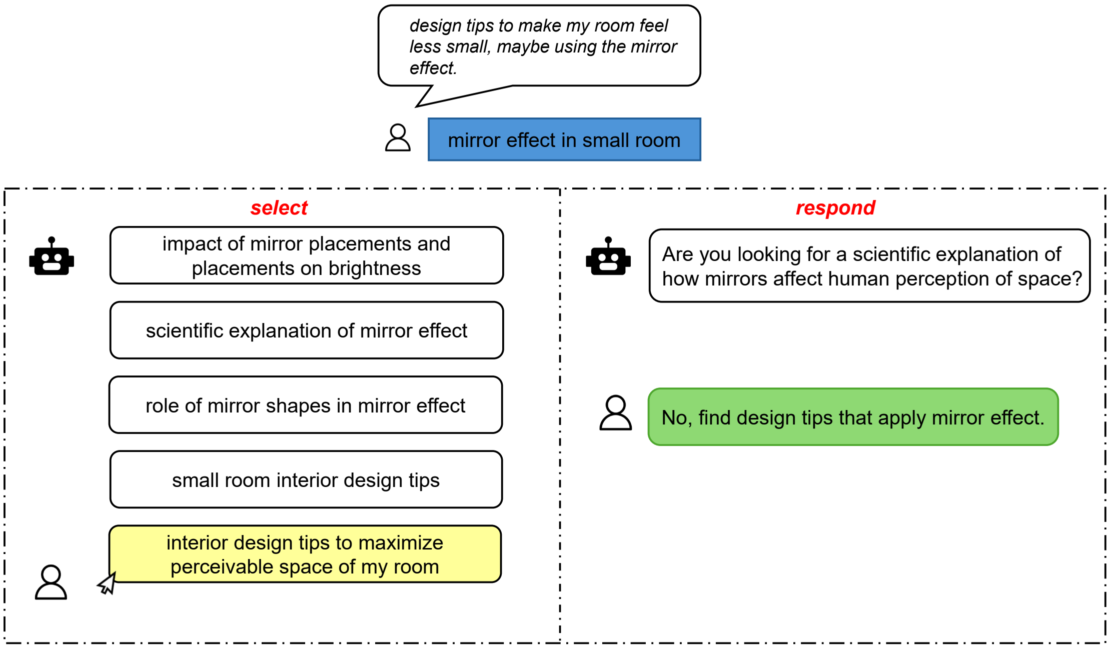
design tips to make my room feel less small, maybe using the mirror effect.
mirror effect in small room
select
impact of mirror placements and placements on brightness
scientific explanation of mirror effect
role of mirror shapes in mirror effect
small room interior design tips
interior design tips to maximize perceivable space of my room
respond
Are you looking for a scientific explanation of how mirrors affect human perception of space?
No, find design tips that apply mirror effect.

Figure 3: Illustration of the two clarification interaction scenarios in our user simulation: *select* and *respond*.

Table 4: Prompts of *respond* and *reformulation*. There is no *reformulation* prompt for the mode *select*.

Prompt type	System instruction

response (scenario *select*)	`Imagine that you are a user seeking information with the help of a conversational assistant. At each turn of the conversation, the assistant provides you several reformulated queries to better understand your intent. Given a conversation history and a paragraph describing the user intent, choose the reformulated query that most accurately reflects the provided user intent.`
	`<chat history>`
	`<user intent>`

response (scenario *respond*)	`Imagine that you are a user seeking information with the help of a conversational assistant. At each turn of the conversation, the assistant asks a clarification question to better understand your intent. Given a conversation history and a paragraph describing the user intent, respond to the clarification question based on the provided user intent.`
	`<chat history>`
	`<user intent>`

reformulation (scenario *respond*)	`Given a conversation history, summarize the conversation as a reformulated query. The conversation history includes the initial query and several clarification turns between the user and a virtual assistant.`
	`<chat history>`

we ask LLMs to regenerate the outputs with a maximum number of retry attempts set to 10. In rare cases, we manually parse the LLM outputs to address persistent parsing errors.

BERTScore. We use the third-ranked pre-trained model of BERTScore[4] based on their experimental results [5] on the WMT16 machine translation task [7].

IR Pipeline. Following [32, 45, 46], we adopt a two-stage retriever-rerank pipeline for IR tasks. Top-k relevant documents are first retrieved from a large-scale document collection using BM25 and then reranked using MonoT5. For the retriever, we use a no-tuning *pyserini*[6] Lucene implementation with k fixed to 100. For the reranker, we use a pre-trained MonoT5 [46].

[4] https://huggingface.co/microsoft/deberta-large-mnli
[5] https://github.com/Tiiiger/bert_score
[6] https://github.com/castorini/pyserini

IR Datasets. We use *ir_datasets* [39], a commonly used Python package in IR community to manage IR datasets. The Python implementation of *pyserini* is used to build BM25 indexes of Clueweb09 and Clueweb12.

5 Task 1: Clarification Generation (CG)

This section aims to evaluate the impact of integrating ambiguities into LLM reasoning on the performance of the clarification generation (CG) task. Table 5 depicts the overall comparison between different prompting methods using various datasets and the BERTScore metric. Results show that AT-CoT consistently outperforms the three baselines with significant margins across all datasets. For instance, AT-CoT reaches a BERTScore of 82 vs. scores ranging from 78.8 to 80 for other baselines on ClariQ. This suggests that ambiguity-oriented reasoning (AT-COT) helps generate better clarifying questions. This improvement is consistent on both specific-domain datasets (RaoCQ) and open-domain datasets (Qulac, ClariQ), showing that our method generalizes to different types of queries. Besides, through the comparison between AT-standard and standard prompting, we find that only informing LLMs of existing ambiguity types is not helpful, and even degrades the CG performance in some cases (e.g. 77 vs 77.9 for resp. AT-standard vs. standard on Qulac). This demonstrates that integrating ambiguity types is only helpful when integrated into LLM reasoning. Our observation on CoT prompting is coherent with previous work [24, 61]: CoT is more effective than standard prompting in terms of generating clarifying questions. We explore even further by claiming that ambiguity-oriented reasoning is more helpful.

Stratification by Ambiguity-level. We further evaluate the performance of CG tasks across different ambiguity levels. We use labels provided in ClariQ for this analysis, and present results in Table 6. Generally, both CoT and AT-CoT outperform standard and AT-standard prompting on the first three ambiguity levels, demonstrating the usefulness of both freely generated LLM reasoning and ambiguity-oriented reasoning for CG when queries are not extremely ambiguous. However, in cases of extreme ambiguity (level-4), the performance of CoT falls below standard prompting (BERTScore of 78 vs. 78.5 and 79.7 for standard and AT-standard resp.), meanwhile, the improvement of AT-CoT is still consistent (BERTScore=82.4). This suggests that ambiguity types could be particularly useful for handling ambiguous queries.

Distribution of Ambiguity Types. To provide more insight into AT-CoT, we analyze the distribution of ambiguity types predicted by AT-CoT. We first investigate the frequency of predicted ATs (namely *Semantic*, *Generalize* and *Specify*), and then focus on the impact of predicting different ATs on the performance of the CG task. Predicted ATs are extracted from the reasoning generated by AT-CoT. Table 7 shows statistics about each group on all CG datasets: the frequency of queries identified as a specific AT and the performance difference in terms of BERTScore between AT-CoT and CoT (in parentheses). Our main conclusions are the following: 1) *Semantic* and *Specify* are the most frequent types for all CG datasets, with *Specify* being slightly more common (i.e. at most 45.9% vs. at most 53.7% for *Specify*). This observation aligns with the fact that most ATs in existing taxonomies can be categorized as *Semantic* or *Specify*. However, though less frequent, the importance of *Generalize* cannot be overlooked. 2) The *Generalize* type is more marginal but the fact that 12% of queries in RaoCQ are predicted to be generalizable justifies our decision to include *Generalize* in our taxonomy. 3) The observation that queries in RaoCQ more often require generalization suggests that the AT predictions of AT-CoT effectively capture the clarification needs of queries and are less likely to be random. Since RaoCQ queries are extracted from user posts on StackExchange, they are generally longer compared to queries in Qulac and ClariQ. It is therefore very likely that a query in RaoCQ does not precisely describe user intents and requires generalization. Differently, queries in Qulac are often short, used for web search, making them less possibly to require generalization. This gap between the frequency of *Generalize* being predicted and improvements caused by predicting *Generalize* reflects that AT-CoT adapts well to datasets with different characteristics.

Table 5: Overall evaluation on CG datasets. *, †, $^\Delta$ marks statistically significant improvements over standard, AT-standard, CoT respectively with $p < 0.01$ under a t-test.

Prompt	Qulac	ClariQ	RaoCQ
standard	77.9	79.3	60.0
AT-standard	77.0	78.8	59.9
CoT	79.2^*	80.0	60.5
AT-CoT	$\mathbf{80.6}^{*\dagger\Delta}$	$\mathbf{82.0}^{*\dagger\Delta}$	$\mathbf{62.4}^{*\dagger\Delta}$

Table 6: CG results on ClariQ stratified by ambiguity levels. *, †, $^\Delta$ marks statistically significant improvements over standard, AT-standard, CoT, respectively.

	level-1	level-2	level-3	level-4
standard	78.7	80.0	78.9	78.5
AT-standard	77.6	79.2	78.4	79.7
CoT	78.6	80.5	$80.7^\dagger$	78.0
AT-CoT	$\mathbf{80.9}^{*\dagger}$	$\mathbf{82.0}^{*\dagger\Delta}$	$\mathbf{82.1}^{*\dagger\Delta}$	$\mathbf{82.4}^{\dagger\Delta}$

Table 7: Distribution of ATs predicted by AT-CoT. In parentheses, we show corresponding CG performance differences between AT-CoT and CoT regarding the BERTScore.

	Qulac	ClariQ	RaoCQ
Semantic	44.6 (⇑ 1.3)	45.9 (⇑ 1.8)	42.4 (⇑ 2.0)
Generalize	1.7 (⇓ 0.6)	1.9 (⇑ 1.4)	12.3 (⇑ 2.0)
Specify	53.7 (⇑ 1.4)	52.2 (⇑ 2.0)	45.3 (⇑ 1.9)

6 Task 2: Information Retrieval (IR)

This section aims to investigate the impact of integrating ambiguity into LLM reasoning on IR performance. Table 8 shows IR results of the two different interaction scenarios (*select* & *respond*) and the baseline without clarification. We detail the result alongside three successive turns. Generally, we observe that AT-CoT > CoT > AT-standard ≈ standard for most of the interaction modes

Table 8: Results on IR datasets based on user simulation. Scores are in nDCG@10 (%) for Trec Web Track 2009-2012 and TREC Web Track 2013-2014; MRR@10 (%) for TREC DL Hard. *, †, $^\Delta$, indicates statistically significant improvements over standard, AT-standard, CoT respectively with $p < 0.01$ under a t-test.

	TREC Web Track 09-12		*TREC Web Track 13-14*		*TREC DL Hard*	
	select	*respond*	*select*	*respond*	*select*	*respond*
w/o clarification	0.123	0.123	0.277	0.277	0.084	0.084
Turn-1						
standard	0.161	0.232	0.336	0.387	0.060	0.120
AT-standard	0.165	0.230	0.337	0.383	0.066	0.113
CoT	0.174*†	0.238	0.341	0.392†	0.063	0.123†
AT-CoT	**0.188**$^{*\dagger\Delta}$	**0.244***†	**0.347***†	**0.397**†	**0.074**$^{*\Delta}$	**0.125**†
Turn-2						
standard	0.152	0.223	0.307	0.379	0.054	0.127
AT-standard	0.149	0.228	0.291	0.376	0.052	0.151*
CoT	0.160*†	0.226	0.310†	0.384	0.062†	0.174*†
AT-CoT	**0.176**$^{*\dagger\Delta}$	**0.233**	**0.320**$^{*\dagger\Delta}$	**0.391***†	**0.071**$^{*\dagger\Delta}$	**0.184**$^{*\dagger\Delta}$
Turn-3						
standard	0.141	0.212	0.295	0.371	**0.056**	0.141
AT-standard	0.149	0.213	0.276	0.367	0.051	0.154
CoT	0.148	**0.216**	0.300†	0.373	0.054	0.184*†
AT-CoT	**0.152**	0.213	**0.305**†	**0.381**	0.052	**0.188***†

and turns. For instance, clarifications obtained with the method AT-COT allow to reach the best IR metrics values for the TREC Web Track 2013-2014 dataset over all turns (0.397, 0.391, and 0.381 for each turn respectively vs 0.392, 0.384, and 0.373 at most for the baselines). We also note that IR performance is always better for the *respond* interaction mode corresponding to the generation of clarifying questions (in contrast to the *select* mode based on query reformulation. Altogether, these results highlight two main conclusions: 1) it aligns with our remarks on the CG performance, demonstrating the benefits of introducing ambiguity-oriented LLM reasoning for clarification, both intrinsically and extrinsically. And 2) this reinforces our hypothesis based on the need for clarification interactions based on ambiguity and reasoning in IR. Our findings also demonstrate the robustness of our methodology in interaction scenarios. For both interaction scenarios *select* and *respond*, AT-CoT consistently provides the best IR performance, implying that our method can adapt to various real-world scenarios such as query suggestion-based scenarios (e.g. search suggestions) or chat scenarios (e.g. chatbot).

Per-turn IR performance. We observe the same pattern of performance changing across multiple conversation turns for all prompting schemes. For example, under *select*, the IR performance reaches the highest value in the first turn, then monotonically decreases; for Trec DL Hard under *respond*, the IR performance steadily increases as the conversation continues. This IR performance changing pattern is coherent to query difficulties. As a collection that contains complex queries from Trec DL 2019/2020 datasets [21, 22], queries in Trec DL Hard are relatively longer, more challenging in terms of resolving ambiguities. Therefore, Trec DL Hard may necessitate multi-turn conversations to fully clarify ambiguities, which is reflected in the increasing scores across conversation turns. Similarly, for Trec Web Track datasets, the peak IR performance appearing at the first turn is reasonable, since queries in these datasets are not highly ambiguous. Nevertheless, in terms of turn-specific IR performances, AT-CoT still outperforms other prompting schemes, demonstrating that there is no need to increase conversation turns to reflect the improvements of AT-CoT. Regardless of how many turns of conversation a user intends to have, AT-CoT is able to provide better clarifications compared to other prompting schemes.

7 Task 3: Alignment between Clarification Generation & Information Retrieval

To mitigate potential bias introduced in user simulation, we further align the performance of CG and IR by using Qulac-Trec Web Track 2009-2012 as mentioned in Section 4.1.3. Since reference CQs in Qulac are only provided for initial queries, we use the CG and IR results from the first turn under *respond.* We compute the Pearson correlation coefficient to measure the strength of the linear relationship between the CG and IR results, and obtain $r = 0.92$, $p = 0.08$, which shows a strong positive correlation. We hypothesize that the insignificance of this correlation may due to the complexity of the document collection, which is insufficient to differentiate the quality of clarifications. A query may be refined by high-quality CQs through user simulation, but there lack of relevant documents to account for this refinement and reflect it by IR performance. However, given that we obtain a correlation coefficient greater than 0.9, it does not undermine our observation that IR performance is correlated to CG, i.e. IR performance improvements brought by AT-CoT are due to better clarifications.

8 Conclusion

In this work, we investigate the integration of ambiguities and reasoning into LLM prompting methods for clarification, proposing a new action-based ambiguity type taxonomy and a new prompting scheme, AT-CoT. Experiments on clarification generation and information retrieval datasets demonstrate the effectiveness of our methodology. Besides, in-depth analyses show that our method is robust in different clarification interaction scenarios and can capture the clarification needs of datasets with different characteristics.

However, our work is not without limitations. First, we establish an ambiguity type taxonomy containing three general ATs for integration with LLM reasoning. We do not experimentally study the impact of AT granularity, particularly investigating whether reasoning over a structured ambiguity taxonomy would be beneficial. Second, we only use Llama-3-8B without testing LLMs of different scales. It would be interesting to study the reasoning capability of different model scales. Nevertheless, we believe that our work acts as a foundation to better understand the role of ambiguity types in LLM prompting methods for clarification and may provide useful insights for future work.

Acknowledgments

This work benefited from support from the French National Research Agency (Project GUIDANCE, ANR-23-IAS1-0003) and SCAI (Sorbonne Center for Artificial Intelligence).