ClarifyMT-Bench: Benchmarking and Improving Multi-Turn Clarification for Conversational Large Language Models

Sichun Luo[1], Yi Huang[2], Mukai Li[1], Shichang Meng[3], Fengyuan Liu[1], Zefa Hu[2], Junlan Feng[2], Qi Liu[1]

[1]The University of Hong Kong   [2]JIUTIAN Research, China Mobile   [3]CityUHK
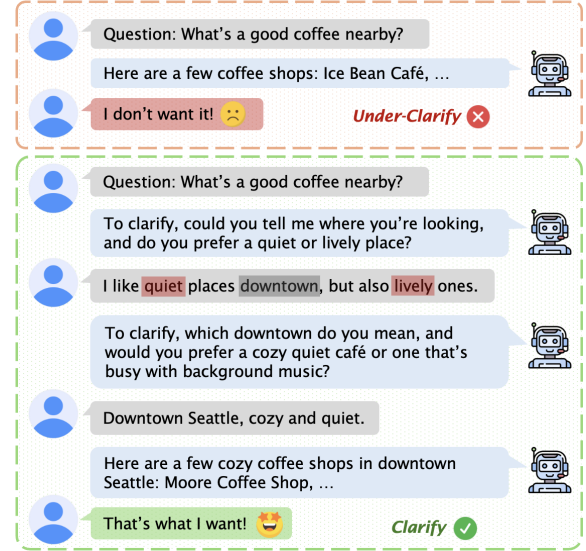
sichunluo2@gmail.com

Abstract

Large language models (LLMs) are increasingly deployed as conversational assistants in open-domain, multi-turn settings, where users often provide incomplete or ambiguous information. However, existing LLM-focused clarification benchmarks primarily assume single-turn interactions or cooperative users, limiting their ability to evaluate clarification behavior in realistic settings. We introduce **ClarifyMT-Bench**, a benchmark for multi-turn clarification grounded in a five-dimensional ambiguity taxonomy and a set of six behaviorally diverse simulated user personas. Through a hybrid LLM–human pipeline, we construct 6,120 multi-turn dialogues capturing diverse ambiguity sources and interaction patterns. Evaluating ten representative LLMs uncovers a consistent under-clarification bias: LLMs tend to answer prematurely, and performance degrades as dialogue depth increases. To mitigate this, we propose **ClarifyAgent**, an agentic approach that decomposes clarification into perception, forecasting, tracking, and planning, substantially improving robustness across ambiguity conditions. ClarifyMT-Bench establishes a reproducible foundation for studying when LLMs should ask, when they should answer, and how to navigate ambiguity in real-world human–LLM interactions.

1 Introduction

Large language models (LLMs) have rapidly become the foundation of modern conversational systems [22, 28, 42]. Yet in real-world interactions, user inputs are rarely complete or unambiguous: users often omit key details, provide vague descriptions, contradict earlier statements, or respond off-topic [24, 34]. Since modern LLMs are primarily optimized for helpfulness and fluency [31, 39], they tend to generate confident answers even when the user intent is underspecified. Such premature answering behavior can lead to misleading or unsafe outputs, disproportionately affecting users with limited domain knowledge or digital literacy [12, 27].

Clarification offers a principled alternative: rather than inferring user intent from underspecified queries, a system should proactively ask targeted follow-up questions until the necessary information is obtained [8, 36]. However, clarification in open-domain, multi-turn dialogue is inherently challenging, as user responses may be vague, contradictory, or off-focus, as illustrated in Figure 1. Recent work has begun to explore clarification-oriented dialogue systems. CLAMBER [46] evaluates ambiguity detection and single-turn clarifying question generation; ClariMM [32] incorporates multimodal signals for disambiguation; and ClarQ-LLM [10] studies clarification in task-oriented domains. Despite these advances, existing efforts remain limited in scope and do not fully address clarification under noisy, unconstrained, and multi-turn interaction patterns.

Figure 1: Illustration of clarification in user–LLM interactions. The upper example shows an under-clarified response that fails to capture user intent, while the lower example demonstrates effective clarification through follow-up questions, leading to user satisfaction. The user response may contain contradictory or vague information, labeled in red and gray respectively.

Existing benchmarks exhibit several key limitations: they typically assume single-turn interactions, cooperative user behavior, or narrowly scoped tasks, limiting their ability to capture the noisy and non-deterministic nature of real-world human-LLM interactions. Moreover, existing evaluations rarely assess when an LLM should ask, what it should ask, when it should stop asking, or how it should remain robust under vague, contradictory, or factually incorrect user feedback, which are the core challenges real-world conversational settings. This motivates a central question: **Can LLMs effectively balance clarification and answering under complex, noisy multi-turn interactions?**

To address this question, we introduce **ClarifyMT-Bench**, a multi-turn clarification benchmark designed to evaluate an LLM's ability to dynamically choose between asking and answering. Our benchmark is grounded in a five-dimensional ambiguity taxonomy that spans linguistic, intent, contextual, epistemic, and interactional ambiguity, providing a principled foundation for controlled multi-turn evaluation. We construct the dataset using a hybrid LLM–human pipeline that generates diverse ambiguous queries,

Table 1: Comparison with related datasets and benchmarks.

Dataset / Benchmark	Clarify	Multi-Turn	Open-Domain	LLM-focused Eval	Noisy User
ClariQ [1]	✓	✗	✓	✗	✗
CLAMBER [46]	✓	✗	✓	✓	✗
ClarQ-LLM [10]	✓	✓	✗	✓	✗
ClariMM [32]	✓	✓	✗	✓	✗
Our ClarifyMT-Bench	✓	✓	✓	✓	✓

filters and refines them for correctness and clarity, and expands them into multi-turn dialogues. To model realistic conversational uncertainty, we further incorporate a user-behavior simulator encompassing six response types via LLM prompting and validated by human annotators for quality and diversity. A comparison of ClarifyMT-Bench with prior benchmarks is provided in Table 1.

We evaluate ten leading LLMs spanning six model families. Across all settings, we observe a consistent under-clarification bias: most LLMs prefer to answer prematurely rather than engage in sufficient clarification, especially as dialogue depth increases. This finding highlights an important gap between current alignment objectives and robust, responsible conversational behavior.

To address this challenge, we propose **ClarifyAgent**, an agent-based framework that formulates multi-turn clarification as a structured reasoning process. Beyond the ask–answer decision task, we introduce a new user persona inference task. Specifically, upon receiving a user query, the Perceiver extracts task-relevant information and identifies potential ambiguity, while the Forecaster infers the user persona to anticipate the user's behavioral tendencies. The Tracker then updates the state representing unresolved ambiguity slots. Finally, the Planner integrates signals from all components and decides whether to continue clarifying or proceed to answering, with the selected action executed by the Output module. Experimental results validate the effectiveness of ClarifyAgent, demonstrating substantial improvements in clarification performance.

In a nutshell, our contributions are threefold.

- **Benchmark.** We present ClarifyMT-Bench, to our best knowledge, the first benchmark that jointly evaluates multi-turn, open-domain clarification under noisy user behavior. Grounded in a five-dimensional ambiguity taxonomy and six behaviorally diverse user personas, ClarifyMT-Bench contains 6,120 multi-turn interactions, enabling controlled evaluation under diverse conversational uncertainty.
- **Analysis.** Through extensive experiments on ten representative LLMs, we identify a consistent under-clarification bias and quantify robustness disparities across ambiguity types, user persons, and dialogue depth. These findings highlight key limitations in current conversational LLMs' ability to balance asking and answering under uncertainty.
- **Method.** We introduce ClarifyAgent, an agentic framework that decomposes clarification into perception, forecasting, tracking, and planning. By introducing extra user persona inference task, ClarifyAgent produces more robust multi-turn clarification behavior, serving as a strong baseline for future research.

2 Related Work

Multi-turn Dialogue. Multi-turn dialogue better reflects real-world conversational settings and plays a central role in user experience [3, 47]. However, recent studies show that LLMs still struggle with state tracking, long-context reasoning, and grounding under evolving dialogue histories [15, 20, 44]. When instructions involve pronouns, ellipses, or cross-turn dependencies, models frequently fail to maintain consistency or follow user intent [4]. Such limitations directly affect clarification, where models must track ambiguous references, integrate new information, and determine when clarification is necessary as the dialogue progresses.

LLM-oriented Clarification Dataset and Benchmark. Traditional clarification methods in IR and NLP treat ambiguity as a static property of the query, resolving lexical, syntactic, or topical underspecification through query expansion, disambiguation, or slot filling [13, 26]. In contrast, LLM-based assistants must reason under open-ended generation, hallucination risk, epistemic uncertainty, and diverse user preferences, making the ask–or–answer decision itself central. Several recent benchmarks study this direction: CLAMBER [46] evaluates clarifying question quality; ClariMM [32] extends the task to multimodal inputs; and ClarQ-LLM [10] examines multi-turn clarification in task-oriented settings. While valuable, these resources typically do not jointly evaluate stop decisions, operate in limited domains, or account for noisy and inconsistent users, leaving open how to assess clarification policies under realistic ambiguity and imperfect user feedback.

Clarification Methods for LLMs. Prompting-based approaches have been explored to elicit better clarification behavior. Deng et al. [8] and Lee et al. [17] investigate prompting strategies for asking clarifying questions. Moreover, Zhang et al. [45] introduce a task-agnostic framework for detecting ambiguity, selecting effective clarifying questions, and integrating new information. Additionally, AT-CoT [36] improves clarification by first identifying ambiguity type and then generating targeted questions. Although effective in single-turn or cooperative settings, these methods do not directly generalize to multi-turn interactions with noisy, inconsistent, or adversarial users, where clarify-or-answer decisions must be made sequentially under uncertainty.

User Simulation with LLMs. User simulators have long been used to evaluate interactive systems, from task-oriented dialogue [35] to interactive IR [25]. However, many classical simulators assume cooperative, truthful, and stationary users, limiting their ability to reflect the variability and noise present in real interactions [40]. Recent LLM-based simulators introduce richer behaviors and improved linguistic naturalness, yet they frequently emphasize coherent and goal-directed responses, which may overlook contradictory, off-focus, or factually incorrect replies [23, 33]. Capturing such behaviors is essential for robust clarification evaluation, where systems must determine when to ask, what to ask, and when to stop under imperfect and unpredictable user feedback.

Limitations of Existing Work. Existing LLM-oriented clarification benchmarks have advanced the study of ambiguity resolution, but they often restrict domain coverage or interaction types, limiting their applicability to open-domain conversational settings. Moreover, current ambiguity taxonomies provide limited expressiveness for modeling the multi-faceted sources of ambiguity that arise in

Table 2: Ambiguity Taxonomy in the era of LLMs, illustrating five major dimensions and representative subtypes with examples. Categories are organized along a continuum from linguistic form to social interaction.

Category	Subtype	Description	Example
Linguistic Ambiguity	*Lexical Ambiguity*	Word has multiple meanings.	Please tell me about the seal.
	Syntactic Ambiguity	Sentence allows multiple parses.	List movies from the 1990s starring actors from Canada.
	Semantic Ambiguity	Unclear semantic role or criteria.	Is New York the largest city?
Intent Ambiguity	*Goal Ambiguity*	The user's goal is vague or incomplete.	Help me write a report.
	Scope Ambiguity	The intended task scope is unclear.	Tell me about quantum computing.
	Intent Conflict Ambiguity	The user expresses incompatible goals.	Summarize 'War and Peace' without omitting anything.
Contextual Ambiguity	*Entity Ambiguity*	Multiple possible referents exist for a term or name.	Who is the real Spider-Man?
	Spatial Ambiguity	The location is unspecified or underspecified.	Tell me how to reach London.
	Temporal Ambiguity	The time is unspecified or underspecified.	I need the weather forecast for New York.
Epistemic Ambiguity	*Knowledge Gap Ambiguity*	The user assumes shared prior knowledge or context.	You remember the new update, right?
	Unfamiliarity Ambiguity	The query involves entities or facts unknown to the model.	Find the price of the Samsung Chromecast.
	Value Ambiguity	Subjective or evaluative terms without clear criteria.	Recommend a good movie.
Interactional Ambiguity	*Partial / Vague reply*	The user provides uncertain or imprecise feedback.	Sort of, I guess.
	Factually Wrong Reply	The user provides information that is clearly incorrect.	Paris is the capital of Germany.
	Contradictory Reply	The user expresses conflicting statements or attitudes.	It's urgent. No rush actually.
	Off-focus Reply	The user diverts the intended topic or clarification.	Let's talk about something else.

LLM-era interactions. Most prior work further assumes cooperative or noise-free user feedback, under-emphasizing robustness to vague, off-topic, contradictory, or factually incorrect responses. In addition, existing benchmarks rarely evaluate state-of-the-art LLMs under such challenging conditions. These limitations highlight the need for a more comprehensive benchmark capable of evaluating clarification policies under realistic, multi-turn, and noise-prone human–LLM interactions.

3 Taxonomy and Task

3.1 Ambiguity Taxonomy in the Era of LLMs

Ambiguity is an inherent property of human communication, and prior work typically categorizes it into syntactic, semantic, and contextual varieties [46]. However, LLM-based assistants encounter a broader range of ambiguity sources that extend beyond traditional linguistic formulations. We therefore propose a five-dimensional taxonomy spanning linguistic, intent, contextual, epistemic, and interactional ambiguity. This taxonomy reflects a progression from surface-level linguistic form to interaction-level behavior, consistent with cognitive models of communication that link symbols, intentions, shared context, and collaborative action [5, 18]. Our five-dimensional taxonomy is defined as follows:

- *Linguistic Ambiguity.* Ambiguity arising from lexical, structural, or semantic indeterminacy in the user's utterance.
- *Intent Ambiguity.* Uncertainty about what the user wants the system to accomplish, including unclear goal, task scope, or potentially conflicting intent.
- *Contextual Ambiguity.* Ambiguity caused by under-specified or shifting references to entities, locations, time, or discourse context.
- *Epistemic Ambiguity.* Ambiguity about the knowledge shared between the user and the model, such as assumptions about background knowledge, unfamiliar entities, or subjective evaluative criteria.

- *Interactional Ambiguity.* Ambiguity introduced by the interaction itself when user replies are vague, off-focus, contradictory, or factually incorrect, making it unclear how the conversation should proceed.

This five-dimensional taxonomy highlights the multifaceted nature of ambiguity in LLM-mediated communication and provides a conceptual foundation for evaluating models' ability to detect, diagnose, and adapt to diverse uncertainty sources. Representative subtypes and examples are provided in Table 2.

3.2 Task Formulation

We view multi-turn clarification as a sequential decision-making problem. Given an under-specified user query q_0, the goal of the dialogue system is to acquire the minimal additional information necessary to produce a well-specified and complete answer.

Conceptually, we assume a set of ambiguity-relevant slots $S = \{s_1, \ldots, s_n\}$ (*e.g.*, destination, budget, time), and a subset $S^* \subseteq S$ denoting the *required slots* that must be resolved for the query. At turn t, the dialogue state can be represented as

$$x_t = [f_t(s_1), \ldots, f_t(s_n)], \tag{1}$$

where each slot is in one of three abstract states:

$$f_t(s_i) \in \{\texttt{unfilled}, \texttt{filled}, \texttt{conflict}\}. \tag{2}$$

An ideal clarification policy would choose an action $a_t \in \{\texttt{Clarify}, \texttt{Answer}\}$ based on x_t, and would stop asking once all required slots are filled and no conflicts remain:

$$\text{Stop if} \quad \left(\forall s \in S^*, f_t(s) = \texttt{filled}\right) \land \left(\nexists s' \in S, f_t(s') = \texttt{conflict}\right). \tag{3}$$

In practice, we do not explicitly annotate slots or S^* for each instance, due to the high cost and ambiguity of manually identifying required information for open-domain queries. Instead, each dialogue turn in ClarifyMT-Bench is associated with a *reference action* $y_t \in \{\texttt{Clarify}, \texttt{Answer}\}$ derived from our construction pipeline.
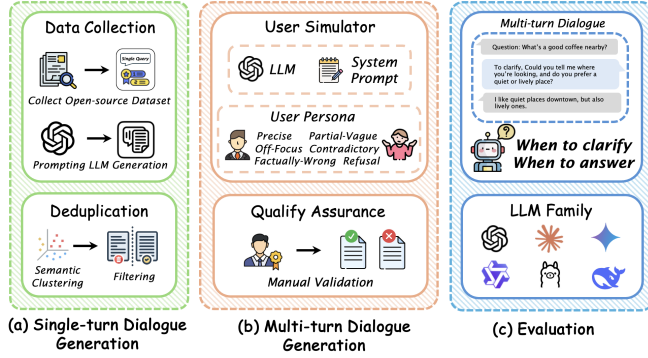
Figure 2: The pipeline of dataset construction.

Given the dialogue prefix up to turn t, the model predicts an action $\hat{a}_t$, and we evaluate its decisions as follows:

- *Under-Clarify*: $\hat{a}_t =$ Answer while $y_t =$ Clarify, meaning the model answers prematurely.
- *Over-Clarify*: $\hat{a}_t =$ Clarify while $y_t =$ Answer, meaning the model asks when an answer is already warranted.

We primarily report overall decision accuracy, *i.e.*, $\mathbb{I}[\hat{a}_t = y_t]$ averaged over all turns and instances, and use the under-/over-clarification categories in our qualitative analysis.

4 Dataset Construction

As illustrated in Figure 2, the construction of ClarifyMT-Bench consists of two stages: (1) *single-turn dialogue generation*, where we collect and synthesize ambiguous queries paired with clarifying questions; and (2) *multi-turn dialogue expansion*, where we simulate user responses of different behavior types to extend each instance into a multi-turn conversation. This pipeline yields a benchmark that reflects the complexity and noise patterns observed in real-world clarification-oriented dialogue.

4.1 Single-Turn Dialogue Generation

Data Collection. We first construct an initial pool of ambiguous user queries and their corresponding clarifying questions. To ensure broad coverage, we prompt diverse LLMs (*e.g.*, GPT-4.1, DeepSeek-V3) to generate ambiguous queries along with clarifying questions aligned with the taxonomy in Section 3.1. Prompt templates and instructions are provided in Figure 7.

Formally, let $\mathcal{M}$ denote an LLM and p a prompt template containing task instructions and few-shot exemplars. The model outputs an ambiguous query q_{amb} paired with a clarifying question q_{clar}:

$$(q_{\text{amb}}, q_{\text{clar}}) = \mathcal{M}(p; \theta), \tag{4}$$

where θ represents model parameters. We repeat this process across multiple models and prompt variants to increase lexical, stylistic, and semantic diversity.

In addition, we leverage samples from existing clarification datasets, such as CLAMBER [46], to improve diversity.

Deduplication. To improve quality and prevent redundancy, we perform semantic-level deduplication. To be specific, each query q_i

is embedded using a pretrained encoder $\mathcal{E}$:

$$\mathbf{v}_i = \mathcal{E}(q_i), \tag{5}$$

and cosine similarity is computed for each pair:

$$S_{ij} = \frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}. \tag{6}$$

Queries with similarity $S_{ij} > \tau_{\text{sem}}$ are grouped into cluster C_k. From each cluster, we retain a single representative:

$$q_k^* = \arg\max_{q_i \in C_k} \text{Quality}(q_i), \tag{7}$$

where $\text{Quality}(q_i)$ reflects fluency and informativeness assessed through human annotation. We use `all-MiniLM-L6-v2`[1] as the encoder and set $\tau_{\text{sem}} = 0.7$ based on a small pilot study to balance recall and diversity. This procedure removes semantic duplicates and enhances dataset variety.

4.2 Multi-Turn Dialogue Generation

User Simulator. To emulate realistic human variability, we model user behavior along five dimensions: information coverage, truthfulness, self-consistency, cooperativeness, and specificity. These dimensions reflect key factors that influence clarification difficulty and commonly vary in real-world interactions. Based on them, we define six canonical user personas:

- *Precise*: provides specific, accurate, and fully relevant information that directly fills the target slot.
- *Partial–Vague*: offers partially relevant but underspecified or indecisive responses with low specificity.
- *Off–Focus*: replies with off-topic or tangential content that does not address the requested slot.
- *Contradictory*: gives internally or cross-turn inconsistent responses that hinder stable intent inference.
- *Factually–Wrong*: provides incorrect factual information.
- *Refusal*: declines to clarify, avoids answering, or explicitly urges the model to proceed without clarification.

This structured simulator supports controlled evaluation across cooperative, noisy, and adversarial user behaviors. Precise and Refusal users typically provide enough information or explicitly decline to provide more, permitting a direct answer, whereas the other four personas generally require additional clarification. Representative examples of user replies are shown in Table 7.

Manual Validation. Following automatic generation, we conduct a human quality check. Two annotators with prior NLP data-curation experience independently review each ambiguous query – clarification pair. Instances that are incoherent, ungrammatical, mislabeled, or inconsistent with the intended ambiguity category are removed. Disagreements are resolved through discussion until consensus is reached. This verification step ensures that all retained samples are linguistically sound and faithfully aligned with their target ambiguity types.

Human Annotation Reliability. To evaluate the reliability of manual annotations, two annotators independently performed a binary filtering decision (retain or remove) on 10% of the randomly sampled instances. We compute Cohen's Kappa (κ) [6], which adjusts for chance agreement. The observed agreement is $P_o = 0.8627$, and

---

[1]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

the resulting Kappa is $\kappa = 0.5980$, indicating moderate to substantial agreement according to the Landis & Koch scale [16]. These results demonstrate that the annotation guideline is clear and that the filtering quality is consistent across annotators.

4.3 Data Statistics

Our dataset covers 12 fine-grained ambiguity subtypes, grouped into four broad categories: linguistic, intent, contextual, and epistemic ambiguity, which characterize the source of ambiguity in the initial query. We generate 85 single-turn instances for each subtype, resulting in 1,020 ambiguous query–clarification pairs. To capture interaction-level variability in user responses, each single-turn instance is further expanded into six multi-turn dialogues, corresponding to six simulated user personas. For the Precise and Refusal user personas, we construct two-turn conversations, while we generate three-turn conversations for the remaining four personas. We limit each simulated dialogue to two or three turns to keep the evaluation focused and controllable. Longer dialogues are certainly possible in real-world settings, but extending them synthetically risks accumulating distributional artifacts (*e.g.*, unnatural conversational drift) and compounding model errors from the user simulator. This design produces a total of **6,120** multi-turn dialogues, with an average of **2.67** turns per dialogue. As shown in Figure 9, the average token length per dialogue is **66.1** tokens. The data source distribution is provided in Appendix B.

5 Evaluation

5.1 Evaluation Setup

Tasks and Metrics. We evaluate models on two complementary tasks that together assess both clarification capability and dialogue quality. (1) Multi-turn ambiguous query clarification. Given an underspecified user query, a model must decide whether to ask a clarifying question or directly provide an answer. We measure clarification accuracy, defined as the proportion of correctly selected actions over all dialogue turns. We further report performance by user persona and ambiguity subtype to capture fine-grained behavioral differences. All experiments are conducted on the full dataset, and we report the average accuracy. (2) Clarifying question quality evaluation. For each clarification turn, we evaluate the semantic quality of the generated question. We adopt an LLM-as-a-Judge evaluation protocol [19] and manual evaluation to assess contextual appropriateness, relevance, and helpfulness.

Evaluated Models. We conduct extensive experiments using representative models from six major LLM families, covering both closed-source and open-source models, including: GPT-4.1 [30], o3 [29], Gemini-2.5-Flash [7], Claude-Sonnet-4.5 [2], Qwen-2.5 [37], Llama-3.1 [9], DeepSeek-V3 [21], and DeepSeek-R1 [11]. These LLMs span a range of training paradigms, allowing for a comprehensive comparison of robustness and interactional behavior under ambiguity. The implementation details are shown in Appendix C.

5.2 Main Results

5.2.1 Task 1: Multi-Turn Ambiguous Query Clarification. We evaluate model performance across user personas, with results shown in Table 3. Qwen-2.5-7B-It tends to answer directly regardless of uncertainty, which boosts its scores on Precise and Refusal cases but

Table 3: Model performance across six user personas: Precise (P), Partial-Vague (PV), Off-Focus (OF), Contradictory (CT), Factually-Wrong (FA), and Refusal (RF). The best and worst results are labeled in green and red, respectively.

Model	P	PV	OF	CT	FA	RF	Avg.
GPT-4.1	69.2	81.9	89.1	82.0	69.9	40.7	72.1
o3	55.9	73.3	76.8	68.1	54.7	56.5	64.2
Gemini-2.5-Flash	48.7	95.5	94.5	97.4	83.7	14.5	72.4
Claude-Sonnet-4.5	36.8	94.8	96.1	91.9	93.7	35.6	74.8
Qwen-2.5-7B-It	96.6	36.5	60.3	42.4	38.8	72.9	57.9
Qwen-2.5-72B-It	93.8	67.7	78.9	72.5	75.4	59.8	74.7
Llama-3.1-8B-It	48.3	75.6	87.5	86.4	84.8	44.4	71.2
Llama-3.1-70B-It	70.4	85.1	90.9	88.7	88.7	39.0	77.1
DeepSeek-V3	70.0	85.0	82.1	85.6	78.6	31.5	72.1
DeepSeek-R1	68.0	77.4	78.8	73.7	64.4	49.1	68.6

causes sharp drops on ambiguity-heavy types where clarification is required. Across models, most achieve an overall accuracy above 70%, yet each displays distinct strengths and weaknesses across personas. Notably, performance on Refusal-style inputs is uniformly low, reflecting a widespread tendency toward over-clarification and underscoring the need for better ambiguity-aware alignment. Moreover, large reasoning-centric models do not consistently outperform instruction-tuned LLMs, suggesting that explicit reasoning traces alone are insufficient for handling ambiguity-driven behaviors. Instead, robustness appears more closely tied to alignment quality and pragmatic inference than to raw reasoning depth. Finally, model scale correlates strongly with robustness: larger models consistently outperform their smaller counterparts, indicating that pragmatic reasoning, uncertainty calibration, and referential grounding all benefit substantially from increased capacity.

Table 4 reports subtype-level results. Across ambiguity subtypes, we observe three consistent trends. First, performance declines markedly as the dialogue progresses: while most models perform reasonably well on the first turn, accuracy drops sharply on the second turn and even more steeply on the third, showing that multi-turn clarification compounds reasoning difficulty. Second, differences between models become more pronounced with increased depth. Gemini-2.5-Flash is the most robust model, achieving the highest accuracy in most subtypes, whereas smaller models (*e.g.*, Qwen-2.5-7B-It, Llama-3.1-8B-It) degrade severely. Third, despite strong first-turn accuracy, variability across subtypes remains large. These findings highlight that ambiguity resolution is fragile under multi-turn interaction and that both model capacity and ambiguity subtype heavily influence robustness.

Takeaway 1. Most LLMs exhibit a strong tendency to under-clarify as dialogue depth increases, particularly for smaller models. In addition, robustness varies across user personas, indicating that user behavior is a major factor shaping clarification performance.

5.2.2 Task 2: Clarifying Question Quality Evaluation. We evaluate clarifying question quality using an LLM-as-a-Judge evaluation protocol. We randomly sample 30 instances from each ambiguity subtype and use GPT-4.1 to assign quality scores. For each instance, the judge assigns a score between 0 and 5, where 5 indicates a question that is fully appropriate for resolving the underlying ambiguity,

Table 4: Model performance by ambiguity subtype across dialogue turns. Subtypes include: Lexical (Lex.), Syntactic (Syn.), Semantic (Sem.), Goal (Goal), Scope (Sco.), Intent Conflict (Con.), Entity (Ent.), Spatial (Spa.), Temporal (Tmp.), Knowledge Gap (Kno.), Unfamiliarity (Unf.), and Value (Val.) ambiguity. Avg. denotes the average accuracy across all subtypes. The best and worst results are labeled in green and red, respectively.

Model	Lex.	Syn.	Sem.	Goal	Sco.	Con.	Ent.	Spa.	Tmp.	Kno.	Unf.	Val.	Avg.
Turn = 1													
GPT-4.1	78.8	67.1	84.7	100.0	63.5	87.1	62.4	61.2	100.0	100.0	70.6	91.8	77.7
o3	90.6	57.6	49.4	97.6	28.2	77.6	74.1	64.7	77.6	100.0	77.6	67.1	71.8
Gemini-2.5-Flash	88.2	100.0	94.1	100.0	85.9	90.6	77.6	69.4	83.5	97.6	87.1	98.8	89.4
Claude-Sonnet-4.5	82.4	74.1	71.8	97.6	27.1	76.5	64.7	62.4	77.6	94.1	94.1	78.8	75.1
Qwen-2.5-7B-It	89.4	94.1	95.3	98.8	67.1	90.6	77.6	76.5	84.7	98.8	83.5	92.9	87.4
Qwen-2.5-72B-It	74.1	70.6	90.6	100.0	57.6	83.5	64.7	61.2	78.8	100.0	83.5	82.4	78.9
Llama-3.1-8B-It	28.2	43.5	41.2	72.9	37.6	69.4	3.5	60.0	72.9	69.4	67.1	75.3	53.4
Llama-3.1-70B-It	77.6	76.5	97.6	97.6	85.9	77.6	76.5	62.4	80.0	96.5	72.9	96.5	83.1
DeepSeek-V3	54.1	61.2	60.0	92.9	16.5	76.5	25.9	61.2	76.5	98.8	94.1	58.8	64.7
DeepSeek-R1	91.8	81.2	88.2	100.0	70.6	87.1	72.9	65.9	81.2	98.8	89.4	88.2	84.6
Turn = 2													
GPT-4.1	56.5	51.6	57.6	72.4	34.9	64.9	44.1	47.5	57.8	68.0	56.3	71.6	56.9
o3	49.0	35.5	28.0	69.0	14.7	54.9	42.5	43.3	55.1	67.3	51.8	42.5	46.1
Gemini-2.5-Flash	73.7	68.2	67.3	67.6	64.5	63.7	63.7	49.8	59.8	59.2	62.0	73.9	64.4
Claude-Sonnet-4.5	66.1	53.7	53.1	72.5	23.3	54.5	58.8	42.9	52.5	63.3	70.0	63.3	56.3
Qwen-2.5-7B-It	50.0	58.4	52.7	63.3	40.4	50.8	43.5	41.8	46.5	53.9	53.3	53.1	50.6
Qwen-2.5-72B-It	52.4	60.8	70.8	77.6	36.7	61.8	44.3	45.9	61.0	70.2	66.1	59.8	59.0
Llama-3.1-8B-It	17.6	29.8	28.6	52.0	27.3	51.6	2.5	43.3	53.7	47.8	46.5	55.5	38.0
Llama-3.1-70B-It	60.6	57.8	76.1	73.9	64.7	56.3	66.3	49.8	59.4	67.5	57.3	80.0	64.2
DeepSeek-V3	42.2	47.3	41.4	66.7	11.0	53.9	20.0	43.9	52.0	66.9	71.6	43.5	46.7
DeepSeek-R1	59.4	62.0	62.7	69.2	45.9	63.5	43.9	44.7	56.9	66.9	59.8	61.4	58.0
Turn = 3													
GPT-4.1	38.5	49.1	47.4	89.1	19.4	74.1	27.4	51.5	68.5	94.4	49.7	61.2	55.9
o3	20.0	33.8	24.1	65.9	4.4	58.2	16.8	50.0	66.5	88.5	37.9	30.9	41.4
Gemini-2.5-Flash	76.2	98.2	85.3	91.5	66.8	83.5	64.7	63.5	73.8	81.8	82.6	89.1	79.7
Claude-Sonnet-4.5	54.1	68.5	64.7	92.6	21.5	71.8	46.8	59.7	75.9	92.9	76.8	72.1	66.5
Qwen-2.5-7B-It	13.8	23.5	15.6	34.4	10.0	21.5	12.6	18.8	22.1	38.5	24.7	14.7	20.9
Qwen-2.5-72B-It	22.4	36.8	46.2	55.6	12.1	46.5	16.8	29.7	42.6	63.5	53.5	22.6	37.4
Llama-3.1-8B-It	7.6	16.2	7.6	34.1	8.2	35.6	0.0	17.6	32.4	51.2	33.8	19.7	22.0
Llama-3.1-70B-It	52.1	57.1	58.8	84.7	37.6	60.3	45.3	55.9	69.7	88.5	62.1	67.6	61.7
DeepSeek-V3	40.6	56.8	45.3	73.8	7.9	60.3	17.9	46.5	62.6	83.8	83.5	37.9	51.4
DeepSeek-R1	41.5	59.1	53.5	82.6	30.6	66.5	22.6	46.2	60.3	81.8	55.3	59.4	55.0

and 0 indicates an irrelevant or unhelpful question. The prompt template is shown in Figure 6.

Figure 3 presents model-wise quality scores across ambiguity subtypes. Across ambiguity subtypes, stronger models consistently produce higher-quality clarifying questions, while smaller models (*e.g.*, Llama-3.1-8B-It) perform substantially worse. Performance is highest on concrete ambiguity types, such as under-specified goals, missing entities, or attribute-level gaps, where the missing information is explicit. In contrast, quality degrades on more abstract forms of ambiguity, such as high-level intent or latent preference

gaps, where the desired clarification requires implicit reasoning. Overall, these patterns suggest that current LLMs are much more reliable at resolving explicit missing information than at inferring latent user intent or addressing discourse-level under-specification.

We additionally conduct a human evaluation using the same 0–5 rating rubric. The results is shown in Figure 4. We observe a Pearson correlation of 0.658 between human and LLM scores, indicating moderately strong alignment. Human ratings tend to be more extreme, exhibiting larger variance across models, whereas LLM scores are more conservative, a pattern consistent with prior
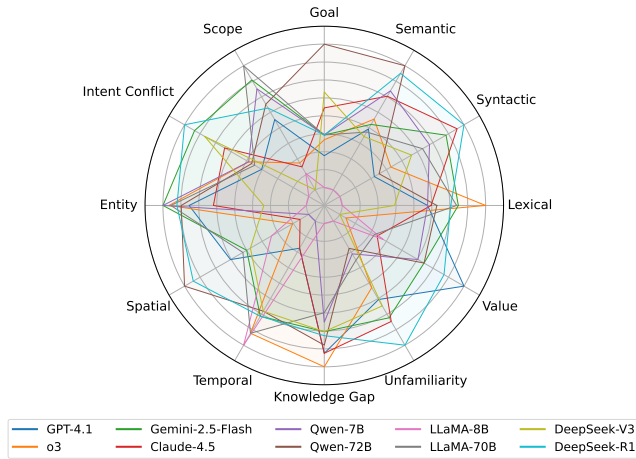
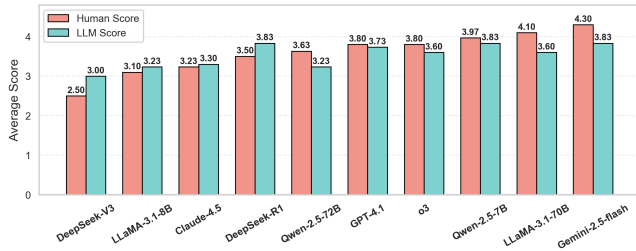Figure 3: Clarifying question quality for each ambiguity subtype evaluated by LLM-as-a-Judge.

Figure 4: Clarifying question quality evaluated by human and LLM-as-a-Judge.

reports that LLM judges avoid assigning very high or very low ratings [19]. Both human annotators and the LLM evaluator consistently assign lower scores to models such as DeepSeek-V3 and Llama-3.1-8B-It, while Gemini-2.5-Flash receives the highest average score from both sources, mirroring its strong performance in Table 4. This convergence further validates the reliability of the LLM-based evaluation.

Takeaway 2. Clarifying question quality largely correlates with ambiguous query clarification accuracy. Meanwhile, different models exhibit distinct strengths across ambiguity subtypes, with variation in both clarification accuracy and question quality.

6 ClarifyAgent: An Agentic Method for Multi-Turn Clarification

To improve multi-turn clarification in conversational LLMs, we propose ClarifyAgent, an agentic framework designed to better handle noisy and inconsistent user behavior. We introduce a new subtask, **user persona inference**, and integrate it into a ReAct-style [43] perception–action loop augmented with a finite-state slot tracker. ClarifyAgent enables dynamic decision-making about when to ask and when to answer across complex multi-turn interactions. The overall pipeline is illustrated in Figure 5.
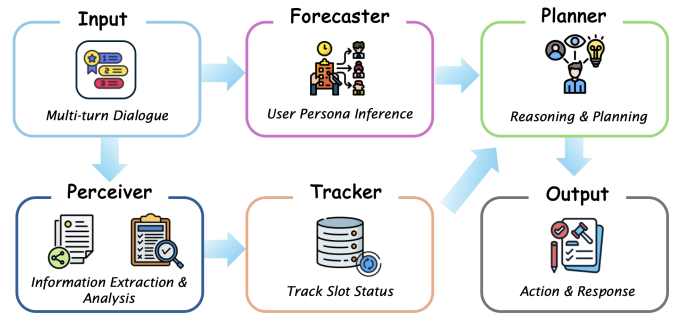

Figure 5: Pipeline of ClarifyAgent.

6.1 Framework

Architecture Overview. ClarifyAgent is composed of six modular components: an *Input* interface, a *Perceiver* for information extraction, a *Forecaster* for user persona inference, a *Tracker* for finite-state slot management, a *Planner* for reasoning and decision selection, and an *Output* module for action and response generation. Together, these components form a closed-loop perception–reasoning–action cycle that supports multi-turn clarification under noisy or inconsistent user behavior.

At a high level, ClarifyAgent executes a goal-directed interaction strategy that tightly couples tracking, forecasting, and planning. At each turn, it parses the user's latest utterance and updates the Tracker's slot states; the Forecaster then re-estimates the user persona and cooperativeness based on newly observed evidence. When important information is missing or conflicting, the Planner produces a focused clarification query. When the tracked slots are judged to be sufficiently resolved and the remaining ambiguity is low, the Planner instead synthesizes a task-completing answer.

Input. The Input module receives the multi-turn dialogue and normalizes it into a structured context representation. This representation serves as the entry point for downstream slot extraction, user inferring, and reasoning, ensuring that ClarifyAgent maintains a coherent view of the evolving conversation.

Perceiver. Given the current dialogue context, the Perceiver identifies candidate slot values and detects inconsistencies. For each slot, it assigns a discrete status label: *filled*, *unfilled*, or *conflict*—thereby converting raw natural-language input into structured perceptual signals. These signals summarize what is known, missing, or contradictory, and are passed to the Tracker for further reasoning.

Forecaster. The Forecaster estimates the user's behavioral persona from six categories: Precise, Partial–Vague, Off–Focus, Contradictory, Factually–Wrong, and Refusal. It outputs a structured persona label that captures interaction-level uncertainty and conditions the Planner's ask–answer decisions. This helps avoid unnecessary clarifications with cooperative users and prevents premature answering under noisy or inconsistent behavior.

Tracker. The Tracker maintains the evolving state of ambiguous slots using a finite-state machine (FSM) functioning as a state keeper. Each slot transitions among *unfilled*, *filled*, and *conflict* as the dialogue unfolds. When all required slots are filled and no conflicts remain, the Tracker notifies the Planner that the *Required Slot Completion (RSC)* condition is satisfied.

Table 5: Accuracy across six user personas: Precise (P), Partial–Vague (PV), Off–Focus (OF), Contradictory (CT), Factually–Wrong (FA), and Refusal (RF). The best results are highlighted in boldface, and * denotes statistical significance at $p < 0.05$ under a paired t-test.

Method	P	PV	OF	CT	FA	RF	Avg.
Backbone: Llama-3.1-8B-It							
Base Model	48.3	75.6	87.5	86.4	84.8	44.4	71.2
Majority Voting	48.3	76.9	86.4	89.0	91.2	43.3	72.4
CoT	12.7	80.2	89.0	88.6	90.1	27.3	64.7
Intent-Sim	22.6	83.3	84.4	86.6	77.6	15.0	61.6
AT-CoT	**74.5**	80.4	84.8	81.8	87.0	29.2	73.0
ClarifyAgent	58.0	**97.2***	**100.0***	**97.1***	**96.5***	**81.7***	**88.4***
Backbone: Qwen-2.5-7B-It							
Base Model	**96.6**	36.5	60.3	42.4	38.8	72.9	57.9
Majority Voting	96.4	42.7	74.6	49.1	43.2	71.6	62.9
CoT	89.0	55.0	63.1	65.6	62.7	63.1	66.4
Intent-Sim	65.2	59.4	60.0	60.8	61.4	50.1	59.5
AT-CoT	93.9	73.2	69.7	54.7	50.8	43.7	64.3
ClarifyAgent	85.3	**85.8***	**97.3***	**91.7***	**84.0***	**83.9***	**88.0***

Table 6: Ablation study of ClarifyAgent with Qwen-2.5-7B-It backbone. We report the accuracy across six user personas, as well as the average accuracy (Avg.) and the standard deviation (Std.) for each variant. The best results are highlighted in boldface, and the runner-up results are underlined.

Variant	P	PV	OF	CT	FA	RF	Avg.	Std.
ClarifyAgent	85.3	85.8	97.3	91.7	84.0	83.9	88.0	4.92
w/o Planner	75.2	79.9	94.1	81.8	81.3	85.7	83.0	5.86
w/o Perceiver	62.0	**91.6**	86.2	89.1	85.2	78.7	82.1	9.85
w/o Forecaster	**91.8**	67.2	90.9	72.1	75.2	**91.9**	81.5	10.28

Planner. The Planner drives ClarifyAgent's decision process via a perception–reasoning–action loop. At each turn, it integrates the FSM state, the inferred user persona, and the dialogue context to select between clarification and answering. When clarification is needed, the Planner selects the target information to query and instructs the Output module to generate a targeted clarifying question. When the RSC condition is satisfied or further clarification is unlikely to yield additional information, the Planner switches to Answer mode and delegates final response generation to the Output module.

Output. The Output module generates the final action, either a clarifying question or an answer. When relevant, it also conveys any remaining uncertainty or conditional assumptions (*e.g.*, "If you mean X, …"). By making such uncertainty explicit, the module provides users with clearer expectations about the reliability and scope of the model's response.

6.2 Empirical Performance Evaluation

Evaluation Setup. We evaluate ClarifyAgent on ClarifyMT-Bench using two representative open-source base LLMs: Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct. We compare against several training-free baselines, including Majority Voting [38], Chain-of-Thought (CoT) [41], Intent-Sim [45], and AT-CoT [36]. All methods operate on the same underlying backbone and are evaluated on

the second-turn dialogue for broad coverage of six user personas. Specifically, Majority Voting uses a decoding temperature of 0.7 and aggregates predictions over k=5 independent samples. CoT follows a standard zero-shot chain-of-thought prompting setup. Intent-Sim also relies on k=5 samples, with a similarity threshold $\tau = 0.5$ to determine whether to query the user. Moreover, AT-CoT is adapted by slightly modifying its prompt templates to ensure appropriate ask-or-answer decisions in the multi-turn clarification setting.

Analysis of Effectiveness and Efficiency. Table 5 reports accuracy for each user persona. Across both backbones, ClarifyAgent consistently outperforms all prompting-based baselines by a notable margin. With Llama-3.1-8B-Instruct, ClarifyAgent achieves an average accuracy of 88.4%, outperforming the strongest baseline by 15.4 absolute points. The improvements are especially pronounced on noisy personas such as Partial–Vague, Off–Focus, Contradictory, and Factually–Wrong. A similar trend is observed on Qwen-2.5-7B-Instruct, where ClarifyAgent improves the average accuracy from 66.4% to 88.0%. Despite these gains, ClarifyAgent does not explicitly optimize for the Precise persona. Instead, it prioritizes balanced and robust behavior across all six personas, achieving the highest average accuracy among compared methods while remaining competitive on precise inputs. This trade-off is desirable in multi-turn clarification, where real-world failures often arise from ambiguous or strategically unhelpful feedback rather than fully specified user queries. In terms of efficiency, ClarifyAgent requires five forward passes of the base LLM (once per module), matching the inference cost of Majority Voting and Intent-Sim, both of which also rely on k=5 samples. Under this comparable computational budget, ClarifyAgent delivers substantially larger accuracy gains, indicating that its performance improvements stem from more effective decision-making rather than increased inference cost.

Ablation Study. Table 6 presents an ablation study on the Qwen-2.5-7B-Instruct backbone, where we systematically disable each module of ClarifyAgent. We observe that removing any component leads to a noticeable drop in overall accuracy. In addition, the standard deviation across personas nearly doubles when either the Perceiver or Forecaster is removed, suggesting that these modules are essential for maintaining balanced performance across heterogeneous user behaviors rather than overfitting to a particular subset. Each module also contributes distinct behavioral effects. For instance, eliminating the Forecaster causes the model to become very strong on Precise and Refusal personas, while its performance degrades markedly on other personas. This indicates that user persona inference is crucial for preserving robustness under noisy or ambiguous feedback.

Overall, the results demonstrate that ClarifyAgent delivers pronounced accuracy gains over strong baselines, remains robust under diverse and noisy user behaviors, and achieves a well-calibrated balance between answering and clarification. The ablations further show that all modules contribute to both accuracy and stability across personas, highlighting their complementary roles within the agentic framework.

7 Conclusion

We presented **ClarifyMT-Bench**, a benchmark for evaluating multi-turn clarification in open-domain human–LLM interactions.

Grounded in a five-dimensional ambiguity taxonomy and six behaviorally diverse user personas, ClarifyMT-Bench enables controlled evaluation of when an LLM should ask, what it should ask, and when it should stop asking. Using this framework, we conducted a systematic study across ten off-the-shelf LLMs and uncovered a consistent under-clarification bias: models tend to answer prematurely, struggle to remain robust under noisy or contradictory user feedback, and degrade substantially as dialogue depth increases. To address this gap, we proposed **ClarifyAgent**, an agentic approach that decomposes clarification into perception, forecasting, tracking, and planning. ClarifyAgent significantly improves ask–answer decisions across user personas, offering a strong baseline for future work. Together, ClarifyMT-Bench and ClarifyAgent reveal key limitations in current conversational LLMs and provide a basis for developing safer, more reliable, and more uncertainty-aware multi-turn interaction.

References

[1] 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* (2020).

[2] Anthropic. 2025. *Claude Sonnet 4.5 System Card.* https://www.anthropic.com/claude-sonnet-4-5-system-card

[3] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7421–7454.

[4] Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan O Arik. 2025. Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training. In *The Thirteenth International Conference on Learning Representations*.

[5] Herbert H Clark. 1992. *Arenas of language use.* University of Chicago Press.

[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[8] Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. 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*. 10602–10621.

[9] 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 e-prints* (2024), arXiv–2407.

[10] Yujian Gan, Changling Li, Jinxia Xie, Luou Wen, Matthew Purver, and Massimo Poesio. 2024. Clarq-llm: A benchmark for models clarifying and requesting information in task-oriented dialog. *arXiv preprint arXiv:2409.06097* (2024).

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* 645, 8081 (2025), 633–638.

[12] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* (2024).

[13] Robert Krovetz and W Bruce Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)* 10, 2 (1992), 115–141.

[14] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*. 611–626.

[15] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120* (2025).

[16] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[17] 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*. 11526–11544.

[18] Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

[19] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2757–2791.

[20] Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717* (2025).

[21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[22] Chung Kwan Lo. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education sciences* 13, 4 (2023), 410.

[23] Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. DuetSim: Building User Simulator with Dual Large Language Models for Task-Oriented Dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 5414–5424.

[24] Yuetian Mao, Junjie He, and Chunyang Chen. 2025. From prompts to templates: A systematic prompt template analysis for real-world LLMapps. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*. 75–86.

[25] David Maxwell and Leif Azzopardi. 2016. Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 1141–1144.

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

[27] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4885–4901.

[28] OpenAI. 2025. ChatGPT. http://chatgpt.com/.

[29] OpenAI. 2025. *Introducing deep research.* https://openai.com/index/introducing-deep-research/

[30] OpenAI. 2025. *Introducing GPT-4.1 in the API.* https://openai.com/index/gpt-4-1/

[31] 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 (2022), 27730–27744.

[32] Kimia Ramezan, Alireza Amiri Bavandpour, Yifei Yuan, Clemencia Siro, and Mohammad Aliannejadi. 2025. Multi-Turn Multi-Modal Question Clarification for Enhanced Conversational Understanding. *arXiv preprint arXiv:2502.11442* (2025).

[33] Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based User Simulator for Task-Oriented Dialogue Systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*. 19–35.

[34] 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*. 1169–1170.

[35] Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023. Metaphorical user simulators for evaluating task-oriented dialogue systems. *ACM Transactions on Information Systems* 42, 1 (2023), 1–29.

[36] Anfu Tang, Laure Soulier, and Vincent Guigue. 2025. Clarifying ambiguities: on the role of ambiguity types in prompting methods for clarification generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 20–30.

[37] Qwen Team. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[38] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. [n. d.]. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

[39] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966* (2023).

[40] Zhenduo Wang, Zhichao Xu, Vivek Srikumar, and Qingyao Ai. 2024. An in-depth investigation of user response simulation for conversational search. In *Proceedings of the ACM Web Conference 2024*. 1407–1418.

[41] 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 (2022), 24824–24837.

[42] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.

[43] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

[44] Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013* (2024).

[45] Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with lms. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 5526–5543.

[46] Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10746–10766.

[47] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

Prompt Template for LLM-as-a-Judge
You are an impartial evaluation assistant. You will be provided with a candidate question, a golden response (the ideal answer), and a candidate response (the model's answer to evaluate). Your task is to assign an alignment score indicating how well the candidate response matches the golden response.
Scoring guidelines (0–5):

- **5** — Fully aligned. Semantically equivalent; no missing key information; no contradictions.
- **4** — Mostly aligned. Minor omissions or differences, but overall meaning preserved.
- **3** — Partially aligned. Contains some correct elements but lacks important information.
- **2** — Weak alignment. Only small portions match; the majority is incomplete or off-target.
- **1** — Barely aligned. Very limited semantic overlap.
- **0** — Not aligned. Irrelevant, contradictory, or does not follow the intent of the golden response (e.g., golden response asks for clarification, but the candidate provides a direct answer).

Instructions:

- Judge semantic meaning, not surface wording.
- Be strict: only credit what the golden response explicitly or implicitly contains.
- Output only a single integer from 0 to 5 with no additional explanation.

Inputs: Candidate question: <question>
Golden response: <golden_clarifying>
Candidate response: <candidate_response>

Figure 6: Prompt template for LLM-as-a-Judge evaluation.

A Ethical Statement

This work does not involve the collection or analysis of real user data. All dialogues in ClarifyMT-Bench are synthetically generated by LLMs or collected from open-source dataset. They are manually validated, and therefore contain no personal, identifiable, or sensitive information. The benchmark poses no privacy or data-protection risks, and does not target any demographic group. Although the dataset models challenging behaviors such as contradictory or factually incorrect replies, these behaviors are abstracted templates rather than reflections of real individuals. Overall, this study introduces minimal ethical risk and aligns with responsible research practices for human–LLM interaction.

B Data Source Distribution

We report the data source distribution of our dataset in Figure 8. As shown, GPT-4.1 and DeepSeek-V3 each contribute roughly one quarter of the samples, open-source dataset (*i.e.*, CLAMBER) account for about 5%, and GPT-5 comprises the remaining 45%. This balanced mixture ensures broad topic diversity and mitigates overfitting to any single model's conversational style.

(a) Prompt Template for Single-turn Dialogue Generation
You are an expert dialogue designer generating short user queries that exhibit <ambiguity_type>.
Task: Produce <number> natural user utterances whose <ambiguity_description>.
For each example, output:

- "question": the ambiguous user request
- "clarifying_question": a natural follow-up asking what scope the user intends
- "explanation": explanation of ambiguity
- Format: JSON per line

(b) Prompt Template for Multi-turn Dialogue Generation
Your task is to continue a short dialogue between a user and an assistant. Given a dialogue where the assistant asks a clarifying question, generate how six different types of users would respond in the next turns. Each response must sound natural. Single-turn types produce one user reply; multi-turn types follow a Q1→A1→Q2→A2 pattern. Output a JSON object with the following keys: <user_personas>
Example: <example>
Now continue for the following dialogue: Q: <user_query>
A: <assistant_query>

(c) Prompt Template for Model Evaluation
You are a helpful conversational assistant. In each turn, given the previous dialogue and the user's latest message, your task is to decide whether to answer the user directly or ask a clarifying question. If the user's request is clear and specific, respond with the final answer. If the request remains ambiguous, underspecified, or missing essential information, respond with an appropriate clarifying question instead.
Your response MUST begin with either 'The answer is' or 'The clarifying question is'.

Figure 7: Prompt templates used for (a) single-turn ambiguity generation, (b) multi-turn dialogue construction across six user personas, and (c) model ask–or–answer evaluation.

C Implementation Details

Unless otherwise stated, we set the decoding temperature to 0, *i.e.*, greedy decoding. Due to resource constraints, Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct are locally deployed and served via the vLLM framework [14]. OpenAI models are accessed through their official APIs, while the DeepSeek models are accessed through the `Bailian` platform.[2] All other models are accessed via the `OpenRouter` API service.[3] Prompt templates follow the general format used in prior clarification benchmarks [10, 46], with minor adjustments to

---

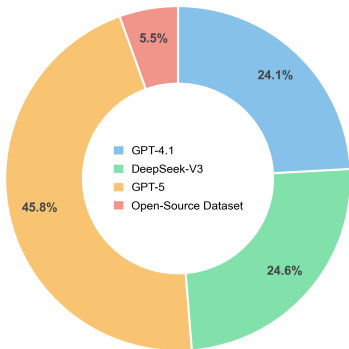[2]https://aliyun.com/product/bailian
[3]https://openrouter.ai/

Table 7: Examples of user responses across six user personas in the travel itinerary scenario.
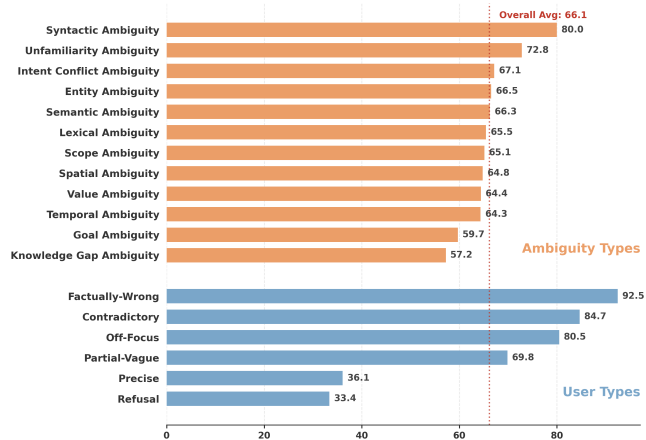
Scenario   Q1: Can you plan a 3-day trip for me?
A1 (Clarification): Could you share the destination, dates, budget, or any preferred activities?

User Persona	Description	Example reply (Q2)
Precise	Provides the missing details clearly and specifically	Kyoto from March 12–14. Budget around $800. I would like to visit temples.
Partial–Vague	Partially related but vague, not decidable	Somewhere in Japan in early spring.
Off–Focus	Responds with information unrelated to the clarification request	Is Japan safe to travel right now?
Contradictory	Provides statements that conflict with each other	Let's keep it cheap. Actually, no need to worry about the budget, make it fancy.
Factually–Wrong	Provides specific but factually incorrect information	Plan something in Kyoto near the Eiffel Tower. I want to visit it on the first day.
Refusal	Declines to clarify and pushes the system to decide	Just make the plan for me. Anything is fine.

Table 8: Ambiguity taxonomy with representative clarifying questions, illustrating how LLMs can engage in pragmatic disambiguation across linguistic, intentional, contextual, epistemic, and interactional dimensions.

Subtype	Example	Clarifying Question
Lexical Ambiguity	Please tell me about the seal.	Do you mean the marine animal or the official stamp?
Syntactic Ambiguity	List movies from the 1990s starring actors from Canada.	Do you mean 1990s films featuring Canadian actors, or Canadian films from that period?
Semantic Ambiguity	Is New York the largest city?	In what respect—by population size, land area, or economic output?
Goal Ambiguity	Help me write a report.	Could you specify the type or purpose of the report—academic, professional, or personal?
Scope Ambiguity	Tell me about quantum computing.	Are you looking for a high-level overview or something more technical?
Intent Conflict Ambiguity	Summarize *War and Peace* without omitting anything.	Should I prioritize completeness or conciseness in the summary?
Entity Ambiguity	Who is the real Spider-Man?	Are you referring to the comic-book character, a film portrayal, or a real person?
Spatial Ambiguity	Tell me how to reach London.	From which location are you starting your journey?
Temporal Ambiguity	When does the meeting start?	Are you referring to today's meeting or a future one?
Knowledge Gap Ambiguity	You remember the new update, right?	Could you clarify which update you are referring to?
Unfamiliarity Ambiguity	Find the price of the Samsung Chromecast.	Did you mean a Samsung streaming device or Google's Chromecast?
Value Ambiguity	Recommend a good movie.	What criteria define "good" for you—genre, popularity, or critical acclaim?
Partial / Vague Reply	Sort of, I guess.	Could you indicate which part you agree with or find uncertain?
Factually Incorrect Reply	Paris is the capital of Germany.	Did you mean Paris, France, or perhaps Berlin, Germany?
Contradictory Reply	It's urgent. No rush actually.	Should I treat this as a high- or low-priority request?
Off-focus Reply	Let's talk about something else.	Of course — would you like to switch to a new topic or pause this discussion?

Figure 8: Distribution of data sources. The dataset is constructed from diverse sources, primarily leveraging high-capability models including GPT-5, DeepSeek-V3, and GPT-4.1, complemented by open-source datasets to ensure high-quality and diverse coverage.

Figure 9: Analysis of average dialogue length, grouped and sorted by ambiguity types (orange) and user types (blue).

ensure consistency across different model families. The complete prompt template is provided in Figure 7.