MTRAG-UN: A Benchmark for Open Challenges in Multi-Turn RAG Conversations

Sara Rosenthal, Yannis Katsis, Vraj Shah,
Lihong He, Lucian Popa, Marina Danilevsky
IBM Research, USA
sjrosenthal@us.ibm.com

Abstract

We present MTRAG-UN, a benchmark for exploring open challenges in multi-turn retrieval augmented generation, a popular use of large language models. We release a benchmark of 666 tasks containing over 2,800 conversation turns across 6 domains with accompanying corpora. Our experiments show that retrieval and generation models continue to struggle on conversations with UNanswerable, UNderspecified, and NONstandalone questions and UNclear responses. Our benchmark is available at https://github.com/IBM/mt-rag-benchmark

1 Introduction

Seeking information continues to be a popular use case for Large Language Models (LLMs) (Wang et al., 2024). Thus, Retrieval Augmented Generation (RAG), particularly in the multi-turn interactions of LLM chat interfaces (Li et al., 2025), remains an important research area. Several benchmarks have been released to evaluate model performance on such tasks (Dziri et al., 2022; Aliannejadi et al., 2024; Kuo et al., 2025). In particular, the recent MTRAG benchmark (Katsis et al., 2025) focused on multi-turn information-seeking conversations, constituting 842 tasks in four domains. They reported several interesting findings that highlighted areas of improvement including unanswerable questions and later conversation turns.

We pick up on these suggested areas by focusing on user goals that are not achievable via a single question-response[1] exchange with an LLM. We show this via a new benchmark, complementary to MTRAG, that focuses on:

- UNanswerable Question - the user question is not answerable (Katsis et al., 2025)
- UNderspecified Question - the user question is ill-formed or ambiguous, lacking the information to determine a clear intent

[1]We use 'question' to refer to any user utterance

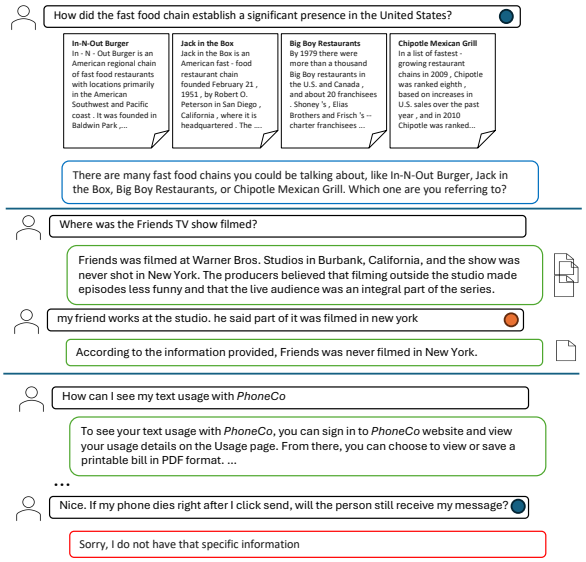


Figure 1: Portions of three conversations highlighting the challenges in MTRAG-UN. The answerability is shown using the assistant response color: answerable, unanswerable, and underspecified. The multi-turn type is shown using the question circle: follow-up and clarification. The last two examples show non-standalone questions.

- NONstandalone Question - the user question cannot be understood without the prior turns
- UNclear Response - the user doesn't understand, or disagrees with the model answer and requires clarification

We thus refer to this new benchmark as MTRAG-UN. An example of each task is shown in Figure 1. Our analysis shows that most frontier models struggle with handling such tasks, jumping to answer based on plausible but assumed interpretations of user intent. These challenges persist in both the retrieval and generation steps of multi-turn RAG.

Our contributions are as follows:

- We present unexplored areas: UNanswerable, UNderspecified and NONstandalone user questions; and UNclear model responses.
- We add multi-turn conversations in two new corpora, Banking and Telco, to explore the use case

Findings of the Association for Computational Linguistics: ACL 2026, pages 10363–10369

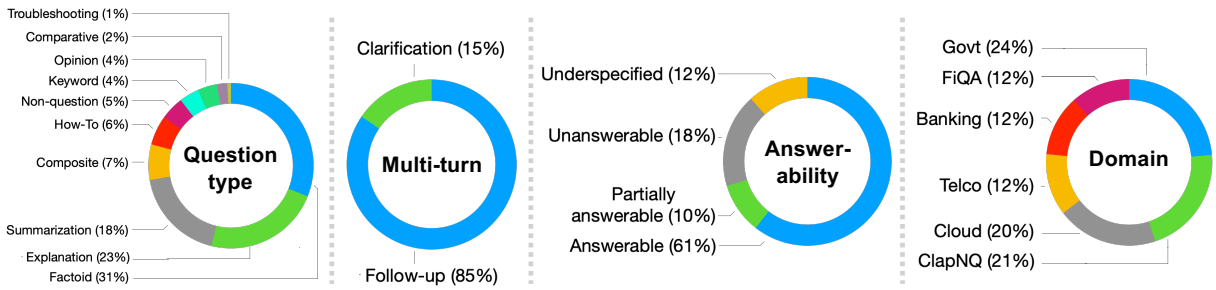


Figure 2: Distribution of tasks in MTRAG-UN based on different dimensions.

of chatbots that are deployed in enterprise settings to support information-seeking questions

- We release MTRAG-UN: A comprehensive benchmark consisting of 666 tasks for evaluating Retrieval, Generation, and the full RAG pipeline. The benchmark is available at: https://github.com/IBM/mt-rag-benchmark

2 Benchmark Creation

We describe the tasks presented in MTRAG-UN, as well as the document corpora used for the reference passages. The conversations were created by human annotators following the process described in (Katsis et al., 2025), using the RAGAPHENE platform (Fadnis et al., 2025). We collect a total of 666 human-generated conversations, with an average of 8 turns per conversation, and we describe the transformation of these conversations into the benchmark tasks at the end of this section.

2.1 Task Definitions

UNanswerable Question. Such a question cannot be answered from retrieved passages, because no relevant passages could be found by the annotator. The MTRAG Benchmark (Katsis et al., 2025) showed that unanswerable questions are challenging for most LLMs. We ask annotators to include at least 2 unanswerable questions in each conversation, to ensure a sufficient and diverse data pool.

UNderspecified Question. A user question may be underspecified, ill-formed, or ambiguous, thus lacking enough information to determine a single clear intent. In such cases, rather than producing a wrong answer or replying with "I don't know", the LLM agent should detect that the user question is unclear and get back to the user, either by pointing out missing details, presenting several plausible interpretations, or listing options based on the underlying passages. Conversations with underspecified questions were created via a combination of human and synthetic generation. In the former, annotators were asked to write conversations that explicitly ended with an underspecified question. In the latter, an underspecified question (also written by a human) was stitched as a last turn on an existing human annotated multi-turn conversation. Relevant passages were added for the underspecified question using query expansion methods with a context relevance filter, in order to generate a rich set of passages to simulate the case of multiple interpretations. The reference response was generated using an LLM followed by human correction. The resulting conversations went through a careful human validation process. Appendix B gives further details.

Corpus	Documents (D)	Passages (P)	Avg P/D
Banking	4,497	33,380	7.4
Telco	4,616	52,350	11.3

Table 1: Statistics of new document corpora in MTRAG-UN.

NONstandalone Question. In a multi-turn conversation, later turns can implicitly reference information in earlier turns. Such questions are considered non-standalone as they require the prior turns to be understood. We directed the annotators to include more non-standalone questions, an interesting challenge for retrieval.

UNclear Response (aka Clarification). In a multi-turn conversation, a user may want to ask a clarification question if they don't clearly understand or disagree with the model answer to their previous question (e.g., "it was filmed in new york" in Figure 1). Though the MTRAG benchmark included some clarification questions, these were not separately called out or evaluated.

2.2 Document Corpora

MTRAG-UN consists of six document corpora: the original four corpora included in MTRAG (CLAPNQ (Rosenthal et al., 2025), FiQA (Maia et al., 2018), Govt, Cloud), and two new corpora

		Recall		nDCG	
		@5	@10	@5	@10
BM25	LT	0.29	0.38	0.27	0.31
	RW	0.36	0.47	0.34	0.39
BGE-base 1.5	LT	0.25	0.32	0.23	0.26
	RW	0.38	0.49	0.35	0.40
Granite R2	LT	0.29	0.38	0.28	0.32
	RW	0.40	0.51	0.37	0.42
Elser	LT	0.40	0.49	0.36	0.40
	RW	0.49	0.60	0.45	0.51

Table 2: Retrieval Performance using Recall and nDCG metrics for Last Turn (LT) and Query Rewrite (RW)

from the domains of Banking and Telco (see Table 1.) These new domains provide enterprise content, an unexplored area in MTRAG and other RAG benchmarks. Each of the corpora was created by crawling ~ 1K web-pages from several companies in the banking and telecommunications sector using seed-pages and crawling their neighborhood to ensure sets of inter-connected pages suitable for writing complex conversations on a given topic.

2.3 Benchmark: Tasks and Statistics

From each conversation, we picked a single turn and created an evaluation task containing the entire conversation up to (and including) the question of the chosen turn, leading to the 666 evaluation tasks comprising the MTRAG-UN benchmark. For conversations with underspecified questions, we chose the turn containing the underspecified question. The remaining conversation turns were picked through a random process biased to give preference to challenging UN-turns. The resulting distribution of tasks is shown in Figure 2. Compared to MTRAG, the MTRAG-UN benchmark includes 6 instead of 4 domains, contains underspecified questions, has a higher representation of unanswerables/partially answerables (a combined 28% vs 15% in MTRAG), and a set of explicitly labeled clarification questions (15% of the tasks). MTRAG-UN is also biased against selecting the first turn of a conversation (8% of the tasks - see Appendix, Figure 4) , which was found to be easier for LLMs (Katsis et al., 2025).

3 Evaluation

We report retrieval and evaluation results on the MTRAG-UN benchmark. Unless otherwise specified, all experiments and settings mimic the MTRAG paper (Katsis et al., 2025).

	Subset	LT	RW
Standalone	No (214)	0.39	0.52
	Yes (254)	0.40	0.46

Table 3: Elser R@5 standalone results

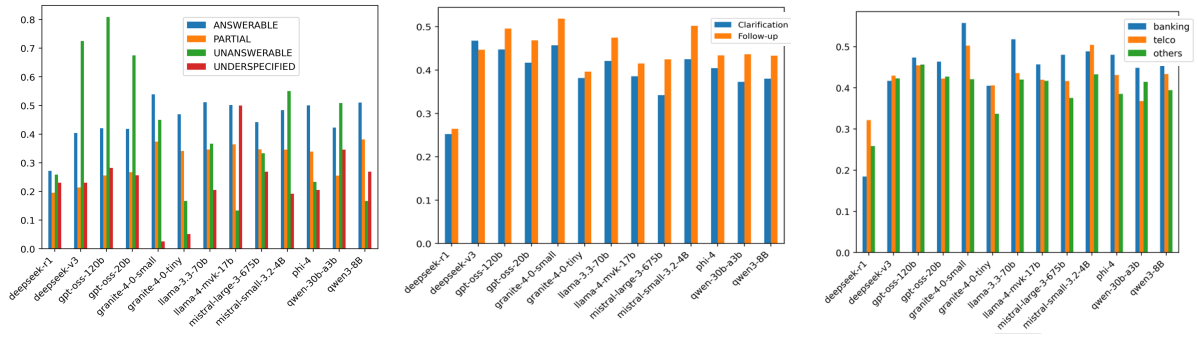
3.1 Metrics

We adopt the evaluation metrics of (Katsis et al., 2025): (1) reference-based RB_{llm} and RB_{alg}, (2) the IDK ("I Don't Know") judge, and (3) faithfulness judge from RAGAS RL_F. All evaluation metrics are conditioned to account for answerability. We use the open-source GPT-OSS-120B instead of the proprietary GPT-4o-mini as judge (Correlation is still aligned with human judgments - see Appendix A). All other judges are consistent with those reported in MTRAG (Katsis et al., 2025). We create a new metric for the underspecified instances run with GPT-OSS-120b (See prompt in Appendix Figure 6). Its accuracy on 80 random llama-4 and gpt-oss-120b model responses from underspecified instances is 96.2%. These instances are not classified using the other metrics.

3.2 Retrieval

We ran retrieval experiments on the 468 answerable and partially answerable questions. We follow the experiments in MTRAG by running on lexical (BM25), sparse (Elser), and dense models. We added a newer SOTA dense embedding model, Granite English R2 (Awasthy et al., 2025), and compare it to BGE-base 1.5 (Xiao et al., 2023) as reported in the original paper. We also experimented with using newer open source models for Query Rewrite (Sun et al., 2023) with the same prompt reported in the MTRAG paper and found that GPT-OSS 20B performed best. In all cases Query Rewrite outperforms the last turn. Granite English R2 performs better than BGE-base 1.5 embeddings, but Elser still performs best. The macro-average results across all domains are shown in Table 2. We also provide a breakdown by standalone as provided in MTRAG in Table 3[2]. Rewrite helps for both standalone and non-standalone questions, but more so for non-standalone questions.

[2]Non-standalone questions were computed differently than in MTRAG. We adjusted our definition to match co-ref in any of the user turns instead of just the last turn. This adjustment is more suitable for MTRAG-UN because each conversation only appears once in contrast to MTRAG where each turn in the conversation is a task. Co-ref in any turn could indicate a rewrite is required. For comparison, in MTRAG-UN 16.6% of the last turns had co-ref compared to 14.3% in MTRAG.



(a) By question answerability (b) By multi-turn type (c) By domain

Figure 3: Generation results in the Reference (•) setting using, RB_{alg}, on three different dimensions.

	RL_F		RB_{llm}		RB_{alg}	
	•	○	•	○	•	○
target	0.85	0.69	0.96	0.92	0.89	0.88
gpt-oss-120b	**0.65**	0.59	**0.76**	**0.65**	**0.46**	0.37
gpt-oss-20b	0.60	0.55	0.67	0.63	0.43	0.36
deepseek-v3	0.63	**0.60**	0.61	0.58	0.42	0.37
deepseek-r1	0.47	0.46	0.54	0.52	0.26	0.23
granite-4-0-small	0.62	0.56	0.55	0.53	0.45	**0.38**
granite-4-0-tiny	0.48	0.46	0.50	0.50	0.35	0.31
qwen-30b-a3b	0.61	**0.60**	0.68	0.60	0.41	0.36
qwen3-8B	0.57	0.55	0.64	0.58	0.41	0.36
llama-4-mvk-17b	0.62	0.58	0.59	0.57	0.42	0.37
llama-3.3-70b	0.62	0.58	0.58	0.55	0.43	**0.38**
mistral-small 24b	0.63	0.59	0.67	0.57	0.45	0.37
mistral-large 675b	0.55	0.52	0.67	0.60	0.39	0.34
phi-4	0.54	0.49	0.64	0.57	0.40	0.34

Table 4: Generation by retrieval setting: Reference (•) and RAG (○). The best result is **bold** and runner-up is underlined.

Our new domains of Banking and Telco perform worse than the other domains with .32 and .39 R@5 respectively (compared to an average of .52 R@5 for the other domains). To investigate this gap, we analyzed corpus-level characteristics and found that Banking and Telco contain substantially longer documents and denser hyperlink structures, suggesting stronger cross-page dependencies typical of enterprise web content. Additionally, these domains include multiple companies with structurally similar pages (e.g., checking accounts or credit card offers), which likely increases retrieval difficulty due to content similarity across sources. Overall, our scores are lower than MTRAG, highlighting that more work is needed for multi-turn retrieval.

3.3 Generation

We ran generation experiments using the original prompt used in MTRAG (Katsis et al., 2025) with an additional sentence to to accommodate the possibility of underspecified questions:

> Given one or more documents and a user question, generate a response to the question using less than 150 words that is grounded in the provided documents. If no answer can be found in the documents, say, "I do not have specific information". If a question is underspecified — e.g., it has multiple possible answers, a broad scope, or needs explanation — include that further clarification/information is needed from the user in your response.

In the reference task we send up to the first 10 relevant passages for generation. In the RAG task, we send the top 5 retrieved passages using Elser with query rewrite.

Table 4 presents the generation evaluation results for both reference and RAG settings. We evaluate a diverse set of LLMs, including GPT-OSS (OpenAI, 2025), DeepSeek-V3 (DeepSeek-AI, 2024), DeepSeek-R1 (DeepSeek-AI, 2025), Granite-4 (IBM, 2025), Qwen3 (Qwen, 2025), Llama (Meta, 2025, 2024), Mistral (Mistral AI, 2025), and Phi-4 (Abdin et al., 2024). Model scores remain significantly lower than target answer scores, indicating room for improvement in multi-turn RAG. Larger models usually perform better within each model family, and performance in the reference setting is consistently higher than RAG, reflecting the added difficulty introduced by retrieval noise. GPT-OSS-120B achieves the best scores, while DeepSeek-V3, Qwen-30B and Mistral-Small-24B remain competitive.

Figure 3 shows the generation quality by different dimensions: answerability, multi-turn type, and domain. While most models perform worse on unanswerables, DeepSeek-V3 and GPT-OSS models exhibit comparatively robust behavior by frequently responding with IDK. This is a stark improvement over the takeaways from prior work (Katsis et al., 2025), where no models handled unanswerables well. Performance on underspecified question is consistently low, as models are generally eager to answer based on a plausible but assumed interpretation of the question. Clarification questions show lower performance than follow-

up questions. This suggests that current models are better at conversational continuation than for intent refinement and self-correction. We find that the performance across the two new domains is largely comparable, while the other domains (average performance reported in Figure 3c) trend lower due to the challenging FiQA corpus (Katsis et al., 2025).

4 Conclusion and Future Work

The MTRAG-UN benchmark of 666 tasks and baseline results provided in our paper highlight existing and ongoing challenges in multi-turn RAG. We release our benchmark[3] to encourage advances in this important topic. In the future, we plan to release multilingual RAG conversations.

5 Acknowledgments

We would like to thank our annotators for their high-quality work in generating and evaluating this dataset: Mohamed Nasr, Joekie Gurski, Tamara Henderson, Hee Dong Lee, Roxana Passaro, Chie Ugumori, Marina Variano, and Eva-Maria Wolfe.

Limitations

Our conversations are limited to English and 6 closed domains. They are created by a small set of human annotators and thus likely contain biases toward those individuals and Elser retriever and the Mixtral 8x7b generator used to retrieve passages and generate the initial response respectively. Expanding the annotator pool and creating conversations in other languages would improve these limitations.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 819–829, New York, NY, USA. Association for Computing Machinery.

Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vignesh P, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, Ken Barker, Martin Franz, Madison Lee, Todd Ward, Salim Roukos, David Cox, Luis Lastras, Jaydeep Sen, and Radu Florian. 2025. Granite embedding r2 models. *Preprint*, arXiv:2508.21085.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv*, abs/2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Kshitij Fadnis, Sara Rosenthal, Maeda Hanafi, Yannis Katsis, and Marina Danilevsky. 2025. Ragaphene: A rag annotation platform with human enhancements and edits. *Preprint*, arXiv:2508.19272.

IBM. 2025. IBM Granite 4.0 models.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Tzu-Lin Kuo, FengTing Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-shan Shiu. 2025. RAD-bench: Evaluating large language models' capabilities in retrieval augmented dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 868–902, Albuquerque, New Mexico. Association for Computational Linguistics.

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. *Preprint*, arXiv:2504.04717.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Meta. 2024. Llama 3 models.

Meta. 2025. Llama 4 models.

[3] https://github.com/IBM/mt-rag-benchmark

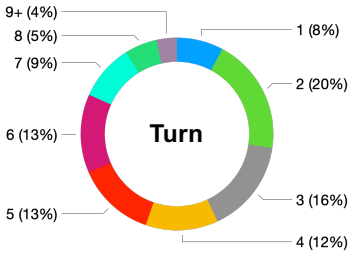
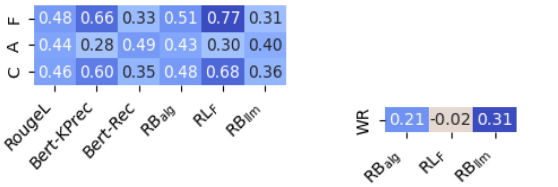


Figure 4: Distribution of tasks in MTRAG-UN based on conversational turn.



(a) With *Faithfulness (F)*, *Appropriateness (A)*, and *Completeness (C)*.

(b) With Win-Rate (WR)

Figure 5: Weighted Spearman correlation: automated judge metrics vs human evaluation metrics.

```
[Instruction]
You are an assistant that determines whether a given response is asking for
clarification. Output "yes" if it is a clarification response, "no" otherwise.

A clarification response is a response whose primary purpose is to request
additional information needed to resolve an ambiguity, underspecification, or
unclear reference in the user's original message.

Clarification responses may appear in two forms:
1. **Direct clarification questions**
  – Explicit questions asking the user to specify or choose among options.
  Example: "Which one are you referring to?"

2. **Directive clarification requests**
  – Imperative or polite statements that ask the user to *provide specific
missing information*.
  – These still count even if they contain no question mark.
  Example: "Please provide the book you are referring to."

Not clarification responses:
- Statements that merely express inability to answer **without requesting the
missing info**.
- Responses that mention missing information but do **not** directly ask
the user to provide it.
  Example: "I can't answer because I don't have your location."
- Responses that ask unrelated questions or introduce new topics.

Response:
{{ response }}
```

Figure 6: Prompt used for clarification judge.

Mistral AI. 2025. Mistral ai open models. Includes Mistral Small and Large models.

OpenAI. 2025. GPT-OSS-120B and GPT-OSS-20B open-weight models.

Qwen. 2025. Qwen 3 models.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025. CLAPnq: Cohesive long-form answers from passages in natural questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Zhongkai Sun, Yingxue Zhou, Jie Hao, Xing Fan, Yanbin Lu, Chengyuan Ma, Wei (Sawyer) Shen, and Chenlei (Edward) Guo. 2023. Improving contextual query rewrite for conversational ai agents through user-preference feedback learning. In *EMNLP 2023*.

Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. Understanding user experience in large language model interactions. *Preprint*, arXiv:2401.08329.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

A Stats and Metrics

A distribution of tasks by turn is provided in Figure 4. MTRAG-UN does not include conversational questions (e.g., "Hi", "Thank you"), since, as noted in MTRAG (which included them in the benchmark but not in the evaluation), more work is required to develop appropriate evaluation metrics for them.

To ensure that using GPT-OSS-120B in place of the GPT-4o-mini as the judge does not negatively affect the quality of the evaluation results, we repeated the correlation analysis of (Katsis et al., 2025) using the open-source model as the judge. The results are depicted in Figure 5. We observe that the correlation between the judges and the humans judgments using the open source model improved slightly or remained consistent compared to using the proprietary model as the judge.

B Details on UNderspecified

Figure 1 shows an example of a conversation where the last user turn is an underspecified question (asking about a vague fast food chain in the US), together with a set of reference passages from the corpus, and a target response for what the model should ask back from the user. The patterns for the model response follow three general categories, each ending with a request to the user to give more information (see also Table 5):

1. Hedging with answers (for the case with few options – e.g., 2-3): list the few options and provide a brief description or answer associated with each.

2. Hedging over list (for the cases with medium number of options – e.g., 4-8): an enumeration of the plausible options without additional explanatory content.

3. Open-domain (for the cases where there are many/unbounded options): directly ask the user for disambiguation over the *type* of entity that they may have in mind.

Pattern	Example
Hedging with answer	There are many astronauts you could be referring to, such as Ellen Ochoa, who was the first Hispanic woman to go to space and has received numerous awards, including the Presidential Medal of Freedom, or Kalpana Chawla, who was the first woman of Indian origin to go to space and tragically died in the Columbia disaster in 2003. Which one are you talking about?
Hedging over list	There are many fast food chains you could be talking about, like In-N-Out Burger, Jack in the Box, Big Boy Restaurants, or Chipotle Mexican Grill. Which one are you referring to?
Open-domain	Which modern smart design segment are you talking about?

Table 5: Types of response to underspecified questions.

B.1 Stitching of the underspecified questions

In Section 2.1, we described underspecified questions written by a human that were stitched as a last turn onto an existing human annotated multi-turn conversation. The process we implemented was a very controlled one, where stitching was done in two ways: a) by finding existing conversations on the same or very similar topic, simulating the case where the new turn is not out of place (75% of the underspecified tasks), and b) by finding existing conversations on a different topic, so that the new turn reflects a topic switch by the user, while still being an underspecified question (25% of the underspecified tasks). The second case could change the flow of the conversation, but we believe that it adds an additional challenge to the models evaluated on such data. In particular, it reflects the realistic scenario where users change topics sometimes randomly, but we still want the models to be able to detect that and react accordingly.

B.2 Validation

The underspecified questions went through careful validation including filtering (e.g., of cases where the last turn intent would accidentally become clear because the addition of the context in which it is being stitched onto), editing of the last turn or of the reference model response to it, and most of the time just plain validation.