MTRAGEval: Evaluating Multi-Turn RAG Conversations

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

1 Overview

Large Language Models (LLMs) have become extremely popular as chat-based assistants. One common use is seeking information where it is important to receive a reliable and trustworthy response by grounding the answer in relevant passages. Retrieval augmented generation (RAG) has therefore become an important and popular field in recent years, including the recent shared task, TREC RAG eval (Pradeep et al., 2024), on a single question.

Work has also shifted beyond question answering to Multi-Turn RAG conversations (Aliannejadi et al., 2024; Dziri et al., 2022; Feng et al., 2021; Kuo et al., 2024; Katsis et al., 2025). This includes the TREC iKAT (Aliannejadi et al., 2024) task, which focused on personalized Multi-Turn RAG. In the more general setting, a recently released new benchmark called MTRAG (Katsis et al., 2025) shows that there is still a need for improvement in RAG-based chat. For instance, it highlights some open areas of improvement such as answerability (knowing when to answer a question or not) and later turns (turns beyond the first turn are more challenging due to non-standalone information). To the best of our knowledge, MTRAG is the first benchmark that uses active retrieval (i.e., real-time retrieval), provides long answers, includes unanswerables, and has multiple domains. These properties make it an ideal benchmark for evaluating all aspects of the RAG pipeline: A) Retrieval, B) Generation, and C) RAG. In this task, we will expand on this work by presenting three subtasks (see Figure 1) aligned with evaluating these RAG properties:

- **Subtask A) - Retrieval**
- **Subtask B) - Generation with Reference Passages (Reference)**
- **Subtask C) - Generation with Retrieved Passages (RAG)**

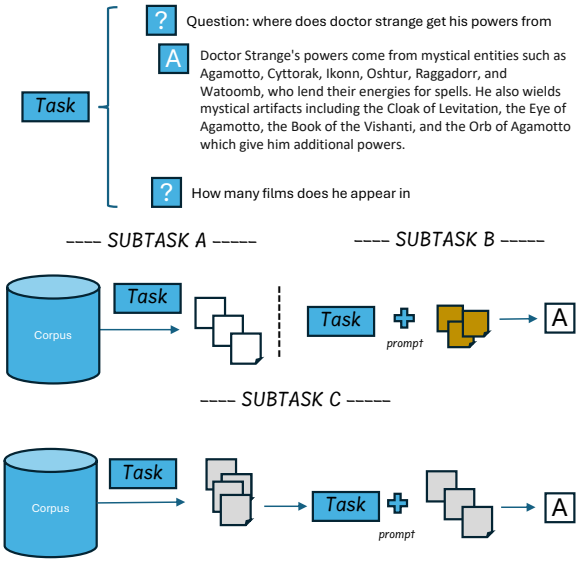

Figure 1: A description of the input and task for the three proposed subtasks. A TASK with two turns is shown. Subtask A) Given a TASK, Retrieve the relevant passages for the last turn in the task. Subtask B) Given A TASK and the reference passages (in gold), generate an answer to the last turn in the task. Subtask C) Given a TASK, retrieve passages and then generate an answer to the last turn in the task. Input provided to participants is shown in blue (and gold), model output that will be evaluated is in white, and intermediate input is in grey.

These tasks will promote research in Multi-Turn RAG, particularly solutions that address challenges such as answerability and later turns as well as other new challenges (to remain hidden) that were not highlighted in the benchmark. Prior RAG tasks, TREC RAG and iKAT, had 45 and 24 submissions respectively, we expect similar participation in our task. This task will also provide additional multi-turn conversations which is a valuable resource which we will release publicly to the community.

2 Data and Resources

We will be providing the MTRAG benchmark (Katsis et al., 2025) and its corresponding corpora

across four domains as trial and development data. All evaluation data for the test phase will come from new unseen conversations targeting areas found challenging in MTRAG (answerability, later turns) as well as new unseen challenges that will remain hidden to participants. The conversations in MTRAG and the new test set are all manually created and reviewed by human annotators to ensure high quality as described in the paper (Katsis et al., 2025).

Given a conversation, a task is a conversation turn containing all previous turns together with the last user question (e.g., the task created for turn k includes all user and agent questions/responses for the first k - 1 turns plus the user question for turn k). An example of a task with two turns is shown in Figure 1. All of our subtasks will be performed at the task level. The test set will consist of approximately 200 tasks, each coming from a different conversation. We will intentionally include tasks that are challenging for SOTA LLMs in the evaluation. We will not evaluate on a full conversation because otherwise a task will reveal answers to a previous task in the conversation. We will release the full conversations, not just the tasks, following the evaluation period.

3 Subtasks

Using the data described in the previous section, we will organize the following three subtasks (See Figure 1): A) Retrieval, B) Generation with reference passages, and C) full RAG - retrieval followed by generation:

- **Subtask A - Retrieval**: Given a task that is answerable and has a set of relevant passages, evaluate whether the retrieval system retrieves the relevant passages.
- **Subtask B - Generation with Reference Passages (Reference)**: Given a task with a set of relevant passages and a target answer, evaluate the predicted answer of the generator system compared to the reference answer.
- **Subtask C - Generation with Retrieved Passages (RAG)**: Given a task with a target answer, first retrieve 5 passages, then generate an answer. Evaluate the predicted answer of the generator system to the reference answer.

Participants can participate in one or multiple subtasks. We encourage creative submissions that use different resources such as new trained models, prompt engineering, query rewriting, and agentic RAG (repeated querying). We will not allow submissions using Mixtral 8x7B as this model was used as the generator during dataset creation.

	Recall		nDCG	
	@5	@10	@5	@10
BM25	0.20	0.27	0.18	0.21
BGE-base 1.5	0.30	0.38	0.27	0.30
Elser	0.49	0.58	0.45	0.49

Table 1: Retrieval Performance baselines of models on our benchmark using Recall and nDCG metrics with the last turn question only. Retrieval performance is only computed on the 777 answerable and partial tasks of MTRAG

4 Evaluation

We will provide the teams with an evaluation script to run on their own on the development set. During the evaluation period, we will provide the teams with the test data. Due to the sequential nature of the task, we will have two phases, a retrieval phase (Subtask A, C), followed by a generation phase (Subtask B). We will only be evaluating one submission per subtask, but participants can keep submitting until the end of the evaluation phase and we will only evaluate their final submission. Once the evaluation phase is complete, we will evaluate each subtask on the same script provided to the participants with the reference passages and answers. The results for the test leaderboard will only become visible after the competition has concluded.

The retrieval subtask will be evaluated using the common retrieval metrics, nDCG and Recall. The two generation subtasks will be evaluated using the three main metrics reported in the paper, (1) $\mathbf{RB}_{alg}$: The harmonic mean of Bert-Recall, Rouge_L, and Bert-K-Prec, (2) $\mathbf{RB}_{llm}$: A Reference-Based LLM judge adapted from RAD-Bench (Kuo et al., 2024) and (3) $\mathbf{RL}_f$: The RAGAS (Es et al., 2024) Faithfulness LLM judge. All metrics will be conditioned on an IDK LLM judge that first determines if the response contains an answer. In addition, due to the reference-less nature of the full RAG task we will have a small human evaluation (approximately 20 tasks) on all participating models. Full details regarding the evaluation metrics are available in the benchmark paper (Katsis et al., 2025).

	RL_F		RB_{llm}		RB_{alg}	
	•	○	•	○	•	○
Reference	0.87	0.65	0.95	0.95	0.88	0.85
GPT-4o	**0.76**	0.71	**0.76**	**0.70**	0.45	0.40
Llama 3.1 405B	0.75	**0.72**	0.74	0.68	**0.48**	**0.42**
Llama 3.1 8B	0.55	0.56	0.59	0.59	0.37	0.35
Qwen 2.5 7B	0.68	0.67	0.66	0.68	0.44	0.39

Table 2: Generation results by retrieval setting on the 842 tasks: Reference (•) and RAG (○), w/ IDK conditioned metrics. Per column, the best result is in **bold** and second best is underlined.

5 Trial Data and Baselines

We will be releasing the MTRAG benchmark (Katsis et al., 2025) as trial/training data along with scripts for evaluation. The benchmark is available at https://github.com/IBM/mt-rag-benchmark. The benchmark contains 110 conversations which is 842 tasks. We provide baselines in Table 1 and Table 2 respectively for retrieval and generation. Even with the best model, the retrieval results are low, showing that there is room for improvement. In the generation results, there is room for improvement compared to the reference answer across both generation tasks. We plan to include in the leaderboard different baselines based on model size (small/large).

6 Task Organizers and Roles

Sara Rosenthal is a Staff Research Scientist at IBM Research in New York. She has considerable experience in building benchmarks and human annotation. She has co-organized eight SemEval tasks: Sentiment Analysis in Twitter, OffensEval, and Table Fact Verification. She is currently an organizer for the SemEval Workshop 2024-2025. She has also served as an Area Chair, SAC, and D&I chair at several *ACL conferences and is an Action Editor for TACL. Sara will be the lead organizer for the task.

Yannis Katsis is a Senior Research Scientist at IBM Research - Almaden with experience in management, integration, and knowledge extraction from both structured and unstructured data. His recent work includes evaluating and improving RAG systems, including the release of the MTRAG benchmark, which will form the basis of this task. Yannis will be a co-organizer of the task with a focus on dataset selection and quality.

Vraj Shah is a Staff Research Scientist at IBM Almaden Research Center with experience in RAG systems, LLM-based evaluation methods, and data management. He has served as program committee member at several data management venues such as VLDB and SIGMOD. Vraj will be a co-organizer of the task with a focus on evaluation metrics and running evaluation.

Marina Danilevsky is a Senior Research Scientist at the IBM Almaden Research Center, and the manager of the Core Language Technologies group. She works on a variety of research projects on language modeling and text understanding within multiple domains, with a particular focus on evaluation, model explainability and human-in-the-loop techniques. She has co-organized a previous SemEval task (Table Fact Verification) and co-led multiple tutorials and online courses. Marina will be the advisory organizer for the task.

7 Ethical Considerations

7.1 Impact

Hallucination is a key challenge associated with LLMs. RAG is an important framework for avoiding hallucination by remaining faithful to the provided context. Improving RAG performance can help avoid the spread of misinformation making it an important task that can have a significant impact for the development of LLMs.

7.2 Data and Annotators

The annotators are highly skilled individuals whose sole job is to perform annotation tasks and they are paid well above minimum wage. All annotator information is anonymized to avoid PII. Further, the questions in the benchmark and test data are general and not specific to any individual. Any mention of information that looks personal is fictitious (e.g. How can I avoid bankruptcy?).

All data will be released under the Apache 2.0 license and all costs associated with the task, including annotation and evaluation resource, will be covered under ongoing work processes at IBM.

References

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.

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.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Preprint*, arXiv:2501.03468.

Tzu-Lin Kuo, Feng-Ting Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. 2024. RAD-Bench: Evaluating large language models capabilities in retrieval augmented dialogues. *Preprint*, arXiv:2409.12558.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnarök: A reusable rag framework and baselines for trec 2024 retrieval-augmented generation track. *Preprint*, arXiv:2406.16828.