Disambiguation in Conversational Question Answering in the Era of LLM: A Survey

Anonymous ACL submission

Abstract

Ambiguity remains a fundamental challenge in Natural Language Processing (NLP) due to the inherent complexity and flexibility of human language. With the advent of Large Language Models (LLMs), addressing ambiguity has become even more critical due to their expanded capabilities and applications. In the context of Conversational Question Answering (CQA), this paper explores the definition, forms, and implications of ambiguity for language driven systems, particularly in the context of LLMs. We define key terms and concepts, categorize various disambiguation approaches enabled by LLMs, and provide a comparative analysis of their advantages and disadvantages. We also explore publicly available datasets for benchmarking ambiguity detection and resolution techniques and highlight their relevance for ongoing research. Finally, we identify open problems and future research directions, proposing areas for further investigation. By offering a comprehensive review of current research on ambiguities and disambiguation with LLMs, we aim to contribute to the development of more robust and reliable language systems.

1 Introduction

The inherent ambiguity in natural language communication presents a fundamental challenge in human-AI interactions, especially in conversational systems. Modern AI Assistants, such as Adobe's AEP AI Assistant[1] and Amazon's Rufus[2], must navigate these ambiguities through advanced language understanding mechanisms. The ability to accurately determine the intended meaning of a term or phrase within a given context is fundamental to enhancing the performance of such conversational systems. This mirrors human cognitive behavior, where communicators must anticipate potential misunderstandings, while recipients engage in active disambiguation through contextual analysis (Anand et al., 2023), clarifying questions (Zamani et al., 2020; Zhang et al., 2024c), and continuous interpretation refinement (Zukerman and Raskutti, 2002; Jones et al., 2006).

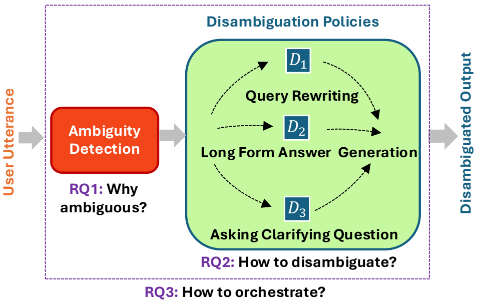

Figure 1: Broadly, we categorize the existing literature to answer three major research questions (RQs), namely, why ambiguous (RQ1), how to disambiguate (RQ2), and how to orchestrate (RQ3).

The advent of Large Language Models (LLMs) has further underscored the importance of understanding and resolving ambiguity to enhance the performance and reliability of language understanding systems. As LLMs become increasingly integral to applications, such as search engines or Information Retrieval (IR) (Anand et al., 2023; Ma et al., 2023), Conversational Question Answering (CQA) (Zhang et al., 2020; Thoppilan et al., 2022; Xu et al., 2023), automated text summarization (Kurisinkel and Chen, 2023; Zakkas et al., 2024) and so on, their ability to manage ambiguous language is essential for effective communication and user satisfaction. This is because their utility can often be compromised by ambiguous user queries, which can lead to incorrect or irrelevant outputs (Kuhn et al., 2022; Deng et al., 2023a).

While disambiguation techniques have witnessed significant advancements over recent decades, driven by sophisticated algorithms (Raganato et al., 2017; Zhang et al., 2018; Rao and Daumé III, 2018, 2019; Xu et al., 2019; Alianne-

[1] business.adobe.com/products/sensei/ai-assistant.html
[2] aboutamazon.com/news/retail/how-to-use-amazon-rufus

jadi et al., 2019; Kumar and black, 2020; Min et al., 2020; Zamani et al., 2020; Guo et al., 2021; Kuhn et al., 2022; Lee et al., 2023), the inherent complexity of natural language and the need for large annotated corpora has been continuing to pose substantial challenges. For these reasons, an emerging and active area of research is to explore the capacity of LLMs themselves to identify and resolve ambiguous queries (Liu et al., 2023; Mehrparvar and Pezzelle, 2024; Zhang and Choi, 2023; Zhang et al., 2024c; Anand et al., 2023). While LLM-based disambiguation techniques are gaining popularity, the field lacks a systematic analysis and categorization of existing methods. This paper addresses that gap by surveying current LLM-based approaches for ambiguity detection and disambiguation, outlining their underlying principles, strengths, and limitations. Among the NLP tasks, we primarily focus on CQA as this task seems to be prominent in majority of the use-cases.

Organization of this Survey. We structure this survey around three core research questions (see Figure 1): RQ1: Why do ambiguities arise in language, and how can we detect them? RQ2: How can we disambiguate, particularly using LLMs? RQ3: How can we automate disambiguation strategies in real-world applications? Section 2 addresses RQ1 by defining key concepts, presenting a taxonomy, and reviewing ambiguity detection methods. Section 3 tackles RQ2 by categorizing LLM-based disambiguation approaches and analyzing their strengths and weaknesses. To support these, Section 4 surveys relevant public datasets used for benchmarking. Finally, Section 5 explores open challenges and outlines future directions, centering on RQ3: how to orchestrate disambiguation effectively in practice.

2 *Why Ambiguous?*

2.1 Definition of Ambiguity

Ambiguous queries are typically those that have multiple distinct meanings, insufficiently defined subtopics (Clarke et al., 2009), syntactic ambiguities (Schlangen, 2004), for which a system struggles to interpret accurately, resulting in inappropriate or unclear answers (Keyvan and Huang, 2022). These ambiguities can arise at lexical, syntactic, or semantic levels, motivating the development of various taxonomies, which we present in the next section.

2.2 Taxonomy of Ambiguity

Existing literature approaches the taxonomy of ambiguities in various ways, often influenced by specific use-cases, public datasets, or the scope defined for new data collection. For instance, Tanjim et al. (2025) focuses on industrial conversation question answering, while Zhang et al. (2024c) examine ambiguities through public datasets. Additionally, Liu et al. (2023) define their own criteria for collecting new datasets, further diversifying the landscape of ambiguity taxonomies. This complexity is compounded by the various NLP tasks to which these taxonomies are applied. For example, Natural Language Inference (NLI), Question Answering (QA), and Machine Translation (MT) each have unique requirements and interpretations of ambiguity, as explored by Zhang and Choi (2023). Consequently, different taxonomies have emerged from these diverse focuses. Moreover, the same example can be treated differently across various studies. For instance, Zhang et al. (2024c) categorized the example *"Real name of gwen stacy in amazing spiderman?"* as an Aleatoric 'What' type of ambiguity. In contrast, Zhang and Choi (2023) classified this as a 'Literal vs. Implied interpretation' ambiguity. This discrepancy underscores the need for a unified approach to taxonomy.

In Table 1, we present a comparative analysis of these taxonomies to highlight common grounds despite their differences. To cater to broader applications and provide clarity, we propose simplifying existing taxonomies into three overarching categories. We argue that these categories can encompass all existing taxonomies, irrespective of the underlying tasks, thereby offering a more cohesive framework for understanding ambiguities.

Syntactic Ambiguity: When a sentence can be parsed in different ways (Church and Patil, 1982; Wasow, 2015). For example, *'I saw the man with a telescope.'* Here the ambiguity arises because it could be interpreted in two ways: did the speaker see the man 'with the telescope' or did the speaker see 'the man' using the telescope? This taxonomy is listed in both Tanjim et al. (2025) and Liu et al. (2023), but it seems to be missing in the other two.

Semantic Ambiguity: When a sentence is grammatically correct but semantically unclear, due to ambiguity in a word, phrase, or the overall interpretation. The more common case involves ambiguity at the word or phrase level, often referred to as *lexical ambiguity* (Navigli, 2009; Beekhuizen et al.,

Literature	Taxonomy		
	Type	**Definition Provided by the Literature**	**Example Given**
Tanjim et al. (2025)	Pragmatic	The meaning of a sentence depends on the context, reference, or scope.	*“How many do I have?”*
	Syntactic	The structure of a sentence is incomplete or allows for multiple interpretations.	*“Business event”*
	Lexical	The meaning of the word/term is not clear or has multiple interpretations.	*“Are we removing abc123 from XYZ?”*
Zhang et al. (2024c)	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?”*
	Lexical	Query contains terms with multiple meanings.	*“Tell me about the source of Nile.”*
	Semantic	Query lacks context leading to multiple interpretations.	“When did he land on the moon?”
	Aleatoric	Query output contains confusion due to missing personal/temporal/spatial/task-specific elements.	*“How many goals did Argentina score in the World Cup?”*
Liu et al. (2023)	Pragmatic	Literal and pragmatic interpretations are present.	*“I’m afraid the cat was hit by a car.”*
	Lexical	A lexical item has different senses.	*“John and Anna are married.”*
	Syntactic	Different syntactic parses lead to different interpretations.	*“This seminar is full now, but interesting seminars are being offered next quarter too.”*
	Scopal	Ambiguity from the relative scopal order of quantifiers or the scope of particular modifiers.	*“The novel has been banned in many schools because of its explicit language.”*
	Coreference	Ambiguous coreference.	*“It is currently March, and they plan to schedule their wedding for next December.”*
Zhang and Choi (2023)	Word-Sense Disambiguation	Word-sense disambiguation for named entities, also commonly surfaces as entity linking ambiguities.	*“Who wins at the end of friday night lights?”*
	Literal vs. Implied Interpretation	A question literally means something different from what the user probably meant to ask.	*“The cake was so dry, it was like eating sand.”*
	Multiple Valid Outputs	Ambiguity due to multiple valid outputs.	*“When did west germany win the world cup?”*

Table 1: Here, we present several taxonomies exactly as they appear in the existing literature, along with their definitions and examples (ambiguous parts of the text are underlined). As can be seen there are redundancies in these definitions, highlighting the need for a unified taxonomy.

2021), where a term has multiple possible meanings. As shown in Table 1, this type is listed across most prior work, with the exception of Zhang and Choi (2023), where they mention it as ‘word sense disambiguation.’ Similarly, the ‘Unfamiliar’ category in Zhang et al. (2024c) aligns with this type, as unknown words are inherently open to interpretation until contextual or domain-specific knowledge is applied. Beyond word-level issues, semantic ambiguity can also stem from interpretive variation at the sentence level. This includes the usage of literal vs. pragmatic words as mentioned by Liu et al. (2023), who refer to it as pragmatic ambiguity, and ‘Literal vs. Implied Interpretations’ by Zhang and Choi (2023). The ‘Figurative’ type in Liu et al. (2023) also falls into this category, as does the ‘Contradiction’ category in Zhang et al. (2024c) because of conflicts with the semantics of previous statements.

Contextual Ambiguity: When the context of the conversation is missing or the answers could be multiple unless no specific context is given (e.g., what/when/where/who type of questions without context) (Sperber and Wilson, 1986; Huang, 2017). Tanjim et al. (2025) name this as pragmatic ambiguity, whereas it is listed as ‘Semantics’ in Zhang et al. (2024c) and as ‘Aleatoric’, ‘Coreference’ and ‘Scopal’ in Liu et al. (2023), and as ‘Multiple Valid Outputs’ in Zhang and Choi (2023). Meanwhile, ‘Knowledge Conflict‘, as described by Neeman et al. (2022); Shaier et al. (2024), also aligns with this type, occurring when a question lacks specific context, such as temporal or locational cues, causing retrieval-augmented models to face conflicts between retrieved and parametric knowledge.

2.3 Ambiguity Detection

The body of work for detecting ambiguity can be broadly categorized into three major groups: traditional methods (not language model-based), language model-based methods, and large language model (LLM)-based methods. In Table 2, we sum-

Literature	Approach	Inputs	Ambiguity Type
Trienes and Balog (2019)	Logistic regression + features	Q, tags, similar Qs	Syntactical
Dhole (2020)	BiLSTM classifier	Dialogue, intents	Contextual
Guo et al. (2021)	BERT classifier	Conv., passage	Semantic, Contextual
Lee et al. (2023)	BERT classifier	Q, passages	Contextual
Tanjim et al. (2025)	ST + rules + features	Q only	Syntactical, Semantic, Contextual
Kuhn et al. (2022)	Prompted LLM	Q Only	Contextual
Zhang et al. (2024c)	Prompted LLM	Q, context (optional)	Semantic, Contextual
Zhang and Choi (2023)	LLM + CoT by ambiguity type	Q, prompt schema	Semantic, Contextual
Kim et al. (2024)	LLM + uncertainty signals	Q only	Semantic, Contextual

Table 2: Summary of ambiguity detection methods. Shaded by method type: traditional (gray), LM (cyan), LLM (pink). Here, ST= Sentence Transformer, Q=Question, Conv.= Conversation.

marize each method's approach, model inputs, and the types of ambiguity it addresses based on our taxonomy. We give more details below.

Traditional Methods: Early research into ambiguity detection primarily concentrated on binary classification methodologies. A significant contribution in this domain was made by Trienes and Balog (2019), who used logistic regression on features from similar questions in community QA forums. Their model and features targeted queries that have a defect in their structure, thereby focusing on *syntactical ambiguity*. While offering interpretability, their scope was limited to single-turn QA and did not account for other ambiguity types such as semantic or contextual ambiguities in dialogue-based settings. To address some of these limitations, Dhole (2020) proposed a two-stage approach for resolving ambiguous user intents in task-oriented dialogue. Their work falls under *contextual ambiguity*, as their classifier disambiguates underspecified user intents.

Language Model-Based Methods: In the realm of language model-based methods, Guo et al. (2021) introduced Abg-CoQA, a benchmark dataset and framework for ambiguity detection and clarifying question generation in conversational QA. Their model addressed both *semantic* and *contextual* ambiguities owing to their framing ambiguity detection as a QA classification task (thus capable of understanding the semantic ambiguity). However, even with BERT-based models, performance remained low (23.6% F1). Similarly, Lee et al. (2023) proposed a BERT-based classifier to detect ambiguity given a passage, but their model also exhibited low performance. Their work primarily focused on *contextual ambiguity*, where a question can lead to multiple valid answers without further specification. A more recent study by Tanjim et al. (2025) employed Sentence Transformers with handcrafted rules and features to detect all three ambiguity types—*syntactic*, *semantic*, and *contextual*—demonstrating that explicit modeling of ambiguity categories can improve detection.

LLM-Based Methods: With the advent of large language models (LLMs), ambiguity detection has increasingly shifted toward prompt-based methods. Kuhn et al. (2022) demonstrated that LLMs could be prompted to decide whether to answer a query or ask for clarification. Their method targeted primarily *contextual ambiguity*, especially in cases of underspecified user queries. Zhang et al. (2024c) introduced CLAMBER, a benchmark with a taxonomy of eight ambiguity types. They showed that LLMs can identify certain *semantic* (e.g., lexical or referential ambiguity) and *contextual* ambiguities, but struggle with systematic disambiguation. Zhang and Choi (2023) proposed a prompting method that asks the model to reason about ambiguity types before generating a clarifying question. Their framework covers both *semantic* and *contextual ambiguity*, aligning clarification strategies with the predicted ambiguity type. Finally, Kim et al. (2024) presented a method where LLMs use their internal uncertainty to decide whether a query is ambiguous. Their alignment framework quantifies information gain through clarification, capturing *semantic* ambiguities (e.g., polysemous terms) and *contextual* ones (e.g., missing scope or domain).

Despite the flexibility of LLMs, these works collectively show that ambiguity detection—particularly fine-grained distinctions among types—remains a complex problem. We will revisit these challenges in Section 5.

3 *How To Disambiguate?*

In the era of LLMs, disambiguation is gaining increasing attention due to their extensive world knowledge and advanced capabilities, surpassing traditional and smaller language models. However, current research in this area often lacks systematic categorization and tends to address various aspects in isolation. To that end, in this paper, we argue existing disambiguation works fall in three major policies, which we present in Figure 2. We describe each of them below.

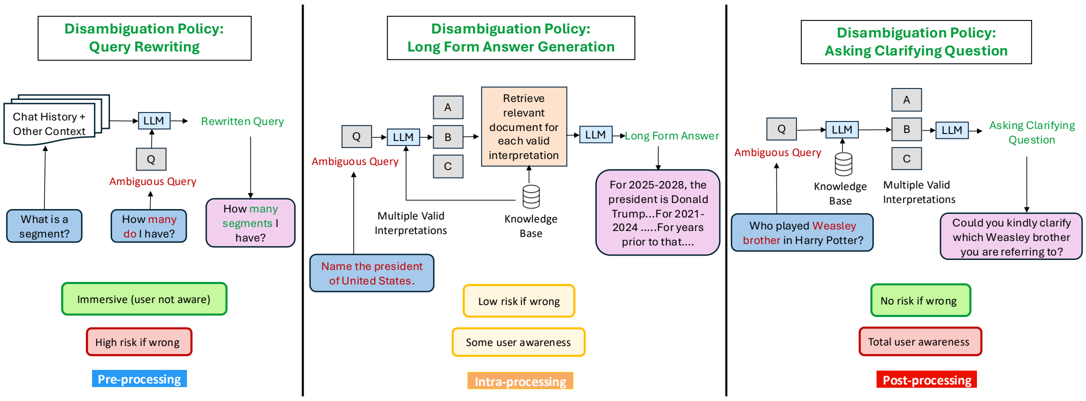
Disambiguation Policy: Query Rewriting
Chat History + Other Context
LLM
Rewritten Query
Q
Ambiguous Query
What is a segment?
How many do I have?
How many segments I have?
Immersive (user not aware)
High risk if wrong
Pre-processing
Disambiguation Policy: Long Form Answer Generation
Q
LLM
A
B
C
Retrieve relevant document for each valid interpretation
LLM
Long Form Answer
Ambiguous Query
Multiple Valid Interpretations
Knowledge Base
For 2025-2028, the president is Donald Trump...For 2021-2024 .....For years prior to that....
Name the president of United States.
Low risk if wrong
Some user awareness
Intra-processing
Disambiguation Policy: Asking Clarifying Question
Q
LLM
A
B
C
LLM
Asking Clarifying Question
Ambiguous Query
Knowledge Base
Multiple Valid Interpretations
Who played Weasley brother in Harry Potter?
Could you kindly clarify which Weasley brother you are referring to?
No risk if wrong
Total user awareness
Post-processing

Figure 2: We find existing disambiguation efforts using LLMs broadly fall into these three major categories: Left. Query Rewriting, Middle. Long Form Answer Generation, Right. Asking Clarifying Questions. These policies have different requirements and also work in different processing steps in CQA pipeline, resulting in unique advantages and disadvantages for each approach. We highlight a couple here and provide a more comprehensive list in Table 5.

3.1 Query Rewriting (QR)

Query rewriting (QR) represents a wide span of techniques that transforms ambiguous or unclear user queries into well-defined, comprehensive expressions (Carpineto and Romano, 2012). Early work focused on query expansion (Carpineto and Romano, 2012; Lavrenko and Croft, 2017), contextual rephrasing (Zukerman and Raskutti, 2002), and synonym-based augmentation (Jones et al., 2006). Prior to LLM, research demonstrates significant advances in neural query rewriting through supervised learning approaches (Elgohary et al., 2019; Anantha et al., 2021) and reinforcement learning frameworks (Vakulenko et al., 2021). Other innovations have explored explicit reasoning patterns (Qian and Dou, 2022) achieving good performance in transforming ambiguous queries into precise, answerable questions.

The emergence of LLMs has enabled more advanced query reformulation, moving beyond term-based edits to deeper semantic understanding and contextual refinement across downstream tasks (Wang et al., 2023). Recent research works, such as Ma et al. (2023); Jagerman et al. (2023), have demonstrated the efficacy of LLM-based query reformulation in zero-shot and few-shot settings, particularly valuable when domain-specific training data is scarce. The principled way of QR is shown in Figure 2 (Left), where an LLM is prompted with previous chat history and other relevant information as context. Some advanced prompting, such as Ye et al. (2023) also includes "rewrite-then-edit" framework. Apart from prompting, LLMs also have been fined-tuned (Peng et al., 2024) or used to generate Supervised Fine Tuning (SFT) dataset to improve QR model either through a re-ranker (Mao et al., 2024) or preference optimization (Zhang et al., 2024b).

3.2 Long Form Answer Generation (LFAG)

Generating long-form answers to ambiguous questions involves presenting all valid interpretations alongside their corresponding answers. For instance, the question "Who has the highest goals in world football?" can refer to either men's or women's football. A well-structured response would be: "Ali Daei holds the record in men's football, while Christine Sinclair does in women's football." As shown in Figure 2 (Middle), this task typically comprises three steps: **1)** Disambiguating the question, **2)** Answering each interpretation, and **3)** Consolidating the results into a single, coherent response. Early methods streamlined these steps into a single model inference. Stelmakh et al. (2022) finetuned T5 to directly produce long-form answers. More recent LLM-based approaches, such as Gao et al. (2023), show that few-shot prompting can be similarly effective without fine-tuning. To reduce reasoning load, Amplayo et al. (2022) proposed a two-step method: first inferring multiple interpretations, then generating a long-form answer from them. RAC (Kim et al., 2023) introduced retrieval-augmented disambiguation to generate answers with supporting evidence (Steps 1–2), while ToC (Kim et al., 2023) extended this via iterative retrieval to capture overlooked interpretations, trading off efficiency. DIVA (In et al., 2024) improved efficiency by modeling a reasoning chain that compresses this process into a single step, maintaining performance while reducing complexity.

Technique	Syntactic	Semantic	Contextual
QR	✓	✓	✓
LFAG	✗	✓	✓
ACQ	✗	✓	✓

Table 3: Disambiguation techniques and the types of ambiguity they are equipped to handle.

3.3 Asking Clarifying Question (ACQ)

This is one of the most extensively studied disambiguation policies, with approaches ranging from rule-based prompts (e.g., "Did you mean A or B?" (Coden et al., 2015), "What do you want to know about QUERY?" (Zamani et al., 2020), or category-based options (Lee et al., 2023)) to traditional machine learning (Zhang et al., 2018; Rao and Daumé III, 2018, 2019) and language model-based methods (Xu et al., 2019; Aliannejadi et al., 2019). Several works also introduce new datasets (Xu et al., 2019; Kumar and black, 2020; Min et al., 2020; Guo et al., 2021), discussed further in Section 4. However, these methods often struggle with complex queries and rely on annotated corpora, which could be difficult to obtain.

With the advent of LLMs, recent studies have leveraged prompt-based approaches (Kuhn et al., 2022; Deng et al., 2023b; Zhang et al., 2024c), typically employing zero-shot or few-shot Chain-of-Thought (CoT) prompting strategies. These methods mirror the Long-form Answer Generation pipeline but focus on analyzing multiple valid interpretations to generate clarifying questions, as shown in Figure 2 (Right). Like QR, they reduce the need for domain-specific data and can be training-free while supporting complex question structures. Some works adopt a two-stage pipeline: first detecting ambiguity, then generating suitable clarification questions. For instance, Zhang and Choi (2023) proposed an innovative uncertainty estimation technique for ambiguity detection that quantifies intent entropy through simulated user-assistant interactions. Finally, similar to QR, LLMs can be also be fine-tuned to generate clarifying questions. For example, Zhang et al. (2024a); Kim et al. (2024) fine-tuned various LLMs, such as `Llama-2-7B` (Touvron et al., 2023), `Gemma-7B` (Team et al., 2024), and `Llama-3-8B` (Dubey et al., 2024).

Table 3 summarizes how disambiguation techniques address different ambiguity types. **QR** handles all three by reformulating queries to fix syntactic issues, resolve semantic confusion through inferred interpretations, and incorporate missing contextual details from prior conversation. **LFAG** handles semantic and contextual ambiguity by presenting multiple plausible interpretations, including those that differ semantically as well as those that are plausible when considering different contexts. **ACQ** resolves semantic and contextual ambiguity by explicitly asking the user to confirm among similar options or supply missing information. While QR might look most appealing for its broad coverage, it still faces key challenges such as semantic drift (Anand et al., 2023) and practical concerns like latency, cost, and error propagation in production (Tanjim et al., 2025). We will discuss the strengths and limitations of each approach further in Section 5.

4 Benchmarks

To evaluate disambiguation strategies, prior work has introduced task-specific benchmark datasets and metrics, which we describe below.

Ambiguity Detection and ACQ. Most existing datasets related to ambiguity fall into the category of detecting the need for clarification and necessary disambiguation by asking clarification questions. Notable datasets in this area include CLAQUA (Xu et al., 2019), ClarQ (Kumar and black, 2020), AmbigNQ (Min et al., 2020), ClariQ (Aliannejadi et al., 2020), Abg-CoQA (Guo et al., 2021), PACIFIC (Deng et al., 2022), CAmbigNQ (Lee et al., 2023), and CLAMBER (Zhang et al., 2024c). These corpora exhibit significant variation across several dimensions, each contributing uniquely to the understanding of ambiguities in dialogue systems, as listed in Table 4. Among them, the CLAMBER benchmark (Zhang et al., 2024c) has emerged as the first comprehensive evaluation benchmark for LLM-based ambiguity detection and ACQ, providing valuable insights into the current limitations of LLM-based approaches and establishing baseline metrics for future research. Statistics for all these datasets, along with their corresponding URLs, appear in Table 4. Metrics typically used for ambiguity detection include classification metrics such as Precision, Recall, F1, Accuracy, and AUROC score (Zhang et al., 2024c; Tanjim et al., 2025). For ACQ, the metrics are usually automatic text evaluation metrics, such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004). However, some studies criticize the limitations of these metrics and favor human judgment instead (Zamani et al., 2020).

Paper	Name	Domain	Core Unit	Scale	# Ambiguous	Link
			Ambiguity Detection and Asking Clarifying Question			
Xu et al. (2019)	**CLAQUA**	Open-domain	Q w/ Ans. (ST + MT)	17K + 22K	7K + 9K	github.com/msra-nlc/MSParS_V2.0
Kumar and black (2020)	**ClarQ**	Stack Exchange	Q w/ Context	6M	2M	github.com/vaibhav4595/ClarQ
Min et al. (2020)	**AmbigNQ**	Wikipedia	Q w/ Ans. (Tr/Vl/Te)	10K / 2K / 2K	4K / 1K / 1K	nlp.cs.washington.edu/ambigqa
Guo et al. (2021)	**Abg-CoQA**	Stack Exchange	P + Q	4K + 8K	800+ / 900+	github.com/MeiqiGuo/AKBC2021-Abg-CoQA
Aliannejadi et al. (2021)	**ClariQ**	TREC, Qulac	Conv. + Clar.Q	11K + 1M	Rated	github.com/aliannejadi/ClariQ
Deng et al. (2022)	**PACIFIC**	TAT-QA	Conv. + Q w/ Context & Ans.	2K + 19K	2K	github.com/dengyang17/PACIFIC
Lee et al. (2023)	**CAmbigNQ**	AmbigNQ	Clar.Q + Ans. + P	4K + 400+ + 400+	All Ambig.	github.com/DongryeolLee96/AskCQ
Zhang et al. (2024c)	**CLAMBER**	Mixed	Q w/ Context	12K	5K	github.com/zt991211/CLAMBER
			Query Rewriting			
Elgohary et al. (2019)	**CANARD**	QUAC	Q + Rewrite	40K + 40K	N/A	canard.qanta.org
Anantha et al. (2021)	**QReCC**	QUAC, NQ, TREC-C	Conv. + Q + Rewrite	13K + 80K + 80K	N/A	github.com/apple/ml-qrecc
			Long Form Answer Generation			
Stelmakh et al. (2022)	**ASQA**	Wikipedia, AmbigNQ	Q w/ LF Ans. (Tr/Vl/Te)	4K / 900+ / 1K	All Ambig.	github.com/google-research/language

Table 4: Publicly available datasets for benchmarking ambiguity in QA, covering both ambiguous and non-ambiguous cases (except ASQA, CANARD, QReCC). Rows are task-grouped and color-coded by size: large (pink), medium (cyan), small (yellow). "Core Unit" abbreviates data structure: Tr=Train, Vl=Val, Te=Test, P=Passage, Q=Question, Ans.=Answer, Clar.Q=Clarifying Q., Conv.=Conversation, LF=Long Form, Context=Passage/Table/Post (depends on the dataset), Rated=All questions rated from 1 (clear) to 4 (ambiguous).

Query Rewriting. There are two prominent benchmark datasets for evaluating the quality of rewritten queries. The pioneering dataset in this area is CANARD (Elgohary et al., 2019), which includes questions with context and their rewritten versions. This was followed by QReCC (Anantha et al., 2021), where each user question is accompanied by a human-rewritten query, and answers to questions within the same conversation may be distributed across multiple web pages. Notably, QReCC is used in recent LLM-based QR approaches such as Ye et al. (2023) and Zhang et al. (2024b). Both of these datasets, along with their statistics and URLs, are listed in Table 4. It is important to note that, unlike datasets related to ACQ, these datasets do not contain specific fields or labels explicitly indicating 'ambiguity' in queries. As for metrics, similar to ACQ, BLEU and ROUGE are popular choices for measuring the quality of rewritten queries. Additionally, since QR is often employed for IR tasks, standard IR metrics such as mean reciprocal rank (MRR), mean average precision (MAP), and Recall@k and Precision@k are used to evaluate whether the rewritten query retrieves the correct information (Ma et al., 2023; Ye et al., 2023). For these purposes, popular open-domain QA datasets like NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and HotpotQA (Yang et al., 2018) are often used as benchmarks. However, we do not list them here as they do not focus specifically on ambiguity and lack corresponding human-rewritten queries.

Long Form Answer Generation. To the best of our knowledge, ASQA (Stelmakh et al., 2022) is the only dataset that falls into this category. ASQA is a long-form QA dataset derived from a subset of ambiguous questions in the AmbigNQ dataset (Min et al., 2020). Its statistics and corresponding URL are provided in Table 4. The dataset is designed to evaluate how well systems can generate comprehensive answers that cover all valid interpretations. Two main metrics are used to assess the generation quality (In et al., 2024): Disambig-F1 (D-F1) (Stelmakh et al., 2022), which assesses the accuracy of responses by verifying correct answers to disambiguated questions using an F1 score, and ROUGE, which evaluates the correctness by comparing them to ground-truth long-form answers.

Disambiguation Policy	Automatic?	Additional LLM Call?	Visible to User?	High Risk?	UX Disrupting?
Query Rewriting	Yes	Yes	No	Yes	No
Long Form Answer Generation	Yes	Maybe	Yes	No	Maybe
Asking Clarification Question	No	Yes	Yes	No	Yes

Table 5: Comparison of disambiguation policies across key dimensions. Trait colors: Green = positive, Red = negative, Yellow = context-dependent. No single policy suffices, motivating an agentic framework to coordinate.

5 Open Problems and Challenges

Detecting Ambiguities. While LLMs have exceptional generative capabilities, recent studies consistently highlight the challenges of using LLMs to detect ambiguous queries with high performance. For example, Zhang and Choi (2023) achieved an AUROC of 0.57 on AmbigNQ (Min et al., 2020) using `LLaMA-2-13B-Chat`, while Zhang et al. (2024c) reported a best F1 score of 0.53 on their dataset using `GPT-3.5-Turbo`. Tanjim et al. (2025) shares a similar study and highlight a relatively lower performance using `GPT-3.5-Turbo` and `LLaMA-3.1-70B`. One potential reason, as suggested by Liu et al. (2023), is that LLMs are not inherently designed to model ambiguities.

How To Orchestrate? This is one of the research questions we posed at the beginning. To first see why we need to ochestrate among the disambiguation policies, in this paper, we systematically analyze the pros and cons of each disambiguation policy, making us the first to do so to the best of our knowledge. We show the list in Table 5, which are: *1) Automatic:* Both QR and LFAG are automatic and do not require human validation, unlike clarifying questions. *2) Additonal LLM Call:* For CQA, at least one LLM call is needed for answer generation, and so LFAG could be integrated into that same LLM call. But both QR and ACQ require dedicated LLMs. *3) Visible to User:* Rewritten queries are not typically visible to the user, whereas users might notice long-form answers and are definitely aware of clarifying questions. *4) High Risk:* Each policy affects different processing steps; for example, QR impacts downstream tasks significantly, as incorrect assumptions can lead to wrong answers. *5) UX Disrupting:* Repeated QR does not affect user experience as it is not visible, but too many clarifying questions can vex users. LFAG falls in between, as overly long answers are sometimes unwelcome. As can be seen, each approach has unique strengths and weaknesses, necessitating the need of coordination. The challenge lies in determining when to use which policy. For example, always asking clarifying questions can disrupt UX while always rewriting queries can lead to errors (Tanjim et al., 2025).

Opportunities. One promising direction for improving disambiguation systems is the development of an agentic framework (Wu et al., 2023; Zeng et al., 2023) that can intelligently select and coordinate various disambiguation strategies. However, this approach faces two key challenges: reliably detecting ambiguities and selecting appropriate disambiguation policies. To address ambiguity detection, recent advances in LLM-based data synthesis and model training can be leveraged (Zhang et al., 2024a,b). For policy selection, further enhancement can be achieved through a dedicated reasoning component. Inspired by DeepSeek's R1 framework (Guo et al., 2025), we can simulate various conversation scenarios using different disambiguation policies, calculate rewards based on performance metrics (e.g., answer accuracy, user satisfaction, IR performance), and learn optimal policy selection strategies through iterative refinement. This two-pronged approach—combining specialized ambiguity detection with intelligent policy selection—could lead to a more robust and effective “Disambiguation Agent."

6 Conclusion

In this paper, we have provided a comprehensive analysis of ambiguity and disambiguation in LLM-based CQA systems through three fundamental research questions. First, we have explored different types of ambiguity and proposed a unified taxonomy using three categories. We also highlighted the challenges of accurately detecting ambiguity, even with LLMs. Next, we have categorized various LLM-based disambiguation approaches and reviewed key benchmark datasets and metrics. Finally, we discussed open challenges and opportunities for LLM-based ambiguity detection and disambiguation strategies. By offering a comprehensive review of current research on ambiguities and disambiguation with LLMs, we hope our survey will contribute to the development of more robust and reliable LLM-based applications.

Limitations

In this work, we aimed to provide a comprehensive review and categorization of recent research on LLM-based ambiguity detection and disambiguation. Through our analysis, we identified three simplified categories of ambiguity types and three primary disambiguation techniques. However, this categorization is not exhaustive and may differ from other frameworks, which often use more granular or task-specific classifications. Despite our thorough literature review, it is possible that some recent or less-publicized works were overlooked, given the rapid advancements in this field. Additionally, our survey focused exclusively on ambiguity in Conversational Question Answering (CQA) tasks. In this survey, we did not cover other important NLP tasks, such as Natural Language Inference (NLI), Machine Translation (MT), Information Retrieval (IR), and Code Generation (e.g., NL2SQL), where ambiguities also arise and pose significant challenges. Future work could benefit from extending the scope to include these tasks, providing a more holistic understanding of ambiguity in NLP applications.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *arXiv preprint arXiv:2009.11352*.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2022. Query refinement prompts for closed-book long-form question answering. *arXiv preprint arXiv:2210.17525*.

Abhijit Anand, Vinay Setty, Avishek Anand, et al. 2023. Context aware query rewriting for text rankers using llm. *arXiv preprint arXiv:2308.16753*.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.

Barend Beekhuizen, Blair C Armstrong, and Suzanne Stevenson. 2021. Probing lexical ambiguity: Word vectors encode number and relatedness of senses. *Cognitive Science*, 45(5):e12943.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)*, 44(1):1–50.

Kenneth Church and Ramesh Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *American Journal of Computational Linguistics*, 8(3-4):139–149.

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*.

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, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. Pacific: towards proactive conversational question answering over tabular and textual data in finance. *arXiv preprint arXiv:2210.08817*.

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.

Kaustubh D Dhole. 2020. Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559*.

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

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. *Can You Unpack That? Learning to Rewrite Questions-in-Context*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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*.

Yan Huang. 2017. *Pragmatics*. Oxford University Press.

Yeonjun In, Sungchul Kim, Ryan A Rossi, Md Mehrab Tanjim, Tong Yu, Ritwik Sinha, and Chanyoung Park. 2024. Diversify-verify-adapt: Efficient and robust retrieval-augmented ambiguous question answering. *arXiv preprint arXiv:2409.02361*.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pages 387–396.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009, Singapore. Association for Computational Linguistics.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sanggoo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. *arXiv preprint arXiv:2404.11972*.

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

Vaibhav Kumar and Alan W. black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. *Preprint*, arXiv:2006.05986.

Litton J Kurisinkel and Nancy F Chen. 2023. Llm based multi-document summarization exploiting main-event biased monotone submodular content extraction. *arXiv preprint arXiv:2310.03414*.

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.

Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *ACM SIGIR Forum*, volume 51, pages 260–267. ACM New York, NY, USA.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11526–11544.

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

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We‘re afraid language models aren‘t modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Rafe: Ranking feedback improves query rewriting for rag. *arXiv preprint arXiv:2405.14431*.

Behrang Mehrparvar and Sandro Pezzelle. 2024. Detecting and translating language ambiguity with multilingual llms. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 310–323.

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

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. 2024. Large language model based long-tail query rewriting in taobao search. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 20–28.

Hongjin Qian and Zhicheng Dou. 2022. Explicit query rewriting for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4725–4737.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746.

Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. *arXiv preprint arXiv:1904.02281*.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143.

Sagi Shaier, Ari Kobren, and Philip Ogren. 2024. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. *arXiv preprint arXiv:2410.04241*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Harvard University Press.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Md Mehrab Tanjim, Xiang Chen, Victor S Bursztyn, Uttaran Bhattacharya, Tung Mai, Vaishnavi Muppala, Akash Maharaj, Saayan Mitra, Eunyee Koh, Yunyao Li, et al. 2025. Detecting ambiguities to guide query rewrite for robust conversations in enterprise ai assistants. *arXiv preprint arXiv:2502.00537*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

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.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Thomas Wasow. 2015. Ambiguity avoidance is overrated[1]. *Ambiguity: Language and.*

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278.

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, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. *arXiv preprint arXiv:2310.09716*.

Pavlos Zakkas, Suzan Verberne, and Jakub Zavrel. 2024. Sumblogger: Abstractive summarization of large collections of scientific articles. In *European Conference on Information Retrieval*, pages 371–386. Springer.

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.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint arXiv:2310.12823*.

Michael JQ Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.

Michael JQ Zhang, W Bradley Knox, and Eunsol Choi. 2024a. Modeling future conversation turns to teach llms to ask clarifying questions. *arXiv preprint arXiv:2410.13788*.

Tianhua Zhang, Kun Li, Hongyin Luo, Xixin Wu, James Glass, and Helen Meng. 2024b. Adaptive query rewriting: Aligning rewriters through marginal probability of conversational answers. *arXiv preprint arXiv:2406.10991*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024c. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *COLING 2002: The 19th International Conference on Computational Linguistics*.