporal information of the reservation time.
4. Missing specific task-related information.
For example, the instruction "convert string to int" misses the information of the programming language.

You should output the ambiguous type, the ambiguous instruction and its corresponding clarifying questions for each instruction.
FORMAT: {"ambiguous type": <STRING>, "ambiguous instruction": <STRING>, "clarifying question": <STRING>}

Data Examples:
An ambiguous example
Query: Give me some Mother’s Day gift ideas
Clarifying Question: What are your mother’s interests, hobbies, or something she has been wanting?

A unambiguous example
Query: Top scorer of uefa champions league of all time?

Table 20: The prompt and data examples of the Dolly dataset