Research Article

MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems

Yannis Katsis[1], Sara Rosenthal[1], Kshitij Fadnis[1], Chulaka Gunasekara[1], Young-Suk Lee[1], Lucian Popa[1], Vraj Shah[1], Huaiyu Zhu[1], Danish Contractor[1], Marina Danilevsky[1]

1. IBM Research

Retrieval-augmented generation (RAG) has recently become a very popular task for Large Language Models (LLMs). Evaluating them on *multi-turn* RAG conversations, where the system is asked to generate a response to a question in the context of a preceding conversation is an important and often overlooked task with several additional challenges. We present mtRAG: an end-to-end human-generated multi-turn RAG benchmark that reflects several real-world properties across diverse dimensions for evaluating the full RAG pipeline. mtRAG contains 110 conversations averaging 7.7 turns each across four domains for a total of 842 tasks. We also explore automation paths via synthetic data and LLM-as-a-Judge evaluation. Our human and automatic evaluations show that even state-of-the-art LLM RAG systems struggle on mtRAG. We demonstrate the need for strong retrieval and generation systems that can handle later turns, unanswerable questions, non-standalone questions, and multiple domains. mtRAG is available at https://github.com/ibm/mt-rag-benchmark.

1. Introduction

Large Language Models (LLMs) play an important role as chat-based assistants[1]. Relying on knowledge-sources during the conversation is an important task that helps improve answer reliability and trust, and hence Retrieval-augmented generation (RAG) has become an important and popular field in recent years[2][3]. The primary focus of RAG benchmarks has been on single turn[4][5][6] which LLMs have become proficient at[7], however multi-turn RAG, where a turn is defined as a question-response pair, has been largely overlooked as a benchmark, and presents additional challenges not

 doi.org/10.32388/PHS6YN

covered in single-turn RAG. A multi-turn conversation benchmark should sufficiently cover several challenging aspects to effect a holistic evaluation of the full RAG pipeline:

Retrieval. The relevant passages should change during the conversation causing repeated retrieval.

Generation. The generator should struggle to answer many of the questions correctly, particularly questions that refer to and rely on previous turns.

We present mtRAG, a diverse and representative multi-turn RAG benchmark of human-generated conversations across 4 different domains that vary in style, topic and source. Our conversations comprise turns that vary along the dimensions of *question type, multi-turn, and answerability*.

Our benchmark is constructed using a novel process where human annotators simulate a real-world conversation, by actually interacting with a live RAG agent via a custom chat application and improving the output in real time. Annotators took care to diversify their questions across the aforementioned different dimensions, including referencing earlier turns, while ensuring a flowing and natural conversation. At every turn, after issuing their questions, annotators checked the passages retrieved by the RAG system and modified the passage set to improve relevance and diversity. Next, they reviewed and repaired the generated response to improve its quality. Figure 1 shows part of a conversation from the benchmark to illustrate the output of our data creation process (described in detail in Section 4). The resulting conversations average 7-8 turns in length and 16.9 unique relevant passages per conversation.

doi.org/10.32388/PHS6YN

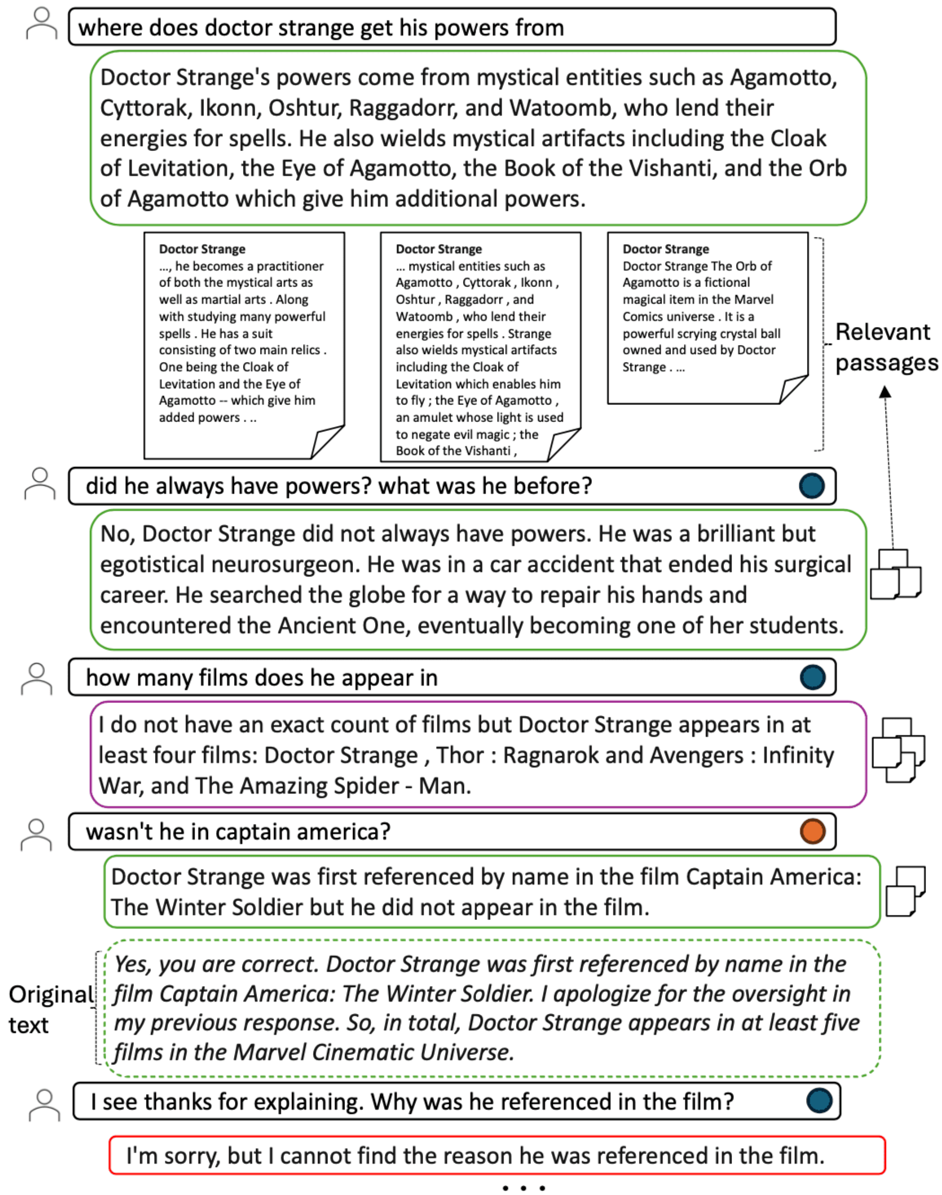


Figure 1. 5/8 turns of a conversation from the CLAPnq domain. The conversation is enriched with question dimensions, passage diversity, and repair. The answerability is shown using the agent response color: answerable, unanswerable, and partial. The multi-turn type is shown using the question circle: follow-up and clarification. The different relevant passages highlight diversity and the original text shows a repair from the model response.

We evaluate our mtRAG benchmark on the retrieval and generation components of RAG systems. We examine retrieval performance of lexical, sparse and dense retrieval under two settings (last turn and

doi.org/10.32388/PHS6YN

query rewrite) and analyze generative performance of 9 LLMs under three retrieval settings (reference, reference+RAG, and full RAG). Our rigorous human evaluation of model responses demonstrates that all models struggle on our tasks, especially on unanswerable questions, and in later turns.

As human data generation and evaluation does not scale well, we systematically explore automation paths. We identify several automated evaluation metrics that correlate with human scores (and many that do not), and demonstrate the need for more work in automatic evaluation. We also construct a companion benchmark, mtRAG-S, of synthetically generated conversations. By providing both human-generated and synthetic conversations over the same corpora, we aim to help the community analyze and understand the relative advantages of the two types of data.

Our contributions are as follows:

1. We present mtRAG, a comprehensive and diverse human-generated multi-turn RAG benchmark, accompanied by four document corpora. The benchmark is available at https://github.com/ibm/mt-rag-benchmark.
2. We evaluate lexical, dense and sparse retrieval models and 9 large language models and conduct a rigorous human evaluation of model responses.
3. We systematically explore automation paths for data and metrics, demonstrating gaps in current automatic evaluation and constructing a companion benchmark, mtRAG-S, of synthetically generated conversations.

To the best of our knowledge, mtRAG is the first end-to-end human-generated multi-turn RAG benchmark that reflects real-world properties of multi-turn conversations.

2. Related Work

Question Answering (QA) and Information Retrieval (IR) have been popular tasks for many years. Prior focus has been on Extractive QA[8][9], Long Form Question Answering[10][11][12], and Open Domain QA[13]. There has also been work on multi-turn conversations, primarily through datasets like MT-Bench[14], which focuses on tasks without retrieval, and Wizard of Wikipedia[15], which is conversational rather than information seeking. Several surveys exist which summarize prior approaches and datasets available to the community[16][17][18][19].

doi.org/10.32388/PHS6YN

Benchmark	Active Retriev.	Long Answer	Unanswerable	Multi-Domain
QuAC[20]	✗	✗	✓	✗
Or-QuAC[21]	✗	✗	✓	✗
CoQA[22]	✗	✗	✓	✗
ShARC[23]	✗	✗	✗	✗
MD2Dial[24]	✗	✗	✗	✗
FaithDial[25]	✓	✗	✓	✗
iKAT[26]	✗	✓	✗	✗
RADBench[7]	✗	✓	✗	✓
mtRAG (Our work)	✓	✓	✓	✓

Table 1. mtRAG compared to prior multi-turn RAG benchmarks

We compare our benchmark to existing multi-turn RAG datasets in Table 1. Except for FaithDial[25], prior work keeps the retrieval component fixed. Retrieval is either performed once at the beginning, restricting the entire conversation to the initial passages; or retrieval is only deployed to find evidence for an existing conversation, which can result in strange mismatches between answers and verifiability. In contrast, we perform *active retrieval*[27], where ongoing passage retrieval influences both follow-up questions and provided answers, more closely reflecting real life scenarios. Excluding RAD-Bench Kuo et al.[7] and iKAT[26], most prior datasets focus on extractive or short answers (1-2 sentences) limiting the kind of questions that can be asked. Further, many existing datasets ignore unanswerable questions - a ripe source of hallucinations in LLMs[28]. Finally, most datasets focus on a single domain/topic, while we explore several domains of different types. Our mtRAG benchmark reflects real-world properties of multi-turn conversations, including active retrieval, long-form answers, unanswerable questions, and multiple domains.

 doi.org/10.32388/PHS6YN

3. mtRAG Benchmark

We now describe mtRAG's characteristics, before delving into how it was created in Section 4.

3.1. Dimensions

To ensure that mtRAG is representative of real RAG use cases, we designed it to be diverse across several important dimensions. For the detailed definitions see Appendix A.

Question types: Similar to contemporaneous work on datasets for RAG[6][29] conversations in mtRAG contain questions of diverse types, including factoid, comparison, explanation, keyword questions, and others. Each question has one or more question type labels.

Multi-turn: In addition to question types that apply to individual questions regardless of the surrounding conversation, there are different types of multi-turn questions. A question can be a follow-up or clarification. All questions in mtRAG beyond the first turn are labeled with one multi-turn type.

Answerability: Models often struggle when confronted with questions or problems that cannot be answered[30][31]. Based on this observation, we include *answerable*, *partially answerable*, *unanswerable* based on the corpora, and *conversational* statements (e.g., "Hi", "That's interesting", "Thank you").

Domain: To test RAG systems over different types of documents, mtRAG is created over four domains and corresponding document corpora (discussed in more detail in Section 4.2).

3.2. Conversation Properties

To ensure that mtRAG is a challenging benchmark, the following properties were incorporated during conversation creation.

High Quality Responses: Following and extending prior work[12][32][26], each reference answer is written to satisfy the following properties: (a) *Faithfulness:* The answer is faithful to the passages or earlier turns, (b) *Appropriateness:* It is appropriate/relevant to the question, (c) *Naturalness:* The answer sounds natural, and (d) *Completeness:* Includes all information in the passages relevant to the question. We refer to these properties as FANC.

Passage Diversity: Conversations contain questions that are diverse enough that the relevant passages do not remain static through the conversation. Our conversations have on average 16.9 unique relevant passages and 20.9 relevant passages in total.

Answer Repair: During conversation creation we employ an LLM to create the initial response, which is then repaired by the annotator as needed. A response that does not require repair can be considered an indication that the question is not challenging for the LLM. Our conversations contain repairs on 92% of the turns.

Non-standalone: The questions occasionally rely on prior turns in the conversation (e.g., by using co-references to prior questions or answers or by clarifying previous turns). On average 1.3 questions per conversation include co-references.

4. Benchmark Creation

4.1. Annotators

The annotators that contributed to this work are highly skilled individuals hired solely to perform language annotation tasks and paid well above minimum wage. Unless otherwise noted, the annotators were used for all annotation tasks in this paper. Great care was taken to ensure random assignment.

4.2. Document Corpora

The first step in creating the benchmark was to assemble the document corpora over which the conversations would be built. mtRAG consists of four document corpora/domains:

- **CLAPnq**[12]: a subset of Wikipedia pages,
- **FiQA**[33]: a set of StackExchange posts discussing financial advice,
- **Govt**: the crawled contents of select web-pages under the .gov and .mil domains, and
- **Cloud**: the crawled contents of select technical documentation pages of a major cloud provider.

CLAPnq and FiQA are existing corpora from QA/IR datasets, while Govt and Cloud are new corpora assembled specifically for this benchmark. To ensure that the new corpora are well suited for generating diverse conversations (i.e., conversations that touch several different passages), both the Govt and Cloud corpora were designed to contain sets of inter-connected pages (see Appendix B.1).

Each corpus was indexed using Elser from ElasticSearch.[1] During ingestion, documents were split into passages of 512 tokens with an overlap stride of 100 tokens. The indexes are used to perform retrieval

doi.org/10.32388/PHS6YN

during conversation creation and all RAG experiments. Table 2 summarizes the corpora.

Corpus	Domain	Provenance	Documents (D)	Passages (P)	Avg P/D
CLAPnq	Wikipedia	Rosenthal et al.[12]	4,293	183,408	42.7
FiQA	Finance	Maia et al.[33]	57,638*	61,022	1.1
Govt	Government	New corpus	7,661	49,607	6.5
Cloud	Technical documentation	New corpus	8,578	72,442	8.4

Table 2. Statistics of document corpora included in the mtRAG benchmark. (*) For FiQA we report the number of individual forum posts, as the dataset does not have the notion of a document

Finally, for each corpus we also assembled a set of seed questions to help human annotators bootstrap the conversation generation process, described in Section 4.3. For CLAPnq and FiQA we leveraged questions from the corresponding QA datasets, while for Govt and Cloud we selected a set of seed documents from each corpus and asked annotators to write seed questions based on them.

4.3. Human-Generated Conversations

Annotators were asked to create multi-turn conversations over the four corpora. To aid in this process, we developed a custom chat application allowing them to interact with a live RAG agent consisting of an ELSERv1 (ElasticSearch 8.10)[1] retriever and Mixtral 8X7b Instruct Jiang et al. [34] generator and correcting the retriever and generator outputs as needed. Using the application, annotators created multi-turn conversations by performing the following actions at every turn: write a question, adjust the set of retrieved passages, edit agent response and enrich with question dimensions. Annotators were directed to write 6-9 (or more) turns per conversation.

1. **Training:** Prior to the main creation task we provided annotators with interactive training sessions, extensive documentation with examples, and pilot tasks with feedback. Once they were comfortable with the task and producing conversations that met our criteria, we began the full annotation task.

2. **Creation:** Annotators started a conversation using a seed question. For subsequent turns, they were encouraged to create questions that naturally extended the preceding conversation while varying in answerability types, question types, and multi-turn patterns (Section 3.1), which were noted during annotation time. Once annotators wrote a question, the chat application queried the retriever for potentially relevant passages, ran the generator to produce an agent response based on the retrieved passages, and presented everything to the user for repair. Annotators were guided to ensure passage diversity and answer quality (Section 3.2) using the application's custom passage search and edit agent response functionalities. In the final benchmark, the similarity between the original and repaired response based on Rouge-L (denoted as the *edit score*) is 60.7 indicating significant amount of repair. This novel process simulates real-time conversations, which is missing in prior work (Section 2).
3. **Review:** The resulting 126 conversations were then reviewed. Annotators could accept or reject conversations, and repair responses, passage relevance, and dimensions as needed. They were not allowed to edit the questions or passages as such changes could negatively affect the conversation flow. During this phase, most conversations were kept (see review process details in Appendix B.3). Our human evaluation (Section 7) shows that the reference responses are preferred by humans.

4.4. Data Statistics

This process yielded a benchmark of 110 conversations (29 ClapNQ, 27 FiQA, 28 Govt, 26 Cloud) with an average 7.7 turns per conversation, leading to 842 tasks. 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). All evaluations described next are performed at the task level. Figure 2 shows the distribution of tasks in mtRAG on the benchmark dimensions.

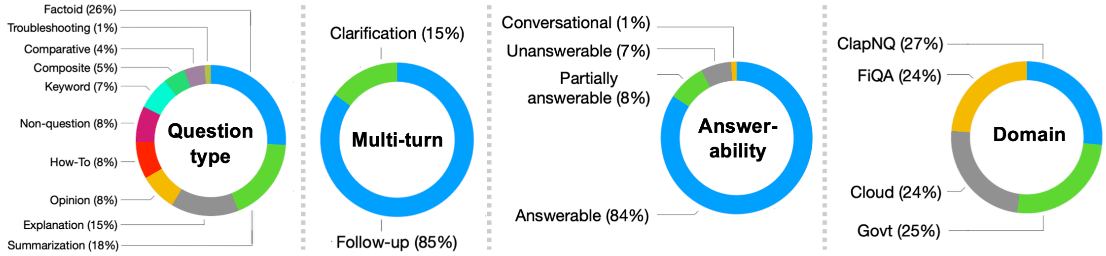


Figure 2. Distribution of tasks in mtRAG based on each of the benchmark's dimensions.

5. Retrieval

We use mtRAG to evaluate both the retrieval and generation components of RAG systems. Retrieval evaluation is described in this section, while generator evaluation is presented in the next section. Note, that this is not intended to be a comprehensive evaluation of all retrievers and generators. Instead our goal is to demonstrate the challenging nature of mtRAG and how it can help evaluate and surface issues in state-of-the-art RAG systems.

5.1. Experimental Setup

We use mtRAG with the indexed corpora from the creation stage (Section 4.2) to evaluate lexical (BM25), dense (BGE-base 1.5[35]) and sparse (Elser) retrieval. The competitiveness of these models is shown on the MTEB leaderboard.[2]

We report the commonly used Recall and nDCG metrics[36] @1, 3, 5, 10. Since we use Elser for retrieval during data creation, there may be some biases towards Elser.

The retrieval task is performed on the reference passages (i.e., those marked as relevant during the creation process). It is therefore only computed on answerable and partially answerable tasks where reference passages exist.

5.2. Retrieval strategies

We experimented with several strategies to query the retriever for relevant passages, including sending the full conversation up to the current user turn, all the user turns without the responses, subsets of the conversation, and only the last user turn. Using the full conversation, or even just a few turns from it consistently under-performed, often causing the retriever to bring back the same passages over and over. The most effective strategy was just using the last user turn. A particular challenge in multi-turn conversations is that a user turn may be non-standalone, employing shortcuts to express intent, or referencing entities or concepts from earlier in the conversation. To alleviate this we used a *query rewrite* strategy, also known as contextual query rewriting[37][38] to rewrite the user turn using an LLM so that it incorporates all necessary parts from the context into an unambiguous, standalone question (see Appendix C.1 for implementation details). An example of query rewrite is shown below:

> User: Who is the CEO of Apple Inc.?
> Agent: The CEO of Apple Inc. is TIM COOK.
> User: its address?
>
> **[Rewriting] What is the address of Apple Inc?**

5.3. Retrieval Results

We highlight the results in Tables 3 and 4. Table 3 shows that the query rewriting strategy consistently outperforms using only the last turn (without rewriting), across all metrics, for all models. Elser outperforms BM25 and BGE-base 1.5. Table 4 shows the Elser results with query rewriting broken down by domain, first turn vs later turn, and whether the question is standalone. Retrieval performance is significantly lower for later turns than for the first turn, and non-standalone questions continue to pose a challenge (though query rewriting helps). These results highlight two key areas of improvement for retrieval components: 1) multi-turn retrieval, and 2) non-standalone questions.

 doi.org/10.32388/PHS6YN

		Recall				nDCG			
		@1	@3	@5	@10	@1	@3	@5	@10
BM25	Last Turn	0.08	0.15	0.20	0.27	0.17	0.16	0.18	0.21
	Query Rewrite	**0.09**	**0.18**	**0.25**	**0.33**	**0.20**	**0.19**	**0.22**	**0.25**
BGE-base 1.5	Last Turn	0.13	0.24	0.30	0.38	0.26	0.25	0.27	0.30
	Query Rewrite	**0.17**	**0.30**	**0.37**	**0.47**	**0.34**	**0.31**	**0.34**	**0.38**
Elser	Last Turn	0.18	0.39	0.49	0.58	0.42	0.41	0.45	0.49
	Query Rewrite	**0.20**	**0.43**	**0.52**	**0.64**	**0.46**	**0.45**	**0.48**	**0.54**

Table 3. Retrieval Performance of models on our benchmark using Recall and nDCG metrics

	Subset	R@5
By Turn	Turn 1 (102)	0.89
	> Turn 1 (675)	0.47
By Standalone	Standalone (555)	0.48
(> Turn 1)	Non-Standalone (120)	0.42
By Domain	CLAPnq (208)	0.56
	FiQA (180)	0.50
	Govt (201)	0.56
	Cloud (188)	0.47

Table 4. Elser retrieval results with query rewrite on subsets of the data to highlight multi-turn properties. Numbers in parentheses denote size of each subset.

6. Generation

We next present the generator experiments. We start with the experimental setup, followed by the results, using automated metrics including LLM judges. Section 7 will complement this with a human evaluation on a subset of mtRAG. Given a task, we send to the model the following information: the question, preceding turns, N passages, and instructions. We choose N=5 passages because it achieves considerable improvement compared to top 3, while remaining a manageable amount of passages (Section 5). For more generation format details, see Appendix D.2.

6.1. Retrieval Settings

We evaluate how LLMs perform under three retrieval settings, simulating ideal/noisy retrieval.

Reference (●): Generation using reference passages or no passages if unanswerable/conversational. No retrieval is performed in this setting; it simulates a perfect retriever.

Reference + RAG (◐): Partial retrieval followed by generation, where the reference passages are supplemented by the top retrieved passages (using Elser with rewrite) to yield a total of 5 passages. We restrict this to the 426 tasks that have at most two reference passages to ensure all passages needed for the reference are included. This can be considered an upper bound, where the retrieval is successful but there are additional noisy passages.

Full RAG (◦): Retrieval using Elser with rewrite followed by generation, where the top N=5 passages are retrieved (the standard RAG setting).

doi.org/10.32388/PHS6YN

	Ans. Acc.			RL_F			RB_{llm}			RB_{alg}		
	•	◐	◦	•	◐	◦	•	◐	◦	•	◐	◦
Reference	0.98	0.97	0.98	0.86	0.87	0.67	0.94	0.94	0.94	0.88	0.88	0.86
Command-R+ (104B)	0.86	0.86	**0.87**	**0.69**	**0.71**	**0.66**	0.66	0.62	0.59	0.43	0.40	0.38
GPT-4o	**0.89**	0.86	0.86	**0.69**	0.69	0.65	**0.73**	0.68	**0.66**	0.46	0.40	0.38
GPT-4o-mini	0.87	0.86	0.84	0.66	0.69	0.64	0.72	0.68	0.64	0.43	0.40	0.37
Llama 3.1 405B Instruct	0.87	0.86	0.85	**0.69**	0.70	0.65	0.70	0.68	0.63	**0.47**	**0.42**	**0.39**
Llama 3.1 70B Instruct	0.78	0.83	0.81	0.63	0.66	0.64	0.62	0.64	0.59	0.43	**0.42**	**0.39**
Llama 3.1 8B Instruct	0.71	0.75	0.74	0.50	0.51	0.53	0.54	0.54	0.52	0.36	0.33	0.34
Mixtral 8x22B Instruct	0.86	**0.87**	0.86	0.54	0.61	0.56	0.66	0.64	0.61	0.39	0.38	0.35
Qwen 2.5 (72B)	0.87	**0.87**	**0.87**	0.65	**0.71**	0.64	0.71	**0.69**	0.65	0.43	0.40	0.37
Qwen 2.5 (7B)	0.88	0.86	**0.87**	0.62	0.66	0.62	0.68	0.65	0.63	0.42	0.38	0.37

Table 5. Generation results by retrieval setting: Reference (•), Reference+RAG (◐), and RAG (◦), w/ IDK conditioned metrics ($n = 426$). Per column, the best result is in **bold** and second best is underlined.

6.2 Models

We evaluate the following auto-regressive models.

Llama 3.1 Models[39]: The Llama 3.1 family of models are instruction-tuned models that support up to 128K tokens. We evaluate the 8B, 70B and 405B models.

Mixtral Mixture-of-Expert[34]: The instruction fine-tuned Mixtral 8x22B model that supports up to 32K tokens.[3]

GPT-4o Models[4]: We use the GPT-4o and GPT-4o-mini model in our experiments. These support context lengths of up to 128K tokens.

Command R+[5]: This is a 104B parameter multi-lingual model optimized for RAG and tool use.

doi.org/10.32388/PHS6YN

Qwen 2.5 models Team[40]: We use the instruct versions of the 7B and 72B models. They support a context length of up to 128K tokens.

6.3 Metrics

We use three metrics, described below, to evaluate the quality of RAG systems and understand if a model response exhibits the desirable FANC properties outlined in Section 3.2. We use these metrics to provide insights into the trends of LLMs on multi-turn RAG rather than declare winners. See Section 8 for a discussion on how we picked the metrics and the challenges associated with finding good evaluation metrics for multi-turn RAG. Moreover, for implementation details, see Appendix F.

6.3.1 Aggregate performance metrics

The first two metrics, RB_{alg} and RB_{llm}, are both reference-based (RB) metrics. They both measure the aggregate performance of model responses by comparing the model response (MR) to the reference answer (RA) utilizing different techniques:

$\mathbf{RB_{alg}}$ is the harmonic mean of three algorithmic metrics[41]: Bert-Recall (*Bert-Rec*), Bert-K-Precision (*Bert-K-Prec*), and *Rouge-L*. The intuition is as follows: Bert-Rec is an approximation for completeness as it measures the semantic overlap between MR and RA, tending to prefer longer answers. Bert-K-Prec compares MR to the passages, P, and is an approximation for faithfulness and completeness. Rouge-L measures whether phrases from RA are in MR and is an approximation for appropriateness.

$\mathbf{RB_{llm}}$ is an LLM judge inspired by RAD-Bench[7]. We adapt RAD-Bench's approach of comparing MR to RA but modify the prompt to add P and anchor the evaluation on the metrics of faithfulness, appropriateness, and completeness. To minimize model biases and improve evaluation reliability, we use several models as judges and use the median as the final score.

6.3.2. Faithfulness metric

We also use a metric specifically for faithfulness, which is important for RAG applications and a challenge for LLMs (see Section 7). $\mathbf{RL_F}$, the Faithfulness LLM judge from RAGAS[32], appears to be a good judge for faithfulness (see Section 8). In contrast to the other two metrics, this is a reference-less (RL) metric, as it does not rely on the reference answer.

6.3.3. Conditioning metrics on answerability

Prior to computing the metrics we employ an IDK ("I Don't Know") judge to detect whether the response has a full or partial answer. It is important to first determine whether a response is IDK because intuitively, words used to indicate not knowing the answer may not match the context; this is also reflected in the metrics, which were not designed to measure IDK correctly. Our IDK judge achieves an accuracy of over 97% (see Appendix F.3). We condition the metric score, $\phi \in \{\mathrm{RB}_{\mathrm{alg}}, \mathrm{RB}_{\mathrm{llm}}, \mathrm{RL}_{\mathrm{F}}\}$, on answerability value A and IDK value IDK as follows:

	IDK **= no, partial**	$IDK =$ **yes**
$A =$ yes, partial	ϕ	0
$A =$ no	0	1

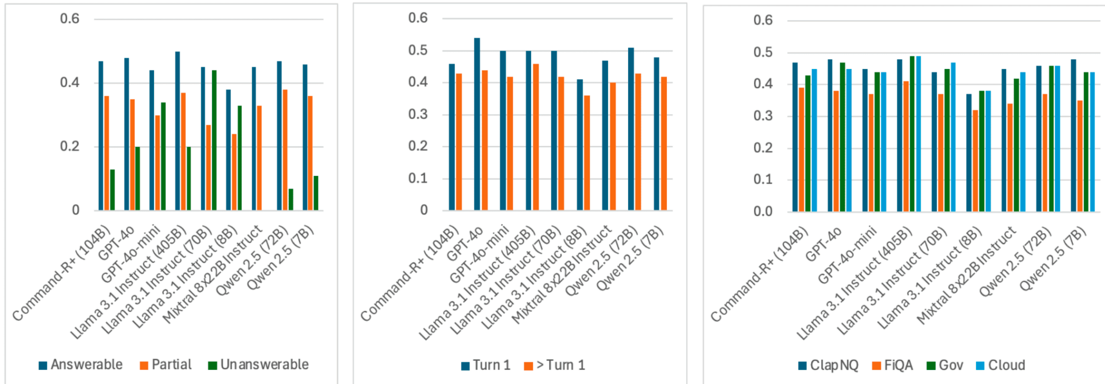
We define *answerability accuracy* (*Ans. Acc.*) of a model as the accuracy of IDK correctly predicting A. Finally, conversational questions also require special handling for evaluation. Since only 10 tasks fall in this category, we exclude them from the experiments, leaving their study as future work.

6.4. Generation: Evaluation Results

The overall results across the different retrieval settings are shown in Table 5[6]. Within model families, the larger model does as well or better than its smaller counterparts across metrics. In general, GPT-4o and Llama 3.1 405B Instruct perform the best across metrics and settings. All models score significantly lower than the reference answer, indicating there is still room for improvement in mtRAG for all LLMs. Comparing performance across retrieval settings, we see that the results degrade as the setting gets more challenging: $\bullet > ◐ > \circ$, indicating the noise has an impact on generation. Interestingly, Qwen 2.5 72B and Command-R+ are more competitive and closer to GPT-4o and Llama models in noisy settings.

Focusing next on the Reference ($\bullet$) retrieval setting, we further explore how LLMs perform on the different dimensions of mtRAG. Figure 3 shows the results. For space reasons we only report the $\mathrm{RB}_{\mathrm{alg}}$ metric. We also leave out the breakdown by question types and multi-turn types, where we did not find interesting patterns. The additional metrics and breakdowns can be found in Appendix I.

 doi.org/10.32388/PHS6YN



(a) By question answerability (b) By first vs subsequent turns (c) By domain

Figure 3. Generation results in the Reference (●) retrieval setting using a single metric, $\mathrm{RB}_{\mathrm{alg}}$, on three different dimensions: (a) answerability, (b) turns, and (c) domains

Answerable vs unanswerable questions. As shown in Figure 3(a), model performance generally drops when questions become partially answerable or unanswerable. In particular, models experience a dramatic drop in performance on unanswerables, struggling to declare they do not know the answer when the answer is not included in the input passages. Interestingly, performance on unanswerables differs widely between model families. While GPT-4o and Llama 405B score low for unanswerables, they still perform much better than models in other families. It is interesting to note that Llama 70B and 8B perform better on the unanswerables, because they say “I don’t know” too often. This can also be seen by their low answerability accuracy in Table 5.

First turn vs subsequent turns. As shown in Figure 3(b), in almost all cases the models perform better on first turn vs subsequent turn questions. This proves our conjecture that answering questions in a multi-turn setting is more challenging, as a model has to interpret a question in the context of the preceding conversation, which was one of the main motivations for this work.

Domains. Figure 3(c) shows that model performance is similar across domains except for FiQA where the results tend to be lower. We suspect this is due to the nature of the corpus; it contains posts from a financial discussion forum, which are typically short, very informal in style, and often subjective.

 doi.org/10.32388/PHS6YN

7. Human Evaluation

We also performed a human evaluation on a subset of the benchmark. This serves several purposes: 1) Verifying that the reference responses are of high quality, 2) correlating our metrics with human judgment, and 3) analyzing frontier models.

We perform the human evaluation on two frontier models: GPT-4o and Llama 3.1 405B Instruct and compare them to the reference answers. We select 5 conversations per domain for a total of 159 evaluation tasks. We ask annotators to measure the quality across the desirable response properties, FANC (Section 3.2), on a scale of 1 (Low) - 4 (High). We also ask them to perform a pair-wise comparison of the models, from which we calculate the *Win-Rate (WR)*. Ties are allowed but we ask them to do so sparingly.

The annotators completed the task using the Appen platform.[7] Each evaluation task comprised the conversation thus far up to the last user question, the relevant passages, and the corresponding model responses (anonymized and in random order). Each task was done by three skilled annotators. The annotator agreement was very high: {F: 89.6, A: 92.1, N: 95.6, C: 90.3, WR: 87.4} (see Appendix E).

Table 6 shows the evaluation results using median to consolidate the scores of the three annotators on the answerable/partial subset (152 tasks) of the human evaluation. The reference answer is exceedingly preferred by annotators over the model responses, as evidenced by the win-rate. It also exhibits the highest score for most individual properties. This highlights the quality of the human-generated reference answers and shows that even frontier LLMs still have room for improvement. The results show that LLMs do a good job of providing natural and appropriate answers; but they still struggle with faithfulness and completeness, where they receive lower scores. Both frontier models are equally preferred for answerable questions.

doi.org/10.32388/PHS6YN

	Human Evaluation						Metrics		
	WR	**F**	**A**	**N**	**C**	**All**	RL_F	RB_{llm}	RB_{alg}
Ref.	**59.3**	**3.8**	3.8	**4.0**	**3.9**	**3.8**	**0.86**	**0.97**	**0.87**
GPT-4o	47.8	3.5	3.8	**4.0**	3.7	3.6	0.80	0.80	0.46
Llama3.1	47.4	3.5	**3.9**	3.9	3.7	3.6	0.78	0.78	0.49

Table 6. Results on the human evaluation for win-rate (WR), **F**aithfulness, **A**ppropriateness, **N**aturalness, **C**ompleteness and All (harmonic mean of FANC) compared to the IDK conditioned metrics. All results are reported on the Answerable subset.

We also explore the few unanswerable questions, and observe that Llama 3.1 405B Instruct is the least preferred by win-rate. Llama still provides answers (hallucinations) to unanswerable questions.

8. Automatic Evaluation

	RB_{alg}	RB_{llm}	RL_F	RL_R	RL_{MTB}
WR	0.24	0.33	0.01	-0.03	-0.17

Table 7. Weighted Win-Rate Spearman correlation with reference-based (RB) and reference-less (RL) metrics

Human evaluation is not feasible as a long term solution for evaluating models as it does not scale easily. We explore reference-less and reference-based automated evaluation via algorithmic metrics and LLM judges. Table 7 shows the correlation of these metrics with the human evaluation's win-rate and Figure 4 shows the correlation of metrics that correlate positively with win-rate for the individual properties of FANC. We employed a weighted correlation on human judgements to correct for data

doi.org/10.32388/PHS6YN

imbalance. We exclude the reference responses from Figure 4 because by design the ranking of the reference is 1 for the Reference-Based metrics. These findings drive our decision to report RB_{alg}, RB_{llm}, and RL_F as our main metrics in Section 6.

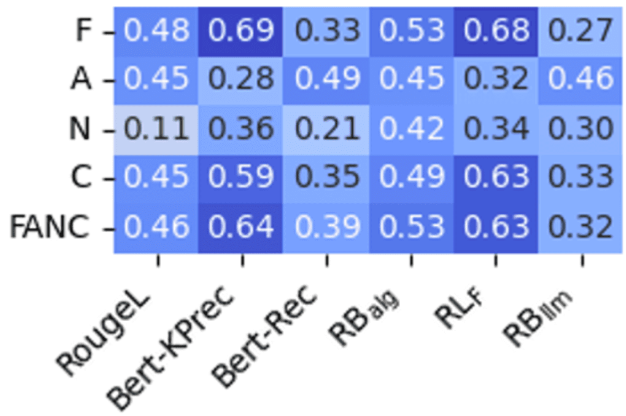


Figure 4. Weighted Spearman correlation of human evaluation with the automated metrics on the answerable subset for the GPT-4o and Llama 3.1 405B Inst. models.

Reference-Based Metrics. We consider an adapted version of RAD-Bench[7], denoted as RB_{llm}, as well as the harmonic mean of algorithmic metrics, denoted as RB_{alg}, as described in Section 6.3. As shown in Table 7, these metrics correlate reasonably well with win-rate; in particular RB_{llm}. Figure 4 shows that they also tend to correlate well with our desired properties; in particular RB_{alg}.

Reference-Less Metrics. We investigate popular metrics in literature: the Faithfulness (RL_F) and Answer Relevance (RL_R) metrics from RAGAS[32], an adapted version of MT-Bench (RL_{MTB} as described in Appendix F.2)[14], as well as Bert-K-Prec. RAGAS Answer Relevance (RL_R), evaluates the model response by asking the LLM to determine the question associated with the response. This question is then compared with the actual question. This is not suitable in a multi-turn setup where the question is often non-standalone. Similarly, RL_{MTB}, while adapted to fit multi-turn conversations, has a very low correlation with win-rate, favoring machine-generated text. In contrast, we find that the two metrics focusing on faithfulness, Bert-K-Prec and RAGAS Faithfulness (RL_F), correlate well with the human faithfulness scores and are therefore useful in settings where no

doi.org/10.32388/PHS6YN

reference is available, although it is clear that evaluating only the faithfulness aspect of a response is incomplete.

9. Synthetic Conversations

Manually creating data is an expensive and time-consuming process that does not scale well. Automating this process has become popular via synthetic data generation[42] and can serve as useful evaluation. To explore this direction, we construct a companion benchmark, mtRAG-S of synthetically-generated conversations. We extend the recently proposed framework of Lee et al.[29] to automatically generate multi-turn conversations (see Appendix G.1 for details). To ensure that the two benchmarks are comparable, we utilize the same corpora, question types, multi-turn patterns, and answerability types (Section 3.1).

In Table 8, we see that synthetic conversations are typically shorter (averaging 5.9 vs 7.7 turns) and exhibit a lower passage diversity (4.6 vs 16.9 unique passages per conversation). Question and response lengths also differ with synthetic data having longer (potentially more detailed) questions but shorter answers. We also found that mimicking several important characteristics posed challenges: our attempts to synthetically generate unanswerable questions were not very successful as the model would often create questions with at least a partial answer. Moreover, increasing the number of turns tended to lead to repetitive user questions and a higher likelihood of hallucinated agent responses.

	mtRAG	mtRAG-S
Avg # turns per conversation	7.7	5.9
Avg # unique passages per conv.	16.9	4.6
Avg # edited responses per conv.	7.3	-
Avg # of question words	8.6	13.7
Avg # of response words *	97.2	72.4
Avg # of relevant passages *	2.8	4.1
Total # of Conversations	110	200
Total # of Tasks	842	1,181

Table 8. Comparing human-generated mtRAG and synthetically-generated mtRAG-S. (Note: Properties with * computed on answerable+partial subset)

Since the aggregate metrics (RB_{alg} and RB_{llm}) rely on a reference answer, which does not exist for synthetic data, we employ the reference-less RL_F and Bert-K-Prec to evaluate faithfulness on mtRAG-S. Using either metric, we see that models across the board receive a higher faithfulness score on the synthetic than on the human-generated data (see Appendix G.2). There are multiple potential explanations for this, ranging from potential idiosyncrasies of synthetic data generation approaches, to the reliability of the presence of desired characteristics, to the quality and biases of automatic evaluation metrics. More work is needed to compare human-generated and synthetic data and we hope that our companion synthetic benchmark serves as a valuable asset towards that goal.

10. Conclusions and Future Work

We present mtRAG, a comprehensive and diverse benchmark of 110 multi-turn human-generated conversations averaging 7.7 turns for a total of 842 tasks[8]. These tasks are used to test the full RAG pipeline. mtRAG is the first end-to-end human-generated multi-turn RAG benchmark that reflects real-world properties of multi-turn conversations. Our experimental results, employing both automated metrics and a human evaluation, highlight the quality of our benchmark and outline several trends and challenges related to multi-turn RAG systems that state-of-the-art retrievers and LLMs face during retrieval and generation. Our findings encourage future research on improving retrieval and generation performance, especially in longer multi-turn conversations, unanswerable questions, and non-standalone user questions. In addition, the ability to scale indicates a clear need for i) more accurate reference-less automatic evaluation metrics, which align more closely with human judgement and can better differentiate model performance; and ii) synthetic data to obtain more conversations. We are also motivated to extend mtRAG in the future to include adversarial turns, additional domains and multilingual conversations.

Appendix A. Dimension definitions

This section provides a detailed list of the mtRAG's dimensions, introduced in Section 3.1. In particular, Tables 9, 10, and 11 show the definitions of the question types, multi-turn types, and answerability types included in the benchmark.

doi.org/10.32388/PHS6YN

Question type	Definition	Example questions
Comparative	Asking for comparison. This can be comparison (a) of multiple entities/concepts, (b) of characteristics of a single entity, or (c) comparison with decision.	(a) “What’s the difference between effective and marginal tax”, (b) “pros and cons of credit cards”, (c) “is X better than Y”
Composite	Comprises several questions. They may be related or dependent.	“Am I eligible for a driver’s license and how do I apply?”
Explanation	Explain the reason behind something.	“Why do I have to do/have X...?”
Factoid	Asking for a specific piece of information, such as a date, quantity, name, yes/no answer, or other singular fact. It can be answered directly and concisely, and does not require an explanation, opinion, interpretation, or subjective judgment to answer.	“Am I eligible for a driver’s license?” , “What is the link to the application portal?”
How-To	Instructions describing how to perform a task.	“How should I do X ...” , “How do I apply for social security disability benefits?” , “I need a license. What do I need to do?”
Keyword	Asking using keywords (not full sentence/phrase). This may be ambiguous.	“vacation days”, “ios 17 upgrade”
Non-question	Not asking a question but instead answering a question or providing information asked by the model.	“I am in Sacramento”, in response to the model saying: "The procedure to file a restraining order depends on the type of restraining order and the court where you want to file it in."
Opinion	Asking model’s opinion on something. The question could also be phrased as leading.	“Don’t you think iphone is better than samsung?”, “Which color car is the best?”
Summarization	Asking to summarize a process or a policy.	“What’s the policy on vacation days?”
Troubleshooting	Finding solutions to issues, problems, challenges.	“I have the error X... what should I do?”

Table 9. Definitions of question types

Multi-turn type	Definition	Example
Follow-up	Ask a question that requests more information or related information to continue the conversation.	User: Can you tell me about the responsibilities of SSA? Agent: Social Security Administration (SSA) assigns Social Security numbers and runs Social Security retirement and disability insurance programs. *User [Follow-up]: I have forgotten my social security number. Should I contact SSA regarding that?*
Clarification	Clarify user's intent or model's answer: (a) Write a statement clarifying the user's intent (typically used when the agent misinterpreted one of the prior questions). (b) Ask a question to clarify the model's previous answer. Clarifications are typically asked when something is unclear or hard to understand.	(a) Clarification of user's intent: User: graphql Agent: GraphQL is an open-source data query and manipulation language for APIs and a query runtime engine. *User [Clarification]: No, I meant, how do I set it up.* (b) Clarification of model's answer: User: Can you tell me about the responsibilities of SSA? Agent: Social Security Administration (SSA) assigns Social Security numbers and runs Social Security retirement and disability insurance programs. *User [Clarification]: Can you explain what you meant by "Social Security retirement and disability insurance programs"?*

Table 10. Definitions of multi-turn types

Answerability type	Definition
Answerable	The question can be fully answered from the passages.
Partially answerable	Only part of the question can be answered from the passages.
Unanswerable	The question cannot be answered neither fully nor partially from the passages.
Conversational	The user turn does not contain a question but is a conversational statement (e.g., "Hello", "Hi, I had a question", "Cool", "That's interesting", "That was all", "Thank you")

Table 11. Definitions of answerability types

Appendix B. Benchmark creation details

B.1. Creating the Govt and Cloud corpora

The Cloud corpus was created by crawling the public technical documentation of a major cloud provider, thus containing inter-connected pages covering various aspects of each cloud offering.

Similarly, the Govt corpus was created starting from 100 seed web pages from the .gov and .mil domains, selected to be suitable to write conversations on (e.g., excluding very short pages or navigation pages) and diverse in terms of topics (e.g., covering parks, NASA, Veteran Affairs, city web-sites, DMV, etc.). For each seed page, we crawled web pages appearing in its neighborhood and thus expected to cover related topics. For each neighborhood the top 150 pages were selected based on their page rank score and additionally filtered to remove duplicates and other low-quality pages, leading to the final set of 7,661 pages/documents.

B.2. Creating conversations

The following instructions were given to the annotators for creating conversations:

doi.org/10.32388/PHS6YN

- Given a seed question, continue a conversation between you and the agent. Continue the conversation to generate a total of 6-9 turns (turn = question + response).
- Aim to have around 25% unanswerable questions. There should be no more than 2 unanswerable questions in one conversation. Most conversations should be completely answerable.
- We encourage diversity of questions across types.
- The questions in your conversation should be connected, but diverse enough to include different passages. All questions shouldn't have the same relevant passages.
- The agent is a bot and the answers may be wrong. In addition to creating questions, you will also edit the answers if they are wrong. Answer correctness is based on the passages provided with the answer. Good conversations will have questions that require edits to repair the conversation.

B.3. Reviewing conversations

As described in Section 4.3, the conversations initially generated by annotators went through a round of review before being finalized. During the review process, 16 out of the 126 conversations were rejected due to lack of turn-wise and topic-level coherency as well as passage diversity, leading to 110 accepted conversations. Additionally, during review, 264 responses (31.3%) were repaired, leading to a slight increase on the number of repaired responses in the final benchmark (from 778 repaired responses before review to 799 responses after review, respectively). The average *edit score* on these responses also increased from 59.5 to 69.2.

Appendix C. Retrieval experiment details

C.1. Query rewriting

For the retrieval experiments, we implemented query rewriting by sending the prompt of Figure 5 to Mixtral 8x7B Instruct. This is only a reference implementation, and better ones could be obtained with larger models (e.g., Mixtral 8x22B Instruct, Llama 3.1 405B, etc.), few-shot prompts with in-context-learning examples, or via fine-tuning. However, this implementation is sufficient to make the point in

doi.org/10.32388/PHS6YN

Section 5 that query rewriting is both necessary and effective in mitigating the challenges of non-standalone questions in multi-turn data.

```
Given the following conversation, please reword the final utterance from the user into
a single utterance that does not need the history to understand the user's intent.
Output in proper JSON format indicating the "class" (standalone or non-standalone)
and the "reworded version" of the last utterance. Use this format: {"class": "type of
last utterance", "reworded version": "the last utterance rewritten into a standalone
question, IF NEEDED"}.

In your rewording of the last utterance, do not do any unnecessary rephrasing or
introduction of new terms or concepts that were not mentioned in the prior part of the
conversation. Be minimal, by staying as close as possible to the shape and meaning of
the last user utterance. If the last user utterance is already clear and standalone, the
reworded version should be THE SAME as the last user utterance, and the class
should be 'standalone'.

{{conversation}}
ASSISTANT:
```

Figure 5. Query rewrite prompt

Appendix D. Generation experiment details

D.1. Model choices

To pick the models used in the generation experiments, we selected a suite of state-of-the-art models of varying sizes. It is not an exhaustive list, but shows the common trends in frontier models. For OpenAI models, we also experimented with the latest models in the o1[9] family. However, initial experiments showed that o1 did not perform as well as GPT-4o. We suspect that this is due to o1's focus on reasoning and multimodal rather than on RAG.

D.2. Model invocation

To evaluate LLMs on mtRAG, for each task we sent to the LLM the following information: question, preceding turns, passages, and instruction. This was composed into the prompt below, adapted to different models using HuggingFace's ChatTemplate[10]. In all cases the prompt fit within the model's context length without need for truncation:

doi.org/10.32388/PHS6YN

Given one or more documents and a user query, generate a response to the query 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”
PASSAGE 1
...
PASSAGE M
User turn 1
Agent Turn 1
...
User Turn N

D.3. Example output

Figure 12 shows an example of an mtRAG task using the Reference (●) retrieval setting together with the Reference response as well as responses from several LLMs.

doi.org/10.32388/PHS6YN

Conversation
User: where does doctor strange get his powers from
Agent: Doctor Strange's powers come from mystical entities such as Agamotto, Cyttorak, Ikonn, Oshtur, Raggadorr, and Watoomb, who lend their energies for spells. He also wields mystical artifacts including the Cloak of Levitation, the Eye of Agamotto, the Book of the Vishanti, and the Orb of Agamotto which give him additional powers.
User: did he always have powers? what was he before?
Agent: No, Doctor Strange did not always have powers. He was a brilliant but egotistical neurosurgeon. He was in a car accident that ended his surgical career. He searched the globe for a way to repair his hands and encountered the Ancient One, eventually becoming one of her students.
User: how many films does he appear in
Passages
Passage 1
The character appears in the Marvel Cinematic Universe. He is first referenced as merely " Stephen Strange " in dialogue by Agent Sitwell in the 2014 film Captain America: The Winter Soldier , and first appears in the 2016 film Doctor Strange , in which he is portrayed by Benedict Cumberbatch . He reprises the role in the films Thor: Ragnarok and Avengers: Infinity War , and will appear again in the untitled sequel to Avengers: Infinity War. Doctor Strange appears in the 2016 animated film Hulk: Where Monsters Dwell, voiced by Liam O'Brien.
Passage 2
In 2008 , Doctor Strange was ranked 83rd in Wizard 's " 200 Greatest Comic Book Characters of All Time " list , and in 2012 was ranked 33rd in IGN 's list of " The Top 50 Avengers " . He was also ranked 38th on IGN 's list of " Top 100 Comic Book Heroes " . The character was first portrayed in live - action by Peter Hooten in the 1978 television film Dr. Strange . Benedict Cumberbatch stars as the character in the Marvel Cinematic Universe , first appearing in the role in the 2016 film Doctor Strange . He reprised the role in the 2017 film Thor : Ragnarok , the 2018 film Avengers : Infinity War , and will return in the fourth Avengers film in 2019 .
Passage 3
Strange appeared as a supporting character for the bulk of the 2000s . He appeared regularly in The Amazing Spider - Man under J. Michael Straczynski , before being cast into a time loop by Baron Mordo . He later appeared on and off in The New Avengers , where he was stated as being part of the secret group known as the Illuminati to deal with future threats to Earth . Ultimately Strange joined the team and allowed them to use his

doi.org/10.32388/PHS6YN

home as a base after the events of Civil War , which he sat out . Doctor Strange was critical of the federal Superhuman Registration Act and aided the anti-registration Avengers team led by Luke Cage .
Responses
Reference
I do not have an exact count of films in which Doctor Strange appears, but here are some: Doctor Strange, Thor: Ragnarok, Avengers: Infinity War, and The Amazing Spider-Man. Doctor Strange, voiced by Liam O'Brien, appears in the 2016 animated film Hulk: Where Monsters Dwell and will appear again in the untitled sequel to Avengers: Infinity War, which will return in the fourth Avengers film in 2019.
Llama 3.1 405B
According to the text, Doctor Strange appears in the following films in the Marvel Cinematic Universe: 1. Doctor Strange (2016) 2. Thor: Ragnarok (2017) 3. Avengers: Infinity War (2018) 4. The untitled sequel to Avengers: Infinity War (Avengers: Endgame, released in 2019, although not explicitly mentioned in the text) Additionally, he appears in the 2016 animated film Hulk: Where Monsters Dwell and the 1978 television film Dr. Strange, played by Peter Hooten.
GPT-4.0
Doctor Strange, portrayed by Benedict Cumberbatch, appears in the following Marvel Cinematic Universe films: 1. **Doctor Strange** (2016) 2. **Thor: Ragnarok** (2017) 3. **Avengers: Infinity War** (2018) 4. **Avengers: Endgame** (2019, referenced as the fourth Avengers film) 5. **Doctor Strange in the Multiverse of Madness** (subsequent appearances expected, although not specified in the provided documents) He also appears in the 2016 animated film *Hulk: Where Monsters Dwell*, voiced by Liam O'Brien.
Command-R+ (104b)
According to the information provided, Benedict Cumberbatch first portrayed the character of Doctor Strange in the Marvel Cinematic Universe in 2016. Cumberbatch reprised the role in two other films, "Thor: Ragnarok" and "Avengers: Infinity War", and will appear again in the untitled sequel to "Avengers: Infinity War". Doctor Strange also appears in the 2016 animated film "Hulk: Where Monsters Dwell", voiced by Liam O'Brien. Doctor Strange was first portrayed in live action by Peter Hooten in a 1978 television film.
Qwen 2.5 (72b)
Doctor Strange, portrayed by Benedict Cumberbatch, appears in at least four films within the Marvel Cinematic Universe: "Captain America: The Winter Soldier" (referenced), "Doctor Strange" (2016), "Thor: Ragnarok" (2017), and "Avengers: Infinity War" (2018). He is also set to appear in the 2019 sequel to "Avengers: Infinity

doi.org/10.32388/PHS6YN

War". Additionally, he appears in the 2016 animated film "Hulk: Where Monsters Dwell", voiced by Liam O'Brien.

Table 12. Example of an mtRAG task using the Reference (●) retrieval setting (this task is part of the conversation shown in Figure 1) with the Reference response as well as responses from several LLMs. It is clear that several models hallucinate.

Appendix E. Annotator agreement

We next describe the computation of the annotator agreement reported in the human evaluation (Section 7). For each property, the annotator agreement was computed as the percentage of tasks on which the 3 annotators had either an *absolute agreement* (i.e., they all agreed on the score) or *high agreement* (i.e., two of the annotators agreed on the score and the third annotator gave a score that was a single point away from the score given by the majority).

Appendix F. Metrics details

F.1. $\mathrm{RB}_{\mathrm{llm}}$ *implementation*

To adapt the RAD-Bench judge[7] to our RAG setting, we make two modifications: (1) We include the retrieved passages as additional input to the evaluation, and (2) we anchor the evaluation on our desired properties of faithfulness, appropriateness, and completeness. Figure 6 shows the final prompt that we used to implement $\mathrm{RB}_{\mathrm{llm}}$. Finally, in order to minimize model biases and improve evaluation reliability, we use four models as judges: GPT-4o-mini (2024-07-18), Qwen 2.5 (72B), Mixtral 8x22B Instruct, and Llama 3.1 405B Instruct, taking the median as the final score.

[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question given the provided document and a reference answer.

Your evaluation should assess the faithfulness, appropriateness, and completeness. Your evaluation should focus on the assistant's answer to the question of the current turn. You will be given the assistant's answer and a reference answer. You will also be given the user questions and assistant's answers of the previous turns of the conversation. You should consider how well the assistant's answer captures the key information, knowledge points mentioned in the reference answer, and how it respects or builds upon the focus and knowledge points from the previous turns.

[Faithfulness]: You are given the full conversation, the question of the current turn, the assistant's answer, and documents. You should evaluate how faithful is the assistant's answer to the information in the document and previous conversation.
[Appropriateness]: You should evaluate if the assistant's answer is relevant to the question of the current turn and if it addresses all the issues raised by the question without adding extra information.
[Completeness]: You should evaluate whether the assistant's answer is complete with information from the documents.

Begin your evaluation by comparing the assistant's answer against the reference answer in this turn. Be as objective as possible, and provide a detailed justification for your rating. After providing your explanation, you must rate the response on a scale of 1 to 10, strictly following this format: "Rating: [[rating]]", for example: "Rating: [[5]]".

[The Start of Previous Conversation]
{previous_conversation}
[The End of Previous Conversation]

[The Start of Current Turn Question]
{current_question}
[The End of Current Turn Question]

[The Start of Reference Answer]
{reference_answer}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{response}
[The End of Assistant's Answer]

[The Start of Document]
{passages}
[The End of Document]

doi.org/10.32388/PHS6YN

Figure 6. Prompt used for the $\mathrm{RB}_{\mathrm{llm}}$ judge

F.2. $\mathrm{RL}_{\mathrm{MTB}}$ *implementation*

The MT-Bench judge was originally designed for evaluating conversations containing exactly two turns (i.e., user question/agent response pairs), with a focus on properties such as helpfulness, depth, and creativity[14]. To use it for assessing general multi-turn conversations with relevant passages, we make the following changes: (1) We focus the evaluation on the last user turn of a multi-turn conversation, (2) we include retrieved passages as additional input, and (3) we modify the instruction to include desired properties, such as faithfulness and completeness. Figure 7 shows the final prompt that we used to implement $\mathrm{RL}_{\mathrm{MTB}}$.

[System]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question given the provided document. Your evaluation should consider factors such as the faithfulness, appropriateness, conciseness, relevance, accuracy, completeness, and correctness of their responses. You evaluation should focus on the assistant's answer to the final user question in the conversation. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: [[rating]], for example: Rating: [[5]].

<|The Start of Assistant A's Conversation with User|>
User:
{question_turn_1}

Assistant A:
{answer_turn_1}
...

User:
{question_turn_last}

Assistant A:
{answer_turn_last}

Document:
{passages}

<|The End of Assistant A's Conversation with User|>

Figure 7. Prompt used for the $\mathrm{RL}_{\mathrm{MTB}}$ judge

F.3. IDK correction to the metrics

To determine whether a model response corresponds to an IDK (I-Don't-know) answer, we use an LLM judge that we developed for this purpose. The judge sends the prompt of Figure 8 to GPT-4o-mini (2024-07-18). To measure the judge's performance, we evaluated its predictions on the human evaluation results across model responses and the reference answer, where the judge achieved over 97% accuracy. We assess the impact of the IDK judge on the unanswerable questions using our reference answers and responses from frontier models (GPT-4o, Llama 3.1 405B Instruct) in Table 13. These observations show that the IDK judge is needed to correct the scores.

doi.org/10.32388/PHS6YN

Read an inquiry and a response to it. Decide if the response contains, in part or in whole, an indication that there is not sufficient information to fully answer the inquiry. Use only one of the following labels: yes, no, partial. Answer "yes" if the response indicates that the inquiry cannot be answered, for example, due to lack of information (in a reference document or text). Answer "partial" if the response indicates insufficient information (in a reference document or text) to answer part of the inquiry, but still attempts to answer other part of the inquiry that can be answered. Answer "no" if the response does not indicate a lack of information preventing answering part of the inquiry, even if some reasoning steps or general knowledge are needed to address the inquiry.

[Inquiry]
How can I learn SQL without taking a course?

[Response]
I'm unable to provide an answer regarding how you can learn SQL without taking a course. However, I can assist with the No SQL databases topic.

[Output]
yes

[Additional ICL examples]

[Inquiry]
{inquiry}

[Response]
{response}

[Output]

Figure 8. IDK judge prompt

Model	RL_F	RL_F	RB_{llm}	RB_{llm}	RB_{alg}	RB_{alg}
Reference	0.18	0.87	0.89	0.87	0.75	0.87
GPT-4o	0.19	0.20	0.66	0.20	0.26	0.20
Llama 3.1	0.21	0.20	0.69	0.20	0.27	0.20

Table 13. Generation results in the Reference (●) retrieval setting before and after conditioning with IDK judge on 55 unanswerable questions

Appendix G. Synthetic conversation details

We next provide additional details on the mtRAG-S benchmark of synthetic multi-turn conversations.

G.1. Synthetic conversation generation

To create mtRAG-S, we extended the recently proposed framework of[29] to automatically generate user questions corresponding to the question types used in mtRAG. To classify question types within the synthetic conversations, we employed a question type classifier trained on metadata derived from the human-generated conversations in mtRAG. Finally, we used Mixtral 8x22B Instruct v0.1 as the LLM for the conversation generation and restricted the conversations to a maximum of 8 turns, as further increasing the number of turns tends to lead to repetitive user questions and a higher likelihood of hallucinated agent responses.

	mtRAG	mtRAG-S
(a) Based on $\mathrm{RL_F}$		
Command-R+ (104B)	0.76	0.83
GPT-4o-mini	0.71	0.81
Llama 3.1 405B Instruct	0.75	0.85
Mixtral 8x22B Instruct	0.61	0.79
Qwen 2.5 (72B)	0.72	0.82
(b) Based on BERT-K-Prec		
Command-R+ (104B)	0.33	0.38
GPT-4o-mini	0.27	0.34
Llama 3.1 405B Instruct	0.33	0.4
Mixtral 8x22B Instruct	0.29	0.41
Qwen 2.5 (72B)	0.3	0.38

Table 14. Comparing faithfulness of models on the human-generated mtRAG and synthetic mtRAG-S using two metrics: $\mathrm{RL_F}$ and BERT-K-Prec

doi.org/10.32388/PHS6YN

		Recall				nDCG			
		@1	@3	@5	@10	@1	@3	@5	@10
CLAPnq (208)	Last Turn	0.20	0.43	0.53	0.65	0.48	0.46	0.49	0.54
	Query Rewrite	0.22	0.45	0.56	0.70	0.54	0.50	0.53	0.59
FiQA (180)	Last Turn	0.15	0.34	0.44	0.55	0.38	0.36	0.40	0.45
	Query Rewrite	0.18	0.39	0.50	0.63	0.43	0.41	0.46	0.52
Govt (203)	Last Turn	0.18	0.42	0.50	0.58	0.42	0.44	0.46	0.49
	Query Rewrite	0.21	0.47	0.56	0.67	0.47	0.48	0.51	0.56
Cloud (189)	Last Turn	0.19	0.38	0.48	0.56	0.40	0.39	0.42	0.47
	Query Rewrite	0.20	0.40	0.47	0.57	0.42	0.41	0.43	0.48
All (780)	Last Turn	0.18	0.39	0.49	0.58	0.42	0.41	0.45	0.49
	Query Rewrite	0.20	0.43	0.52	0.64	0.46	0.45	0.48	0.54

Table 15. Elser Retrieval Performance of models on our benchmark using Recall and nDCG metrics per domain

G.2. Faithfulness of models on synthetic data vs human-generated data

We next provide detailed evaluation results on the comparison of faithfulness of models on mtRAG and mtRAG-S, discussed in Section 9. Tables 14(a) and 14(b) show faithfulness on the two benchmarks based on the RL_F and BERT-K-Prec metrics, respectively. Using either metric, we see that models across the board receive a higher faithfulness score on the synthetic than on the human-generated data.

Appendix H. Detailed retrieval results

Expanding on Table 3, which shows average retrieval results, in Table 15 we provide detailed retrieval results per domain for Elser, the best performing retriever.

doi.org/10.32388/PHS6YN

Appendix I. Detailed generation results

Expanding on the generation experiment results of Section 6 we provide two additional sets of results:

Table 16: Detailed generation results in the Reference (●) retrieval setting using three metrics, RL_F, RB_{llm} , and RB_{alg} , on three different dimensions: (a) answerability, (b) turns, and (c) domains

	Overall			**Answerable**			**Partial**			**Unans.**
	RL_F	RB_{llm}	RB_{alg}	RL_F	RB_{llm}	RB_{alg}	RL_F	RB_{llm}	RB_{alg}	
Reference	**0.87**	**0.95**	**0.88**	**0.88**	**0.96**	**0.88**	**0.71**	**0.88**	**0.83**	**0.87**
Command-R+ (104B)	**0.76**	0.69	0.44	**0.82**	0.74	0.47	**0.59**	0.63	0.36	0.13
GPT-4o	0.75	**0.76**	0.45	**0.82**	**0.81**	**0.48**	0.53	0.71	0.35	0.20
GPT-4o-mini	0.71	0.75	0.43	0.77	0.79	0.44	0.39	0.62	0.30	0.34
Llama 3.1 405B Instruct	0.75	0.74	**0.47**	0.81	0.79	0.50	0.58	0.66	0.37	0.20
Llama 3.1 70B Instruct	0.69	0.66	0.44	0.74	0.69	0.45	0.42	0.47	0.27	**0.44**
Llama 3.1 8B Instruct	0.55	0.59	0.36	0.59	0.62	0.38	0.34	0.47	0.24	0.33
Mixtral 8x22B Instruct	0.61	0.69	0.41	0.68	0.75	0.45	0.41	0.68	0.33	0.00
Qwen 2.5 (72B)	0.72	0.74	0.44	0.79	0.80	0.47	0.53	**0.72**	**0.38**	0.07
Qwen 2.5 (7B)	0.68	0.72	0.43	0.74	0.77	0.46	0.44	0.67	0.36	0.11

Table 16a. By question answerablity

doi.org/10.32388/PHS6YN

	RL_F		RB_{llm}		RB_{alg}	
	TURN 1	**> TURN 1**	**TURN 1**	**> TURN 1**	**TURN 1**	**> TURN 1**
Reference	**0.89**	**0.86**	**0.97**	**0.95**	**0.89**	**0.87**
Command-R+ (104B)	0.83	**0.75**	0.70	0.69	0.46	0.43
GPT-4o	**0.86**	0.74	0.78	**0.76**	**0.54**	0.44
GPT-4o-mini	0.84	0.69	**0.79**	0.74	0.50	0.42
Llama 3.1 405B Instruct	0.81	0.74	0.74	0.74	0.50	**0.46**
Llama 3.1 70B Instruct	0.80	0.68	0.69	0.66	0.50	0.42
Llama 3.1 8B Instruct	0.66	0.53	0.56	0.59	0.41	0.36
Mixtral 8x22B Instruct	0.82	0.58	0.73	0.69	0.47	0.40
Qwen 2.5 (72B)	0.84	0.70	0.77	0.74	0.51	0.43
Qwen 2.5 (7B)	0.82	0.65	0.71	0.72	0.48	0.42

Table 16b. By first vs subsequent turns

	RL_F				RB_{llm}				RB_{alg}			
	CLAPnq	**FiQA**	**Govt**	**Cloud**	**CLAPnq**	**FiQA**	**Govt**	**Cloud**	**CLAPnq**	**FiQA**	**Govt**	**Cloud**
Reference	**0.86**	**0.89**	**0.85**	**0.87**	**0.96**	**0.95**	**0.95**	**0.95**	**0.88**	**0.88**	**0.88**	**0.87**
Command-R+	**0.78**	**0.76**	0.76	0.73	0.72	0.70	0.68	0.66	0.47	0.39	0.43	0.45
GPT-4o	0.73	0.74	**0.77**	**0.78**	**0.77**	**0.74**	**0.78**	**0.74**	**0.48**	0.38	0.47	0.45
GPT-4o-mini	0.69	0.71	0.71	0.75	0.76	**0.74**	0.75	0.73	0.45	0.37	0.44	0.44
Llama 3.1 405B Inst.	0.72	**0.76**	**0.77**	0.76	0.75	0.73	0.75	0.72	**0.48**	**0.41**	**0.49**	**0.49**
Llama 3.1 70B Inst.	0.66	0.67	0.73	0.71	0.66	0.62	0.69	0.65	0.44	0.37	0.45	0.47
Llama 3.1 8B Inst.	0.52	0.56	0.56	0.56	0.62	0.55	0.60	0.57	0.37	0.32	0.38	0.38
Mixtral 8x22B Inst.	0.60	0.60	0.61	0.64	0.72	0.68	0.70	0.68	0.45	0.34	0.42	0.44
Qwen 2.5 (72B)	0.68	0.70	0.74	0.76	0.75	0.73	0.76	0.73	0.46	0.37	0.46	0.46
Qwen 2.5 (7B)	0.70	0.63	0.68	0.69	0.72	0.69	0.74	0.71	**0.48**	0.35	0.44	0.44

Table 16c. By domain

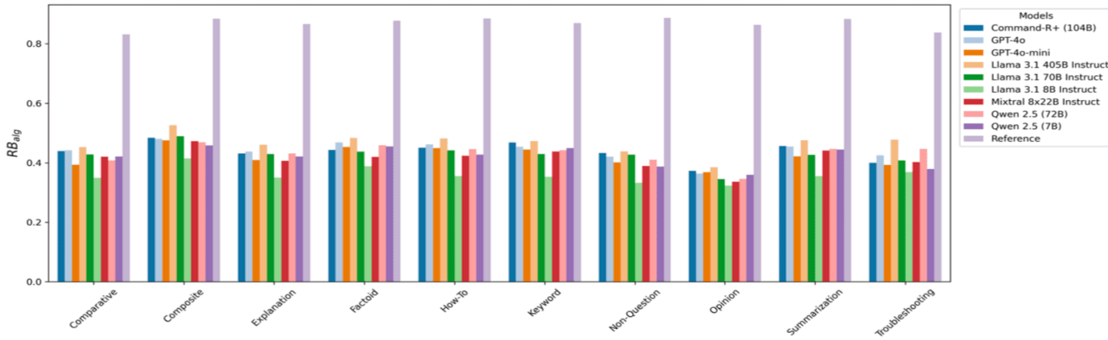
I.1. Generation results for all three metrics

Expanding on Figure 3, which shows generation results in the Reference Generation (•) setting using a single metric, RB_{alg}, we now show generation results for the same setting using all three metrics: RB_{alg}, RB_{llm}, and RL_F. Tables 16(a), 16(b), and 16(c) present the results broken down by answerability, turns, and domain, respectively.

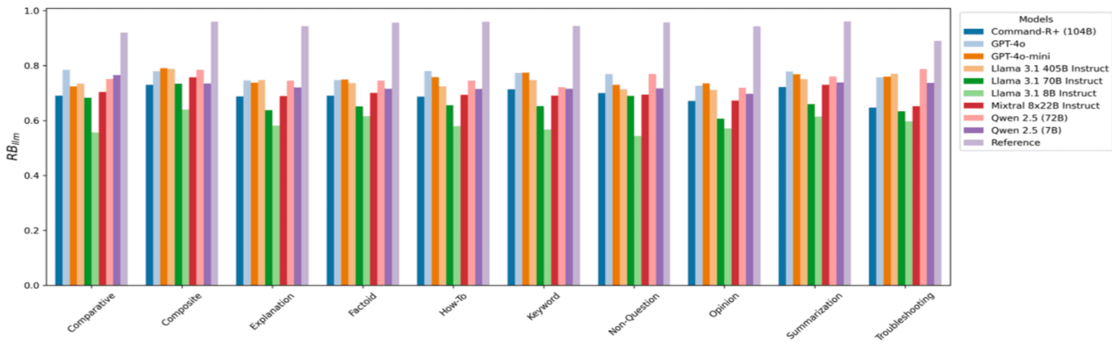
I.2. Generation results by question type and multi-turn type

Using the same three metrics, we also include generation results broken down by question and multi-turn type, shown in Figures 9 and 10, respectively.

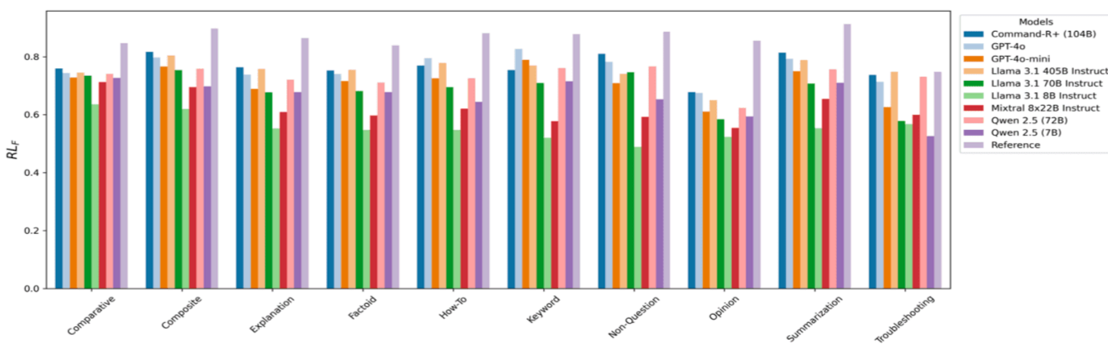
doi.org/10.32388/PHS6YN



(a) Based on RB_{alg}



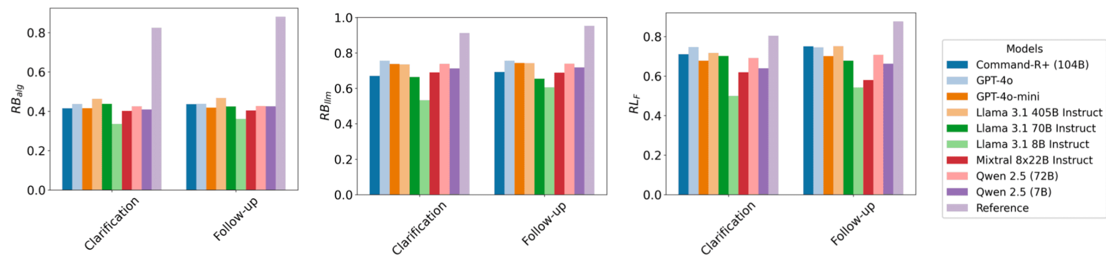
(b) Based on RB_{llm}



(c) Based on RL_F

Figure 9. Generation results in the Reference (●) retrieval setting by question type based on three metrics: RB_{alg}, RB_{llm}, and RL_F

 doi.org/10.32388/PHS6YN



(a) Based on RB_{alg} (b) Based on RB_{llm} (c) Based on RL_F

Figure 10. Generation results in the Reference (●) retrieval setting by multi-turn type based on three metrics: RB_{alg}, RB_{llm}, and RL_F

Footnotes

[1] https://www.elastic.co/guide/en/machine-learning/current/ml-nlp-elser.html

[2] https://www.elastic.co/search-labs/blog/elasticsearch-elser-relevance-mteb-comparison

[3] We use the 22B and not the 7B variant of Mixtral, due to the fact that the latter was used during conversation generation

[4] https://openai.com/index/gpt-4o-system-card/

[5] https://huggingface.co/CohereForAI/c4ai-command-r-plus

[6] These results are limited to the 426 Reference + RAG (◐) tasks to be consistent across retrieval settings. Appendix I shows that the trends persist throughout the benchmark.

[7] https://www.appen.com/

[8] mtRAG is publicly available at https://github.com/ibm/mt-rag-benchmark

[9] https://platform.openai.com/docs/models#o1

[10] https://huggingface.co/docs/transformers/main/en/chat_templating

References

1. ^Lin CC, Huang AYQ, Yang SJH (2023). "A review of AI-driven conversational chatbots implementation methodologies and challenges (1999–2022)". Sustainability. 15(5). doi:10.3390/su15054012.

2. ^Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, et al. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks". Advances in Neural Information Processing Systems. 33: 9459–9474. proceedings.neurips.cc.
3. ^Chen J, Lin H, Han X, Sun L (2024). "Benchmarking large language models in retrieval-augmented generation". Proceedings of the AAAI Conference on Artificial Intelligence. 38(16): 17754–17762. doi:10.1609/aaai.v38i16.29728.
4. ^Friel R, Belyi M, Sanyal A (2024). "RAGBench: Explainable benchmark for retrieval-augmented generation systems". Preprint, arXiv:2407.11005. arXiv:2407.11005.
5. ^Niu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, Song J, Zhang T (2024). "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models". In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.585.
6. a, bYang X, Sun K, Xin H, Sun Y, Bhalla N, Chen X, et al. (2024). CRAG – Comprehensive RAG benchmark. arXiv preprint arXiv:2406.04744.
7. a, b, c, d, e, fKuo TL, Liao FT, Hsieh MW, Chang FC, Hsu PC, Shiu DS (2024). "RAD-Bench: Evaluating large language models capabilities in retrieval augmented dialogues". Preprint, arXiv:2409.12558. arXiv:2409.12558.
8. ^Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016). "SQuAD: 100,000+ questions for machine comprehension of text". In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. doi:10.18653/v1/D16-1264.
9. ^Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang MW, Dai AM, Uszkoreit J, Le Q, Petrov S (2019). "Natural questions: A benchmark for question answering research". Transactions of the Association for Computational Linguistics. 7: 452–466. doi:10.1162/tacl_a_00276.
10. ^Fan A, Jernite Y, Perez E, Grangier D, Weston J, Auli M (2019). "ELI5: Long form question answering". In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3558–3567, Florence, Italy. Association for Computational Linguistics. doi:10.18653/v1/P19-1346.
11. ^Stelmakh I, Luan Y, Dhingra B, Chang MW (2022). "ASQA: Factoid questions meet long-form answers". In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages

8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.566.

12. a, b, c, dRosenthal S, Sil A, Florian R, Roukos S (2024). "CLAPNQ: Cohesive long-form answers from passages in natural questions for RAG systems". Preprint, arXiv:2404.02103. arXiv:2404.02103.
13. ^Lee K, Chang MW, Toutanova K (2019). "Latent retrieval for weakly supervised open domain question answering". In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6086–6096, Florence, Italy. Association for Computational Linguistics. doi:10.18653/v1/P19-1612.
14. a, b, cLianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, Ion Stoica (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems, volume 36, pages 46595–46623. Curran Associates, Inc.
15. ^Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J (2019). "Wizard of wikipedia: Knowledge-powered conversational agents". In International Conference on Learning Representations. openreview.net.
16. ^Bai Y, Wang DZ (2021). "More than reading comprehension: A survey on datasets and metrics of textual question answering". CoRR, abs/2109.12264. arXiv:2109.12264.
17. ^Cambazoglu BB, Sanderson M, Scholer F, Croft B (2021). "A review of public datasets in question answering research". SIGIR Forum. 54(2). doi:10.1145/3483382.3483389.
18. ^Zhen Wang (2022). Modern question answering datasets and benchmarks: A survey. Preprint, arXiv:2206.15030.
19. ^Rogers A, Gardner M, Augenstein I (2023). "QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension". ACM Comput. Surv.. 55(10). doi:10.1145/3560260.
20. ^Choi E, He H, Iyyer M, Yatskar M, Yih WT, Choi Y, Liang P, Zettlemoyer L (2018). "QuAC: Question answering in context". In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics. doi:10.18653/v1/D18-1241.
21. ^Qu C, Yang L, Chen C, Qiu M, Croft WB, Iyyer M (2020). "Open-retrieval conversational question answering". In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, page 539–548, New York, NY, USA. Association for Computing Machinery. doi:10.1145/3397271.3401110.

22. ^Reddy S, Chen D, Manning CD (2019). "CoQA: A conversational question answering challenge". Transactions of the Association for Computational Linguistics. 7: 249–266. doi:10.1162/tacl_a_00266.
23. ^Saeidi M, Bartolo M, Lewis P, Singh S, Rocktäschel T, Sheldon M, Bouchard G, Riedel S (2018). "Interpretation of natural language rules in conversational machine reading". In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics. doi:10.18653/v1/D18-1233.
24. ^Feng S, Patel SS, Wan H, Joshi S (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. doi:10.18653/v1/2021.emnlp-main.498.
25. a, bDziri N, Kamalloo E, Milton S, Zaiane O, Yu M, Ponti EM, Reddy S (2022). "FaithDial: A faithful benchmark for information-seeking dialogue". Transactions of the Association for Computational Linguistics. 10: 1473–1490. doi:10.1162/tacl_a_00529.
26. a, b, cAliannejadi M, Abbasiantaeb Z, Chatterjee S, Dalton J, Azzopardi L (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, page 819–829, New York, NY, USA. Association for Computing Machinery. doi:10.1145/3626772.3657860.
27. ^Jiang Z, Xu F, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, Yang Y, Callan J, Neubig G (2023). "Active retrieval augmented generation". In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7969–7992, Singapore. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.495.
28. ^Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, Soujanya Poria (2024). Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. Preprint, arXiv:2409.11242.
29. a, b, cLee YS, Gunasekara C, Contractor D, Fernandez Astudillo R, Florian R (2024). "Multi-document grounded multi-turn synthetic dialog generation". Preprint, arXiv:2409.11500. arXiv:2409.11500.
30. ^Rajpurkar P, Jia R, Liang P (2018). "Know what you don't know: Unanswerable questions for SQuAD". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 784–789, Melbourne, Australia. Association for Computational Linguistics. doi:10.18653/v1/P18-2124.

31. ^Yuhong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, Hui Zhao (2024). Benchmarking hallucination in large language models based on unanswerable math word problem. Preprint, arXiv:2403.03558.

32. a, b, cEs S, James J, Espinosa Anke L, Schockaert S (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. aclanthology.org.

33. a, bMaia M, Handschuh S, Freitas A, Davis B, McDermott R, Zarrouk M, Balahur A (2018). "WWW'18 open challenge: Financial opinion mining and question answering". In Companion Proceedings of the The Web Conference 2018, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. doi:10.1145/3184558.3192301.

34. a, bJiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, de las Casas D, Bou Hanna E, Bressand F, Lengyel G, Bour G, Lample G, Lavaud LR, Saulnier L, Lachaux MA, Stock P, Subramanian S, Yang S, Antoniak S, Le Scao T, Gervet T, Lavril T, Wang T, Lacroix T, El Sayed W (2024). "Mixtral of experts". Preprint, arXiv:2401.04088. arXiv:2401.04088.

35. ^Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff (2023). C-Pack: Packaged resources to advance general chinese embedding. Preprint, arXiv:2309.07597.

36. ^Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

37. ^Yingxue Zhou, Jie Hao, Mukund Rungta, Yang Liu, Eunah Cho, Xing Fan, Yanbin Lu, Vishal Vasudevan, Kellen Gillespie, Zeynab Raeesy, Wei (Sawyer) Shen, Chenlei (Edward) Guo, Gokhan Tur (2023). Unified contextual query rewriting. In ACL 2023.

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

39. ^Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, Goyal A, Hartshorn A, Yang A, Mitra A, Sravankumar A, Korenev A, Hinsvark A, Rao A, Zhang A, Rodriguez A, Gregerson A, Spataru A, Roziere B, Biron B, Tang B, Chern B, Caucheteux C, Nayak C, Bi C, Marra C, McConnell C, Keller C, Touret C, Wu C, Wong C, Ferrer CC, Nikolaidis C, Allonsius D, Song D, Pintz D, Livshits D, Esiobu D, Choudhary D, Mahajan D, Garcia-Olano D, Perino D, Hupkes D, Lakomkin E, AlBadawy

E, Lobanova E, Dinan E, Smith EM, Radenovic F, Zhang F, Synnaeve G, Lee G, Anderson GL, Nail G, Mialon G, Pang G, Cucurell G, Nguyen H, Korevaar H, Xu H, Touvron H, Zarov I, Ibarra IA, Kloumann I, Misra I, Evtimov I, Copet J, Lee J, Geffert J, Vranes J, Park J, Mahadeokar J, Shah J, Linde Jv, Billock J, Hong J, Lee J, Fu J, Chi J, Huang J, Liu J, Wang J, Yu J, Bitton J, Spisak J, Park J, Rocca J, Johnstun J, Saxe J, Jia J, Alwala KV, Upasani K, Plawiak K, Li K, Heafield K, Stone K, El-Arini K, Iyer K, Malik K, Chiu K, Bhalla K, Rantala-Yeary L, Maaten Lvd, Chen L, Tan L, Jenkins L, Martin L, Madaan L, Malo L, Blecher L, Landzaat L, Oliveira Ld, Muzzi M, Pasupuleti M, Singh M, Paluri M, Kardas M, Oldham M, Rita M, Pavlova M, Kambadur M, Lewis M, Si M, Singh MK, Hassan M, Goyal N, Torabi N, Bashlykov N, Bogoychev N, Chatterji N, Duchenne O, Çelebi O, Alrassy P, Zhang P, Li P, Vasic P, Weng P, Bhargava P, Dubal P, Krishnan P, Koura PS, Xu P, He Q, Dong Q, Srinivasan R, Ganapathy R, Calderer R, Cabral RS, Stojnic R, Raileanu R, Girdhar R, Patel R, Sauvestre R, Polidoro R, Sumbaly R, Taylor R, Silva R, Hou R, Wang R, Hosseini S, Chennabasappa S, Singh S, Bell S, Kim SS, Edunov S, Nie S, Narang S, Raparthy S, Shen S, Wan S, Bhosale S, Zhang S, Vandenhende S, Batra S, Whitman S, Sootla S, Collot S, Gururangan S, Borodinsky S, Herman T, Fowler T, Sheasha T, Georgiou T, Scialom T, Speckbacher T, Mihaylov T, Xiao T, Karn U, Goswami V, Gupta V, Ramanathan V, Kerkez V, Gonguet V, Do V, Vogeti V, Petrovic V, Chu W, Xiong W, Fu W, Meers W, Martinet X, Wang X, Tan XE, Xie X, Jia X, Wang X, Goldschlag Y, Gaur Y, Babaei Y, Wen Y, Song Y, Zhang Y, Li Y, Mao Y, Coudert ZD, Yan Z, Chen Z, Papakipos Z, Singh A, Grattafiori A, Jain A, Kelsey A, Shajnfeld A, Gangidi A, Victoria A, Goldstand A, Menon A, Sharma A, Boesenberg A, Vaughan A, Baevski A, Feinstein A, Kallet A, Sangani A, Yunus A, Lupu A, Alvarado A, Caples A, Gu A, Ho A, Poulton A, Ryan A, Ramchandani A, Franco A, Saraf A, Chowdhury A, Gabriel A, Bharambe A, Eisenman A, Yazdan A, James B, Maurer B, Leonhardi B, Huang B, Loyd B, De Paola B, Paranjape B, Liu B, Wu B, Ni B, Hancock B, Wasti B, Spence B, Stojkovic B, Gamido B, Montalvo B, Parker C, Burton C, Mejia C, Wang C, Kim C, Zhou C, Hu C, Chu CH, Cai C, Tindal C, Feichtenhofer C, Civin D, Beaty D, Kreymer D, Li D, Wyatt D, Adkins D, Xu D, Testuggine D, David D, Parikh D, Liskovich D, Foss D, Wang D, Le D, Holland D, Dowling E, Jamil E, Montgomery E, Presani E, Hahn E, Wood E, Brinkman E, Arcaute E, Dunbar E, Smothers E, Sun F, Kreuk F, Tian F, Ozgenel F, Caggioni F, Guzmán F, Kanayet F, Seide F, Medina Florez G, Schwarz G, Badeer G, Swee G, Halpern G, Thattai G, Herman G, Sizov G, Zhang G, Lakshminarayanan G, Shojanazeri H, Zou H, Wang H, Zha H, Habeeb H, Rudolph H, Suk H, Aspegren H, Goldman H, Damlaj I, Molybog I, Tufanov I, Veliche IE, Gat I, Weissman J, Geboski J, Kohli J, Asher J, Gaya JB, Marcus J, Tang J, Chan J, Zhen J, Reizenstein J, Teboul J, Zhong J, Jin J, Yang J, Cummings J, Carvill J, Shepard J, McPhie J, Torres J, Ginsburg J, Wang J, Wu K, U KH, Saxena K, Prasad K, Khandelwal K, Zand K, Matosich K, Veeraraghavan K, Michelena K, Li K, Huang K, Chawla K, Lakhotia K, Huang K, Chen L, Garg L, A L, Silva L, Bell L, Zhang L, Guo L, Yu L, Moshkovi *ch L, Wehrstedt L, Khabsa M, Avalani M, Bhatt M, Tsimpoukelli M, Mankus M, Hasson M, Lennie M, Reso M, Groshev M, Naumov M, Lathi M, Keneally M, Seltzer ML, Valko M, Restrepo M, Patel M, Vyatskov M, Samvelyan M, Clark M, Macey M, Wang M, Jubert Hermoso M, Metanat M, Rastegari M, Bansal M, Santhanam N, Parks N, White N, Bawa N, Singhal N, Egebo N, Usunier N, Laptev NP, Dong N, Zhang N, Cheng N, Chernoguz O, Hart O, Salpekar O, Kalinli O, Kent P, Parekh P, Saab P, Balaji P, Rittner P, Bontrager P, Roux P, Dollar P, Zvyagina P, Ratanchandani P, Yuvraj P, Liang Q, Alao R, Rodriguez R, Ayub R, Murthy R, Nayani R, Mitra R, Li R, Hogan R, Battey R, Wang R, Maheswari R, Howes R, Rinott R, Bondu SJ, Datta S, Chugh S, Hunt S, Dhillon S, Sidorov S, Pan S, Verma S, Yamamoto S, Ramaswamy S, Lindsay S, Lindsay S, Feng S, Lin S, Zha SC, Shankar S, Zhang S, Zhang S, Wang S, Agarwal S, Sajuyigbe S, Chintala S, Max S, Chen S, Kehoe S, Satterfield S, Govindaprasad S, Gupta S, Cho S, Virk S, Subramanian S, Choudhury S, Goldman S, Remez T, Glaser T, Best T, Kohler T, Robinson T, Li T, Zhang T, Matthews T, Chou T, Shaked T, Vontimitta V, Ajayi V, Montanez V, Mohan V, Kumar VS, Mangla V, Albiero V, Ionescu V, Poenaru V, Mihailescu VT, Ivanov V, Li W, Wang W, Jiang W, Bouaziz W, Constable W, Tang X, Wang X, Wu X, Wang X, Xia X, Wu X, Gao X, Chen Y, Hu Y, Jia Y, Qi Y, Li Y, Zhang Y, Zhang Y, Adi Y, Nam Y, Wang Y, Hao Y, Qian Y, He Y, Rait Z, DeVito Z, Rosnbrick Z, Wen Z, Yang Z, Zhao Z (2024). "The Llama 3 herd of models". Preprint, arXiv:2407.21783. arXiv:2407.21783.*

40. ^*Qwen Team (2024). Qwen2.5: A party of foundation models.*
41. ^*Adlakha V, BehnamGhader P, Lu XH, Meade N, Reddy S (2024). "Evaluating correctness and faithfulness of instruction-following models for question answering". Transactions of the Association for Computational Linguistics. 12: 681–699. doi:10.1162/tacl_a_00667.*
42. ^*Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, Faegheh Hasibi (2024). A survey on recent advances in conversational data generation. Preprint, arXiv:2405.13003.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.