Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA

Nirmal Roy[†*], **Leonardo F. R. Ribeiro**[‡], **Rexhina Blloshmi**[‡], **Kevin Small**[‡]

[†]TU Delft, [‡]Amazon

nroy000@gmail.com, {leonribe, blloshmi, smakevin}@amazon.com

Abstract

Augmenting Large Language Models (LLMs) with information retrieval capabilities (i.e., Retrieval-Augmented Generation (RAG)) has proven beneficial for knowledge-intensive tasks. However, understanding users' contextual search intent when generating responses is an understudied topic for conversational question answering (QA). This conversational extension leads to additional concerns when compared to single-turn QA as it is more challenging for systems to comprehend conversational context and manage retrieved passages over multiple turns. In this work, we propose a method for enabling LLMs to decide *when* to retrieve in RAG settings given a conversational context. When retrieval is deemed necessary, the LLM then rewrites the conversation for passage retrieval and judges the relevance of returned passages before response generation. Operationally, we build on the single-turn SELF-RAG framework (Asai et al., 2023) and propose SELF-*multi*-RAG for conversational settings. SELF-*multi*-RAG demonstrates improved capabilities over single-turn variants with respect to retrieving relevant passages (by using summarized conversational context) and assessing the quality of generated responses. Experiments on three conversational QA datasets validate the enhanced response generation capabilities of SELF-*multi*-RAG, with improvements of $\sim$13% measured by human annotation.

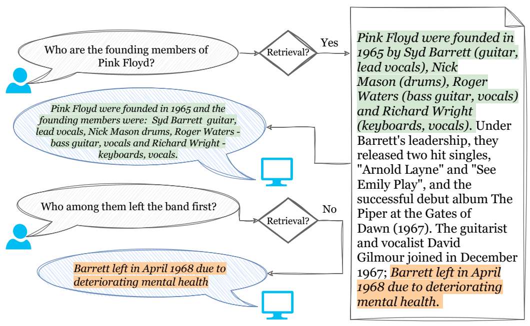


Figure 1: **Understanding conversational context.** In multi-turn conversations user questions often refer to responses in previous turns based on passages already retrieved, as shown in the example above. To answer the follow-up question, it is not necessary to retrieve new passages and the LLM should refer back to the previously retrieved passage, which contains the response.

1 Introduction

Recent advances in LLM technology has made the conversational search paradigm (Culpepper et al., 2018) increasingly viable for mainstream use as next generation search technology. Unlike traditional search engines which primarily process keyword queries, users can treat conversational search systems as a knowledgeable expert and directly engage in a multi-turn natural language conversation to better resolve their search needs. However, despite impressive abilities in a variety of tasks including response generation and conversational understanding, factual errors and hallucinations remain persistent problems for LLM-based systems (Mallen et al., 2023; Wei et al., 2024).

Retrieval-Augmented Generation (RAG) methods have been shown to partially ameliorate these issues by augmenting the input of LLMs with relevant retrieved passages, aiming to reduce factual errors in knowledge-intensive tasks (Lewis et al., 2020). Nonetheless, these approaches also can impede the flexibility of LLMs, introducing extraneous or unrelated passages or providing conflicting information with previous context/turns (Adlakha et al., 2022), resulting in low-quality generation (Shi et al., 2023). Specifically, retrieving passages indiscriminately, without considering whether the factual grounding is beneficial can compromise the quality of the generated content (Shi et al., 2023; Oh and Thorne, 2023).

Thus, understanding *if retrieval is necessary*

[*] Work done as an intern at Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10604–10625
November 12-16, 2024

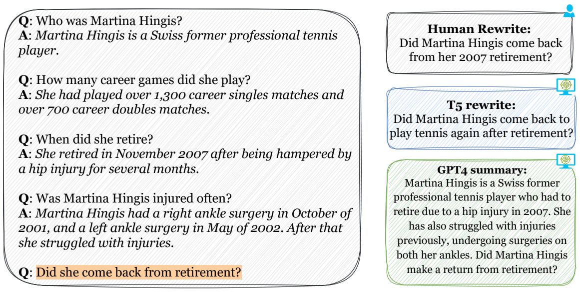


Figure 2: **Summarizing conversational context.** While using the entire conversational context as query to a retrieval model might introduce noise, using traditional rewriting methods might miss on important aspects of the conversations. Conversation summaries provide an adaptable approach as a retrieval query.

for high-quality response generation is an important research question, especially in the context of conversational QA. In multi-turn question answering, comprehending users' contextual intent and generating responses pose significant challenges due to complexities introduced by the extended context window containing previous user interactions (Aliannejadi et al., 2020; Mao et al., 2023b; Wu et al., 2024). The system, when deciding whether to retrieve or estimating the usefulness of its own response, must process a longer context of the conversation history, understand the user intent of current turn, ensure prevention of information repetition, maintain user engagement, etc. An example is provided in Figure 1, which shows that the decision to retrieve or not might depend on the conversational context rather than last turn only.

Furthermore, when retrieval is expensive or noisy, it is beneficial to utilize already retrieved documents in the conversation, given they are relevant and contain the necessary answers. Additionally, detrimental context as a result of noisy retrieval can degrade response generation quality (Shi et al., 2023; Oh and Thorne, 2023). Lastly, when conversation history is longer, traditional conversational query rewriting methods (Anantha et al., 2020; Ishii et al., 2022; Ye et al., 2023) that typically emphasize co-reference resolution, might be insufficient to contain all the information required for an effective retrieval (Bai et al., 2024). As shown in Figure 2, both the gold and a T5-based query rewriting miss potentially important signals (e.g., hip injury) for retrieving correct passages. In that case, representing the conversation history in a summarized form can lead to more effective retrieval.

In this work, we propose `SELF`-*multi*-`RAG`, an approach for efficient retrieval during a multi-turn conversation for improved response generation. `SELF`-*multi*-`RAG` provides refined contextual signals for retrieval and enhances multi-turn answer generation by potentially reusing already retrieved passages in previous turns. In particular, `SELF`-*multi*-`RAG` enables: (i) better *understanding of the conversational context* to decide whether retrieval is needed and to generate useful responses accordingly in a conversational setting and (ii) *summarizing the conversational context* such that it can be used as a query to retrieve relevant documents, when needed, with high retrieval effectiveness. Our specific contribution are as follows:

- We propose `SELF`-*multi*-`RAG`, a framework to train LLMs to respond to question-answer turns by adaptively retrieving passages and reflecting on the retrieved passages in multi-turn setting. `SELF`-*multi*-`RAG` determines the necessity of retrieval given a conversational context and summarizes the conversation into a query for use with off-the-shelf retrieval models.

- We conduct extensive experiments to observe that response quality of `SELF`-*multi*-`RAG` significantly outperforms `SELF-RAG` with an average improvement of ~13% for conversational datasets, measured by human annotations in Table 4. Moreover, `SELF`-*multi*-`RAG` summarization capabilities improves the retrieval effectiveness by 13.5% on average (R@5), compared to query rewriting baselines (Table 5).

2 Related Work

RAG. Retrieval-Augmented Generation (RAG) augments the LM input with retrieved text passages (Lewis et al., 2020; Guu et al., 2020), leading to large improvements in knowledge-intensive tasks (Ram et al., 2023). However, the improved task performance of such approaches have been shown to come at the expense of runtime efficiency (Mallen et al., 2023), robustness to irrelevant context (Oh and Thorne, 2023; Shi et al., 2023), and lack of attributions (Liu et al., 2023; Gao et al., 2023). Yoran et al. (2023) use a natural language inference model and Xu et al. (2023) employ a summarization model to filter out or compress retrieved passages before using them to prompt the LM to generate the output. In comparison, `SELF-RAG` (Asai et al., 2023) processes passages in parallel and filters out irrelevant ones through self-reflection, without relying on external models at inference. The self-reflection mechanism of SELF-RAG also evaluates other aspects of the model output quality, including factuality and attribution. However, SELF-RAG is not trained to comprehend conversational context, which we specifically equip SELF-*multi*-RAG to do. Kulkarni et al. (2024) propose a reinforcement learning (RL) based approach where the policy model can perform two actions: fetch conversation context or skip retrieval. Their approach was shown to save costs by reducing tokens when the model decides retrieval is not needed, while also slightly improving response generation. In contrast, the goal of SELF-*multi*-RAG is not only to decrease retrieval redundancy but also increase retrieval effectiveness.

LLMs and Multi-turn Conversations. In order to enable LLMs to interact with humans in a dialogue-based settings, the standard approach is to collect multi-turn instructions (Chiang et al., 2023; Ji et al., 2023), often synthetically generated using strong LLMs, and used to fine tune the LLMs for the task of response generation. This process is known as instruction fine-tuning, which enables LLMs to generate responses in a multi-turn dialogue setting. LLMs have also been used to perform conversational history modeling by rewriting user question (Mao et al., 2023a; Ye et al., 2023; Wang et al., 2023). Such query rewriting using LLMs have been shown to improve effectiveness for the retrieval of grounding passages. However, none of these works explicitly train the LLMs to reflect whether retrieval is needed or not (given the conversation history) or how to deal with irrelevant passages while generating RAG responses.

Conversational Query Rewriting. Query rewriting plays a vital role in enhancing conversational search by transforming context-dependent user queries into self-contained forms. Existing approaches (Wu et al., 2021; Mo et al., 2023) primarily leverage human-rewritten queries as labels to train query rewriting models and typically aim to convert the conversational into a single question. (Ye et al., 2023) proposed to rewrite queries using the conversation history to make more informative queries. They show that rewriting queries by prompting ChatGPT with information from context helps in more effective retrieval performance. Kaiser et al. (2024) performed RL based reformulations for better retrieval of entities for conversational QA over knowledge graphs. Their reformulations are entity focused where the answers to the questions are entities (as compared to sentences in the case of open domain QA). Ishii et al. (2022) proposed query rewriting based on a reward based system. The current question + conversation history is passed to the QA model (e.g., RoBERTa) to extract answer span from provided evidence document. Jang et al. (2024) rely on information retrieval signals directly to perform conversational query rewriting instead of relying on human-rewritten query as supervision signal. While the above works show improved retrieval effectiveness as compared to a human rewrite of the conversational context they *do not* evaluate how the retrieved passages affect response generation of the models. Furthermore, we hypothesize that summarizing the conversational context instead of rewriting them to a single question will help in improving retrieval effectiveness and consequent response generation performance.

3 SELF-*multi*-RAG

In this work, we propose SELF-*multi*-RAG, which extends SELF-RAG (Asai et al., 2023) to generate responses in a conversational setting. Our proposed methodology trains a LLM to comprehend longer conversational contexts to learn when retrieval is needed and also critiquing the quality of the passages and its own generation given the previous turns by generating *special tokens*. Importantly, when the LLM decides retrieval is needed to generate a response, we train it to summarize the conversational context for more effective passage retrieval, which consequently leads to better response generation. SELF-*multi*-RAG has two important functions:

Understanding Conversational Context. In a multi-turn context, evaluating the relevance of a retrieved passage involves considering both the current question and the previous conversation history, unlike a single-turn question. Furthermore, retrieved passages from previous turns can also be provided in the conversational context or a user may simply provide a passage and ask questions based on that. When the cost of retrieval is high, a model should *not* decide to retrieve if the question in the current turn can be answered from passages retrieved in previous turns (as shown in Figure 1). In that case, the model should not only comprehend the conversational context but also the previously retrieved passages to be able to determine the

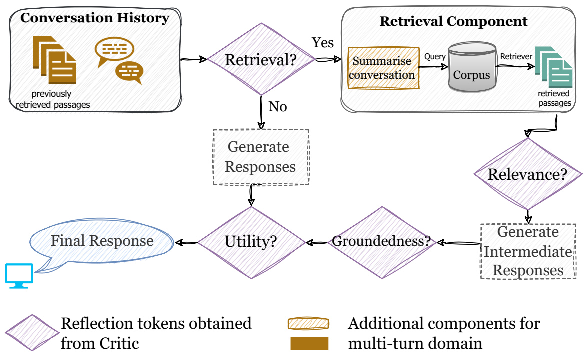


Figure 3: **`SELF`-*multi*-`RAG` framework.** Components of the pipeline highlighted in yellow are specific to multi-turn conversations. The critic model is used to obtain the special reflection tokens that the generator model is trained to predict while generating response.

necessity of retrieval. Utilizing already retrieved passages (given they are relevant) not only mitigates the harmful effects of noisy retrieval (Oh and Thorne, 2023), but also saves costs by reducing the number of context tokens (Kulkarni et al., 2024).

Summarising Conversational Context. Representing conversational context using a single question is difficult (Anantha et al., 2020) and might result in loss of information while retrieving relevant passages (as shown in Figure 2). This is especially true in case of long conversations. Traditional conversation query rewriting methods (Ye et al., 2023; Mo et al., 2023; Lin et al., 2020) are trained to select important parts of the conversation but, typically, in a single question format. Hence, we hypothesize that summarising a conversational context can potentially include more relevant signals when retrieving passages without adding noise. This can be beneficial for both sparse and bi-encoder based dense retrieval.

3.1 Components

Analogous to `SELF-RAG`, `SELF`-*multi*-`RAG` has three main components: (i) Critic, (ii) Generator, and (iii) Retriever.

Critic. The task of the critic model is to output the special reflection tokens given a conversational context (as compared to a single question like in the original `SELF-RAG` framework). We employ the five critic tasks introduced by Asai et al. (2023). However, we redesign the framework to include a conversational history instead of a single-turn question. The important distinction of our approach is that it teaches the critic model to judge whether

Task	Definition	# instances	Tokens	Total
`Retrieval`	whether retrieval is needed	232; Q=96, U=136	[Retrieve] [No Retrieve]	69% 31%
`3-way Retrieval`	whether retrieval is needed or to use conv. history	210; Q=116, U=94	[Retrieve] [No Retrieve] [Continue to Use Evidence]	36% 21% 43%
`Relevance`	of retrieved passages given conv. history	301; Q=213, U=88	[Relevant] [Non Relevant]	41% 59%
`Groundness`	generated response is supported by retrieved passage	259; Q=229, U=30	[Fully supported] [Partially supported] [No support]	43% 13% 44%
`Utility`	usefulness of generated response	269; Q=143, U=126	[Utility:1] [Utility:2] [Utility:3] [Utility:4] [Utility:5]	34% 6% 9% 23% 29%

Table 1: Critic training data (QU-MTC) statistics. Q= and U= denotes number of instances sampled from QReCC and UltraChat, respectively.

retrieval is needed or not, and relevance of retrieved documents based on the entire conversation history. The critic tasks and special tokens are shown in Table 1.

Generator. The first task of the generator is to generate responses with the special reflection tokens. Given a conversation history as input x, we augment the response y to create $\hat{y}$ by including the reflection tokens that is generated by our trained critic model. The generator is trained to generate $\hat{y}$ given x using next token prediction objective. The second task of the generator is to summarise a conversational context to extract important aspects of the conversation *and* pose a question. Given a conversation history, when `SELF`-*multi*-`RAG` decides retrieval is necessary, it is further prompted to create a summary of the conversation which can be used as query to any retrieval model to obtain passages. We *do not* create a separate critic task for summarization. Rather, we *only* train the generator model since the summarization task is performed *whenever* retrieval is deemed necessary.

Retriever. The retriever is the third component of `SELF`-*multi*-`RAG` and can be used as a separate black-box component. In particular, we use 54M passages from Wikipedia and Common Crawl as the knowledge base.[1] We use an off-the-shelf Contriever model trained on MS-MARCO as the retriever.[2]. The retrieved passages are used by `SELF`-*multi*-`RAG` to generate responses during inference when it adaptively decides to call retrieval.

3.2 Overall Framework

Given a conversation history (that may also include passages retrieved in previous turns), the LLM (generator model) first decides whether retrieval

[1]The corpus is released as part of the QReCC dataset.

[2]`https://github.com/facebookresearch/contriever/`

is needed or not by generating one of the three special retrieval tokens: [Retrieve], [No Retrieve], [Continue to use evidence] based on the conversation history and previously retrieved passages. [Retrieve] is typically generated when the response needs to retrieve new facts to respond to a factual question. [No Retrieve] is generated when the question in the conversation requires the model to answer with a creative response. [Continue to use evidence] is typically generated when that the facts needed to answer the current factual question is already present in conversation history or the previously retrieved passages. So no new retrieval is needed and it can rely on the context to generate the response.

If retrieval is needed, SELF-*multi*-RAG then rewrites the conversational history which will be used as a query to retrieve passages from a corpus. The retriever retrieves K passages which the generator process in parallel and retrieves K different candidate outputs conditioned on the conversation history and the retrieved passages. SELF-*multi*-RAG then indicates the relevance of each passage to the conversation history by generating the special relevance tokens [Relevant] or [Non Relevant]. Following which SELF-*multi*-RAG judges whether the generated responses are [Fully supported], [Partially supported] or [No support] by the respective retrieved passages by generating the corresponding groundedness tokens. Finally, SELF-*multi*-RAG gives a usefulness score of [1-5] to the generated response using the utility special token.

The final response is selected out of the candidate responses, using the one which has the highest score in terms of its usefulness, groundedness and the relevance of the passage from which it was generated. Following Asai et al. (2023), we conduct a segment-level beam search (with the beam size=B) to obtain the top-B segment continuations and return the best sequence at the end of generation. The score of each sequence y with respect to passage d is updated with a score S that is the linear weighted sum of the normalized probability of each special token type:

$$\begin{aligned} S = p(y_t|x, d, y_{t-1}) &+ w_1 * S(\texttt{Relevance}) \\ &+ w_2 * S(\texttt{Groundedness}) \\ &+ w_3 * S(\texttt{Utility}) \end{aligned}$$

where x is the conversation history, y_{t-1} is the generated response so far and w_i are hyperparameters that can be tuned to enable custom behavior during inference.[3] $S(.)$ indicates the generation probability of the most desirable reflection token, e.g., [Relevant] in case of Relevance or [Fully Supported] in case of Groundedness.

SELF-*multi*-RAG is thus capable of comprehending longer conversational contexts (for the various critic tasks) than SELF-RAG and also summarizing the conversational context that when used as a query can improve retrieval effectiveness. The framework is depicted in Figure 3.

4 Experimental Setup

We evaluate SELF-*multi*-RAG on three benchmarks: QReCC (Anantha et al., 2020), UltraChat (Ding et al., 2023) and MT-Eval (Kwan et al., 2024).

QReCC contains conversational questions answers to which can be found within a collection of 10M web pages. Answers to questions in the same conversation may be distributed across several web pages. QReCC provides gold passage annotations which indicates the passage from where a question in a conversation can be answered from.

UltraChat is traditionally used as supervised fine tuning (SFT) data for LLMs. It contains diverse and informative instructional conversations and covers a wide range of topics and instructions. While conversations in QReCC are knowledge-grounded and ideally RAG should be beneficial for every turn of the conversation, it is not so for UltraChat.

MT-Eval is similar to UltraChat as it also contains diverse instructional conversations. It forms our out-of-domain evaluation benchmark as we use samples freom QReCC and UltraChat for training SELF-*multi*-RAG. Additional details on the benchmarks are provided in Appendix B.

4.1 Training Data

The training data for our critic and generator models are sampled from QReCC and UltraChat. We employ GPT-4 to collect the labels for training data for critic model. The prompt for collecting GPT-4 labels are provided in the Appendix A. We denote this training data as QReCC-UltraChat Multi-turn Critic Data (QU-MTC), details of which are provided in Table 1. Furthermore, we also create a single-turn variant (QU-STC) by flattening the conversation history to a single-turn using a T5 based

[3]We use default values defined by Asai et al. (2023).

Dataset	Model	`Ret. w/o P.`	`Ret. w P.`	`3-way Ret.`	`Relevance`	`Groundedness`	`Utility`
STC single-turn	Critic_o	0.58	-	0.52	0.66	0.58	0.69
	Critic_s	0.47	-	0.53	0.43	0.58	0.62
	Critic_m	0.50	-	0.52	0.62	0.40	0.68
	Critic_{sm}	**0.84**	-	**0.69**	**0.78**	**0.71**	**0.75**
QU-MTC: multi-turn	Critic_o	0.43	0.51	0.28	0.61	0.35	0.46
	Critic_s	0.68	0.28	0.62	0.72	0.55	0.45
	Critic_m	0.76	0.58	0.48	0.75	0.58	0.79
	Critic_{sm}	**0.83**	**0.77**	**0.63**	**0.77**	**0.61**	**0.80**

Table 2: Critic performance on self-reflection tasks on the test splits of each dataset. We report classification accuracy of predicting the correct special tokens for each task as outlined in Table 1. For the `Retrieval` task, we evaluate critic accuracy, either with passages in the conversation history (`Ret.` w P.) or without (`Ret.` w/o P.).

query rewriter T5QR (Lin et al., 2020). Thus QU-MTC and QU-STC come from the same data distribution, with the difference being the representation of the conversation history. Lastly, we also use the *original* single-turn critic training dataset, STC, released by Asai et al. (2023).

We employ our trained critic models to create training data for the generator using samples from QReCC and UltraChat (different from those sampled for training the critic), referred to as QReCC UltraChat Multi-turn Generator Data, QU-MTG. As in the case of the critic, we create a single-turn variant of the data, QU-STG, by rewriting the conversation history to a single-turn. Lastly, we sample data from both datasets to create the Conversation Summarization Data (CSD) where we prompt GPT-4 to generate ground truth summaries of ∼5000 conversations.[4]

4.2 Models

We train our own critic models from scratch using `mistralai/Mistral-7B-Instruct-v0.2` as initial checkpoint. The different versions of critic models are outlined in Table 3. Critic_o is equivalent to the original critic model trained by Asai et al. (2023) that is sampled from a number of single-turn benchmarks. However, to understand the impact of training in the conversational setting, Critic_{sm} must be compared with Critic_s as they are trained on the same data distribution.

We also train our generator model from the `mistralai/Mistral-7B-Instruct-v0.2` checkpoint. The different versions of generator models, that we use to compare the performance of SELF-*multi*-RAG with, are also outlined in Table 3. Henceforth, we use SELF-*multi*-RAG to refer to the final model that is trained end-to-end on single-turn, multi-turn conversation data and also trained to summarise conversational context and compare its performance against the other generator models.

Model	Critic	Training Data
Critic_o	-	STC
Critic_s	-	STC + QU-STC
Critic_m	-	QU-MTC
Critic_{sm}	-	STC + QU-MTC
SELF-RAG_s	Critic_s	QU-STG
SELF-RAG_{sm}	Critic_{sm}	QU-MTG
SELF-*multi*-RAG	Critic_{sm}	QU-MTG + CSD

Table 3: Configurations explored as training data for our approaches; ST = single-turn; MT = multi-turn; C = Critic, G = Generator. CSD refers to the conversation summarization dataset.

4.3 Evaluation

We first evaluate critic performance on the self-reflection tasks by calculating the accuracy of the predicted tokens described in Table 1. We use held out test split of STC and QU-MTC for evaluating critic. To measure the quality of the responses of the generator models, we employ both automatic metrics and human annotations. We collect GPT-4 and human evaluation scores that rate the responses generated by the models on a scale of 1-5 that includes different dimensions such coherence, understandability, and overall quality. The prompt for GPT-4 evaluation is provided in Appendix D. We employ BERTScore (Zhang et al., 2019) to measure similarity of generated response with ground truth response. To measure coherence of response given the conversation history and grounding of response given retrieved documents, we employ UniEval (Zhong et al., 2022).

We employ Amazon Mechanical Turk for human evaluation of generated responses. Crowd-

[4]The prompt for GPT-4 summary collection is provided in Appendix A.

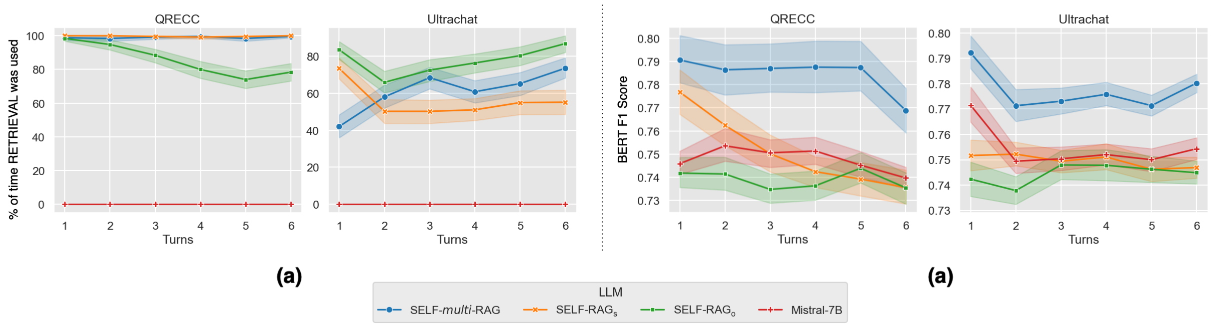


Figure 4: (a) Relation between retrieval calls and number of turns considered in QReCC and UltraChat. (b) Answer quality measured by BERTScore for different turn configurations.

Dataset	Model	GPT4	BERT	Coherence	Groundness*	Human
QReCC	`Mistral-7B-Instruct-v0.2`$^{-}$	2.72	0.75	**0.98**	-	4.08
	`SELF-RAG`†	1.60	0.74	0.90	0.77	3.50
	`SELF-RAG`$_s$†	1.51	0.75	0.96	0.78	3.71
	`SELF-`*multi*`-RAG`†	**4.15**	**0.80**	0.97	**0.87**	**4.20**
UltraChat	`Mistral-7B-Instruct-v0.2`$^{-}$	1.71	0.75	0.99	-	3.95
	`SELF-RAG`†	1.57	0.74	0.95	0.89	4.00
	`SELF-RAG`$_s$†	1.45	0.75	0.98	0.95	4.05
	`SELF-`*multi*`-RAG`†	2.70	**0.78**	0.99	**0.99**	**4.46**
MT-Eval	`Mistral-7B-Instruct-v0.2`$^{-}$	2.04	0.62	0.92	-	3.49
	`SELF-RAG`†	1.50	0.71	0.94	0.82	3.84
	`SELF-RAG`$_s$†	1.20	0.72	0.95	0.80	4.01
	`SELF-`*multi*`-RAG`†	**3.12**	**0.74**	**0.96**	**0.92**	**4.23**

Table 4: Performance of response generation models on the three dataset. QR = rewritten conversation history as context; FC = full conversation history as context. † indicates adaptive retrieval; $^{-}$ indicates no retrieval. * We measure groundness with retrieved documents only for the cases when the model decides to call retrieval.

workers annotate the quality of the generated responses on a scale of 1 to 5 on the dimensions of coherence, engagingness, and understandability. We collect scores from 3 annotators and aggregate the score using majority voting for each of those dimensions. The overall score is the average of the scores across the three dimensions. In order to ensure high-quality human judgment, we use several mitigation strategies such as simplified task setups, clear annotation guidelines, and time checks to exclude potential spammers. Further details of the human annotation guidelines and disaggregated scores are provided in Appendix E.

5 Results

Critic Performance. Table 2 shows that the critic model trained on both single and multi-turn data Critic$_{sm}$ has overall the best accuracy (generating *correct* reflection tokens) on the critic tasks (based on GPT 4 labels), even improving in the single-turn setting. This suggests that Critic$_{sm}$ is better at handling longer context of conversations while judging whether retrieval is needed or not, judging relevance of retrieved passages, and utility of an answer.

Response Generation. As shown in Table 4, `SELF-`*multi*`-RAG`, trained on data created by Critic$_{sm}$, leads to improvement on all conversational benchmarks according to the evaluation metrics. Comparing its performance with `SELF-RAG`$_s$, that has been trained on data from the same dataset but using single-turn contexts, `SELF-`*multi*`-RAG` performs better in comprehending the conversational context. The decision of whether retrieval is required is more accurate for `SELF-`*multi*`-RAG` as compared to the baselines. As evidence, we see Critic$_{sm}$ has higher accuracy on retrieval tasks (both with and without passages included in the conversation history) than Critic$_s$. Figure 4 shows that `SELF-`*multi*`-RAG` decides to call retrieval $\sim$ 100% of time for QReCC, however not so for UltraChat. This is the expected behaviour as conversations in QReCC are mostly knowledge grounded, whereas in UltraChat there are more instructional conversations that not always require retrieved knowledge. This suggests that the decision to call retrieval or not is indeed important for conversational QA and adapting the model to better handle conversational context is beneficial. Figure 4 further shows that SELF-*multi*-RAG generated responses are better at all turns (upto 6) of the conversations further providing evidence to its ability to understand long conversational context. Some of the cases where SELF-*multi*-RAG called retrieval and could not provide a satisfactory answer is typically the cases where it could not find relevant answers within the retrieved documents. Groundedness of the generated response to the conversational context and retrieved passages, as measured using UniEval, is also higher for SELF-*multi*-RAG as compared to its single-turn counterparts. Lastly, we see SELF-*multi*-RAG perform the best on MT-Eval, indicating strong performance on held-out conversational benchmarks.

5.1 Summarizing Conversations

Since QReCC provides ground truth labels of relevant passages (to a conversational context), we use it to evaluate retrieval effectiveness of different representations of the conversational context. Table 5 shows that summaries generated by our approach perform better than rewrites in the form of single-questions (using T5QR) in case of both sparse and dense retrievals. This is in line with research that show expanding queries and documents help in improving retrieval effectiveness (Ayoub et al., 2024; Mackie et al., 2023; Nogueira et al., 2019).

To better understand observed superior performance of SELF-*multi*-RAG as compared to its other variants, we perform ablations to narrow down the causes of gain. Table 6 reports the performance of different conversation history representations as query to retrieve relevant passages. Note that when we use the SELF-*multi*-RAG generated summary as a query to the retrieval model, the response generation is the best for both datasets. Other forms of conversation context as query representations (e.g., T5QR) have lower performance. Overall, SELF-*multi*-RAG improves in two directions, (i) it generates summaries as query with better retrieval effectiveness, and (ii) enhances response generation quality taking into account more suitable retrieval knowledge and conversational context.

Retrieval	Conv. representation	R@5	R@10
BM25	Full conv.	0.50	0.58
	Gold Rewrite	0.50	0.61
	T5QR Rewrite	0.45	0.55
	SELF-*multi*-RAG Summary	0.56	0.66
	GPT4 Summary	**0.60**	**0.70**
Contriever	Full conv.	0.53	0.61
	Gold Rewrite	0.58	0.69
	T5QR Rewrite	0.53	0.64
	SELF-*multi*-RAG Summary	0.61	0.71
	GPT4 Summary	**0.62**	**0.72**

Table 5: Retrieval effectiveness of different conversational context representation. GPT-4 summaries are the ground truth summaries that we collect for training SELF-*multi*-RAG for the summarization task.

5.2 Handling Previous-turn Retrieved Passages

Returning to Table 2, we also explore the value of previously retrieved passages into the context history of the input for the critic model in the `Retrieval` critic task. We evaluate critic accuracy, either with previous retrieved passages in the conversation history (`Ret.` w P.) or without (`Ret.` w/o P.). For QReCC, we include ground truth passages of previous turn in the context. For UltraChat, we sample instances where passages are present as part of a question in a conversation. As shown in Table 2, Critic$_{sm}$ performs the best in `Ret.` w P., where the model has to judge whether retrieval is needed or not given both the conversation history and passages retrieved in previous turn. This indicates superior ability of Critic$_{sm}$ to comprehend not only the conversation history but also previously retrieved passages to deem the necessity of retrieval in a multi-turn setting. Overall, the critic models have lower performance in this task compared to the *Retrieval without passages* (`Ret.` w/o P.) indicating the increased difficulty of the task.

Table 7 compares response generation performance with and without retrieved passages from previous turns are included in the input together with the conversation history. We observe that SELF-*multi*-RAG not only (correctly) decides to call retrieval less number of times, as compared to its single-turn baseline, but is also better at generating responses when the conversation context is composed of both the dialogue and previously retrieved passages indicating its ability to comprehend more complex contexts. Moreover, either configuration (`Ret.` w P., `Ret.` w/o P.) can be chosen based on the desired balance between efficiency and accu-

Dataset	Conv. Representation	GPT4	BERT
QReCC	Full	3.75	0.786
	T5QR	3.52	0.782
	SELF-*multi*-RAG summ.	**4.15**	**0.808**
UltraChat	Full	2.23	0.756
	T5QR	2.12	0.751
	SELF-*multi*-RAG summ.	**2.70**	**0.780**
MT-Eval	Full	2.65	0.720
	T5QR	2.52	0.712
	SELF-*multi*-RAG summ.	**3.12**	**0.743**

Table 6: Response generation quality of SELF-*multi*-RAG with different conversation history representations as query to retrieve relevant passages.

Dataset	Model	Ret.?	Groundness	BERT
QReCC	SELF-RAG$_s$ w/o P.	99%	0.78	0.75
	SELF-RAG$_s$ w P.	99%	0.73	0.74
	SELF-*multi*-RAG w/o P.	98%	0.87	0.80
	SELF-*multi*-RAG w P.	**54%**	0.79	0.78
UltraChat	SELF-RAG$_s$ w/o P.	56%	0.95	0.75
	SELF-RAG$_s$ w P.	67%	0.91	0.73
	SELF-*multi*-RAG w/o P.	62%	0.99	0.78
	SELF-*multi*-RAG w P.	**37%**	0.97	0.75
MT-Eval	SELF-RAG$_s$ w/o P.	60%	0.80	0.72
	SELF-RAG$_s$ w P.	69%	0.78	0.70
	SELF-*multi*-RAG w/o P.	58%	0.92	0.74
	SELF-*multi*-RAG w P.	**40%**	0.90	0.73

Table 7: Comparison of response generation performance with and without passages included in the context. Ret? = % of instances retrieval was used.

racy; `Ret.` w/o P. is useful when performance is prioritized over efficiency, whereas `Ret.` w/o P. is suitable when efficiency is crucial and a slight reduction in performance is acceptable. We present examples when passages are included in the context in the Appendix F.

6 Conclusion

In this work, we propose SELF-*multi*-RAG, a framework to train LLMs to learn when to retrieve and generate response for better conversational QA. We perform extensive evaluation on three conversational QA benchmarks and demonstrated improved performance over previous approaches. This is achieved by overcoming the previous limitations to accurately critic when to retrieve or whether the retrieved documents are relevant or the usefulness of generated response given a multi-turn dialogue. SELF-*multi*-RAG is better at comprehending the longer contexts of multi-turn conversations resulting in better critic and consequently generator performances. Finally, we observe that summaries of conversational history generated by SELF-*multi*-RAG, increase retrieval effectiveness when used as query to retrieve passages and consequently leads to improved response generation. As future work, we would like to (i) consider longer and multi-threaded conversations, (ii) including diversity as a metric while considering passages for response generation.

7 Limitations

In this paper, we propose an approach to enhance the ability of retrieved augmented models on conversational settings. While this is not specific to any particular language, we conducted all of our experiments and analysis exclusively on English-language QA datasets. Hence, this paper does not offer insights into the range of style variations found in non-English and datasets, nor does it ascertain the generalizability of our findings to other datasets and domains. Second, we limit our experiments using one model for critic and generator, `mistralai/Mistral-7B-Instruct-v0.2` and one retrieval model, Contriever. Extending SELF-*multi*-RAG to other models is left for future work. Finally, we perform retrieval in an offline manner for reducing computation overhead. In a more realistic scenario, retrieval will be performed online during response generation.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing evolution of multi-turn conversations for effective answer retrieval. In *Proceedings of the 2020 conference on human information interaction and retrieval*, pages 33–42.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Michael Antonios Kruse Ayoub, Zhan Su, and Qiuchi Li. 2024. A case study of enhancing sparse retrieval using llms. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1609–1615.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Etsuko Ishii, Bryan Wilie, Yan Xu, Samuel Cahyawijaya, and Pascale Fung. 2022. Integrating question rewrites in conversational question answering: A reinforcement learning approach. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 55–66.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. Itercqr: Iterative conversational query reformulation with retrieval guidance. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8114–8131.

Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023. Towards better instruction following language models for chinese: Investigating the impact of training data and evaluation. *arXiv preprint arXiv:2304.07854*.

Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2024. Robust training for conversational question answering models with reinforced reformulation generation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 322–331.

Mandar Kulkarni, Praveen Tangarajan, Kyung Kim, and Anusua Trivedi. 2024. Reinforcement learning for optimizing rag for domain chatbots. *arXiv preprint arXiv:2401.06800*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative relevance feedback with large language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2026–2031.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023a. Large language models know your contextual search intent: A prompting framework for conversational search. *arXiv preprint arXiv:2303.06573*.

Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023b. Learning denoised and interpretable session representation for conversational search. In *Proceedings of the ACM Web Conference*, pages 3193–3202.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645*.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to doctttttquery. *Online preprint*, 6(2).

Philhoon Oh and James Thorne. 2023. Detrimental contexts in open-domain question answering. *arXiv preprint arXiv:2310.18077*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Ante Wang, Linfeng Song, Ge Xu, and Jinsong Su. 2023. Domain adaptation for conversational query production with the rag model feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9129–9141.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2021. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. *arXiv preprint arXiv:2112.08558*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. *arXiv preprint arXiv:2310.09716*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Appendix

A GPT-4 Prompt: Critic Training Data & Conversation Summarization Training Data

Prompts to collect critic training data for the different tasks using GPT-4 are outlined in Table 8 (`Retrieval`), Table 9 (`3-way Retrieval`), Table 10 (`Relevance`), Table 11 (`Groundedness`), Table 12 (`Utility`), We modify the prompts of Asai et al. (2023) to be compatible with multi-turn dialogues (as compared to single-turn QA). Furthermore, we collect conversation summarization training data for our generator model using GPT-4 prompt detailed in Table 13.

B Benchmark Statistics

While QReCC is traditionally used as the primary benchmark for evaluating conversational query rewriting models or response generation, we note a few differences with UltraChat and MT-Eval. Firstly, as mentioned earlier, responses to all conversation turns are grounded in passage from Wikipedia/Common Crawl. Hence, ideally retrieval should be needed at every turn, and hence RAG should lead to the best response. Whereas in the other two datasets there are examples where a passage is already included in a conversation turn or ones where the response to a turn should be based on response generated in previous turns or conversations where users ask creative questions (examples provided in Table 14). Hence, in these two datasets the decision to call retrieval is more important where a model should understand from the context whether retrieval is needed or not. Secondly, questions and gold responses in QReCC are an order of magnitude shorter than UltraChat as shown in Table 14).

C Implementation Details

We use `Mistral-7B-Instruct-v0.2` as our critic an generator models. Critic training is performed using 4 nodes, each containing 8 NVIDIA A10G Tensor Core GPUs. The models are trained for 3 epochs, learning rate of $2e^{-5}$ and warm up ratio of 0.01. Generator training is performed using 5 nodes, 2 epochs, learning rate of $2e^{-5}$ and warm up ratio of 0.03. The distributed training is performed using deepspeed[5]. For computing BERT-Score, we use `roberta-large-mnli` as our model.

[5] https://github.com/microsoft/DeepSpeed

`Retrieval`

Given ***a conversation history***, please make a judgment on whether finding some external documents from the web (e.g., Wikipedia) helps to generate a better response. Please answer [Retrieval] or [No Retrieval].

Conversation History
How did the Boer war start?
Many historians stress that in reality the contest was for control of the rich Witwatersrand gold-mining complex located in the SAR.
What were the Boer Commandos?
Rating: [Retrieval]
Explanation: Retrieving documents will help in generating a good response to the conversation.
Conversation History
How does the taste and texture of swordfish change depending on how it's prepared, and what are some popular cooking methods?
The taste and texture of swordfish can vary greatly depending on how it is prepared. Here are some popular cooking methods and how they affect the taste and texture of the fish: 1. Grilling: Grilling swordfish gives it a smoky flavor and a slightly charred texture. The meat becomes firmer and less flaky. 2. Broiling: Broiling swordfish can give it a crispy exterior while keeping the inside moist and tender. 3. Baking: Baking swordfish at a high temperature can create a crust on the outside of the fish while keeping the inside moist and tender. 4. Pan-searing: Searing swordfish in a hot skillet with oil can give it a crispy exterior while keeping the inside moist and flaky. 5. Poaching: Poaching swordfish in a liquid such as wine, lemon juice or stock can add flavor to the fish and keep it moist and tender. 6. Raw: Swordfish can also be served raw as sushi or sashimi. Raw swordfish has a firm texture and a mild, sweet flavor.
I've only ever had swordfish grilled, what other methods do you recommend trying?
Rating: [No Retrieval]
Explanation: The response doesn't require any external evidence as it can be generated based on the conversation history.

Conversation History
when did Joe Satriani start his music career?
Satriani started playing in a San Francisco-based band called the Squares what instrument does Joe Satriani play?
He was inspired to play guitar at age 14, after hearing of the death of Jimi Hendrix.
what year did he start his music career?
GPT-4-Rating: [Retrieval]

Table 8: GPT-4 prompt for collecting training data for Retrieval critic task.

3-way-Retrieval

You will be provided with ***a conversation history***, evidence, response to the ***conversation***, and preceding sentences (optional). If the preceding sentence is given, the response should be the sentence that follows those preceding sentences. Your task is to determine whether the information in the response sentence can be fully verified by the evidence or the ***conversation history***. There are three cases:
- If the response can be verified solely with the evidence or the ***conversation history***, then respond with [Continue to Use Evidence].
- If the sentence doesn't require any factual verification (e.g., a subjective sentence or a sentence about common sense), then respond with [No Retrieval].
- If additional information is needed to verify the output sentence, respond with [Retrieval]. Please provide explanations for your judgments.

Conversation History
Given the text: that truth might finally win. The candidates deceive us all. in jets and campaign buses. The smiles they wore were painted on. So sly, those hungry foxes. on top of their soapboxes. "Hey, I'll do much more taxing. so you can be relaxing." I'll give you fruits of their hard work. "You've heard me, one and all! We need a woman president! We need a leader NOW! We need to have a first "first man" to be since smoking pot in college with Bill.. because I JUST WANT TO BE PRESIDENT!! We do not need a woman. but someone true who can. for answers to be found. so maybe we should look. So why is he not jailed? Now here's another "great" debate. What difference does it really make? It tells me this: Who cares?
Can you summarize the text material to describe the main message and theme it conveys?
The text criticizes political candidates and their deceitful tactics during campaigns, highlighting their insincerity and lack of concern for the public. The message suggests a need for a truthful and effective leader to bring about change and progress, rather than focusing on superficial qualities such as gender or past indiscretions. The theme centers around the importance of honesty and integrity in leadership
Can you provide examples from the text that show the candidates' insincerity and deceitful tactics during campaigns?
Preceding sentences: Here are a few examples from the text that demonstrate the candidates' insincerity and deceitful tactics during campaigns.
Evidence: The charisma of the sender of a message may affect how the message is received. Political candidates are often chosen more for their possession of this quality than for their other attributes. A charismatic person can often make tired, trivial messages seem new and important to the recipient; however, this too can become detrimental to communication, as the receiver of the message is less likely to question or ask for clarification of the message.
Response: 'The smiles they wore were painted on' suggests that the candidates are not genuinely happy or friendly, but are instead putting on a façade to deceive the public. - 'So sly, those hungry foxes' implies that the candidates are cunning and opportunistic, willing to say or do whatever it takes to win. - 'Hey, I'll do much more taxing. So you can be relaxing.' This statement is a classic political promise that is often made but rarely kept, highlighting the candidates' tendency to make unrealistic claims in order to garner support.
Rating: [Continue to Use Evidence]
Explanation: The response can be generated solely using the conversational history.

Conversation History
In a small sauce pot on low heat, melt butter and add the vegetable oil. Add onions and cook on low heat until they are browned and soft. Approximately 20 minutes. Add sugar and cook for 2 minutes. Stir in vinegar and stock and cook for an additional 10 minutes or until gravy has reduced by half. Stir in mustard and pepper. Season to taste with salt. Remove from heat and using a hand blender, blend until smooth. Is there a substitute for vegetable oil that can be used in this recipe?
Yes, you can use any other type of neutral oil such as canola oil, grapeseed oil, or sunflower oil as a substitute for vegetable oil in this recipe.
Can you give me some tips on how to make sure the onions are properly browned and soft?
Preceding sentences: Sure! Here are some tips to ensure your onions are browned and soft: 1. Use a low heat: Cooking the onions on low heat allows them to slowly caramelize and become sweet and tender. 2. Stir regularly: Stirring the onions regularly will help prevent them from burning and ensure that they are evenly browned. 3. Use a heavy-bottomed pot: A heavy-bottomed pot ensures even heat distribution and helps prevent burning. 4. Don't overcrowd the pot: Overcrowding the pot will cause the onions to steam instead of caramelize, so make sure there is enough room for them to spread out.
Evidence: Palm oil adds some great firmness to your soap, and there is no direct replacement for it. Grapeseed and sunflower oil are both liquid at room temperature. They're considered soft oils, and will make this recipe softer than if it had palm oil. You can leave palm oil out.
Response: 5. Be patient: Properly caramelizing onions takes time, so don't rush the process.
GPT-4-Rating: [No Retrieval]

Table 9: GPT-4 prompt for collecting training data for Multi-Retrieval critic task.

Relevance
You'll be provided with ***a conversation history***, along with an evidence.Your job is to determine if the evidence is relevant and provides useful information to generate the response of the given ***conversation history***. If the evidence meets this requirement, respond with [Relevant]; otherwise, generate [Irrelevant].

Conversation History:
How did the Boer war start?
Many historians stress that in reality the contest was for control of the rich Witwatersrand gold-mining complex located in the SAR.
What were the Boer Commandos?
Evidence: Boer Commando Not to be confused with Commando System (South Africa) or Kommandokorps. The Boer commandos or " Kommandos " were volunteer military units of guerilla militia organized by the Boer people of South Africa . The term came into English usage during the Second Boer War of 1899-1902. Boer Commando in action during the First Boer War , 1881 In 1658, war erupted between the Dutch settlers at Cape Colony and the Khoi-khoi . In order to protect the settlement, all able bodied men were conscripted. After the conclusion of this war, all men in the colony were liable for military service and were expected to be ready on short notice.
Rating: [Relevant]
Explanation: The evidence explicitly talks about Boer commandos from which the response to the conversation can be generated.
Conversation History:
What was the origin of the Olmec?
The beginnings of Olmec civilization have traditionally been placed between 1400 and 1200 BCE. It seems that the Olmec had their roots in early farming cultures of Tabasco.
What can you tell me about the Olmec at El Manati?
Past finds of Olmec remains were ritually deposited at El Manati shrine.
How did they start?
Evidence: It is a theory that according to many, could explain the incredible technologies and skills of this enigmatic Ancient Civilization. Even though the Olmec civilization is surrounded by numerous mysteries, researchers believe that all the classical cultures of Mesoamerica originated from this mysterious civilization. But where did this ancient civilization originate? And why is it that we know so little about one of the most influential ancient civilizations of Mesoamerica.
Rating: [Irrelevant]
Explanation: Although the evidence talks about Olmecs, they do not provide information as to how the Olmec civilisation started.

Conversation History
Where did J. R. Jayewardene live in his early life?
J. R. Jayewardene was born in Colombo, British Ceylon.
Where did he go to school?
Evidence: Highly respected in Japan for his call for peace and reconciliation with post-war Japan at the Peace Conference in San Francisco in 1951, a statue of Jayewardene was erected at the Kamakura Temple in the Kanagawa Prefecture in Japan in his honor. [18] J.R Jayewardene Centre [edit] In 1988, the J.R. Jayewardene Centre was established by the J.R Jayewardene Centre Act No. 77 of 1988 by Parliament at the childhood home of J. R. Jayewardene Dharmapala Mawatha, Colombo. It serves as archive for J.R Jayewardene's personal library and papers as well as papers, records from the Presidential Secretariat and gifts he received in his tenure as
President.
GPT-4-Rating: [Irrelevant]

Table 10: GPT-4 prompt for collecting training data for Relevance critic task.

Groundedness
You will receive ***a conversation history***, evidence, and a response to the ***conversation***. Your task is to evaluate if the response is fully supported by the information provided in the evidence or ***in the conversation history***. Use the following entailment scale to generate a score: [Fully supported] - All information in output is supported by the evidence, or extractions from the evidence or **the conversation history**. [Partially supported] - The response is supported to some extent, but there is major information in the response that is not discussed in the evidence or the conversation history. [No support / Contradictory] - The response completely ignores, is unrelated to, or contradicts the evidence and the conversation history. This can also happen if the evidence is irrelevant to the conversation history. Make sure to not use any external information/knowledge to judge whether the response is true or not.
Conversation History I was thinking of buying a cheesecake, can you tell me some information about them? Sure! Cheesecakes are actually my speciality. Usually it is a baked dessert but it can also be unbaked. I had no idea you didn't have to bake them, is there any interesting history behind the cheesecake? Forms of it go back all the way to greece! That's fascinating, can you tell me more about the Greek cheescake origins? Sure. The earliest mentions of it were in a Greek book. Essentially it was a cookbook about the art of making cheesecakes I wonder why they were so fond of them. Is there a traditional recipe for modern cheesecakes? **Response**: My favourite layer is the biscuit base, particularly if ginger biscuits are mixed in with the graham crackers. **Evidence**: Cheesecakes, having a crust that is separately prepared and baked. A more modern version is found in Forme of Cury', an English cookbook from 1390. On this basis, chef Heston Blumenthal has argued that cheesecake is an English invention. Cheesecake did not evolve into the dessert that we see today up until somewhere around the 18th century. Europeans began removing yeast and adding beaten eggs to the cheesecake instead. **Rating**: [No support / Contradictory] **Explanation**: The response is neither supported by the evidence or the conversation history **Conversation History** What does it mean to be a vegan? Veganism is a stricter form of vegetarianism; like vegetarians, vegans don't eat any animals that have been killed and also avoid animal products altogether, so that means no eggs or dairy products and, usually, no honey either. What is the environmental argument for being a vegan? Studies on world food security estimate that an affluent diet containing meat requires up to 3 times as many resources as a vegetarian diet Do you need milk for strong bones? If you want to drink milk for strong bones, I recommend no more than one glass a day What are other good sources of calcium **Response**: However, many non-dairy sources are also high in this mineral. These include seafood, leafy greens, legumes, dried fruit, tofu and various foods that are fortified with calcium. **Evidence**: Top 15 Calcium-Rich Foods (Many Are Non-Dairy) Subscribe Nutrition Evidence Based Top 15 Calcium-Rich Foods (Many Are Non-Dairy) Written by Kerri-Ann Jennings, MS, RD on July 27, 2018 If you buy something through a link on this page, we may earn a small commission. How this works. Calcium is very. **GPT-4-Rating**: [Partially supported]

Table 11: GPT-4 prompt for collecting training data for Groundedness critic task.

Utility

Given a conversation history and a response, rate whether the response appears to be a helpful and informative answer to the query, from 1 (lowest) - 5 (highest). We call this score perceived utility.The detailed criterion is as follows:
5: The response provides a complete, highly detailed, and informative response to the conversation, fully satisfying the information needs.
4: The response mostly fulfills the need in the conversation, while there can be some minor improvements such as discussing more detailed information, having better structure of the response, or improving coherence.
3: The response is acceptable, but some major additions or improvements are needed to satisfy users' needs.
2: The response still addresses the main request, but it is not complete or not relevant to the conversation.
1: The response is barely on-topic or completely irrelevant.

Conversation History
What was snake river canyon for Evel Knievel?
Knievel hired subcontractor and aeronautical engineer Doug Malewicki to build him a rocket-powered cycle to jump across the Snake River, and called it the Skycycle X-1.
What was the Evel Knievel's snake river canyon jump a launch of?
The decision was then made to have Truax build the Skycycle X-2 and have it take off and fly more like a rocket than a motorcycle.
When was the launch?
Response: It was launched on September 7, 1974.
Perceived utility: 5
Explanation: The response is accurate, brief, and directly addresses the user's question. It provides the exact date of the launch, which is the specific information that the user was asking for.
Conversation History:
How do elephants use their trunks for communication and sensing their environment?
Elephants use their trunks for communication and sensing their environment in several ways: 1. Smelling: Elephants have a highly developed sense of smell and can detect scents from miles away. They use their trunks to smell for food, water, mates, predators, and other elephants. They can also smell the scent left by other elephants as a means of identifying them. 2. Touch: Elephants use their trunks to touch and feel their environment. They can use their trunks to brush away leaves or branches to get a clear view of their surroundings, and they can touch other elephants to communicate their mood or intentions. 3. Vocalizations: Elephants can produce a wide range of vocalizations, from low rumbles to high-pitched trumpets. These sounds are produced by muscles in their trunks and are used for communication with other elephants.
I wonder if elephants have ever been observed using their trunks to communicate with other animals besides other elephants?
Response: I'm not sure, but elephants are known to be highly intelligent and social animals, so it's possible that they have developed complex communication systems with other animals as well.
Perceived utility: 2
Explanation: The assistant did provide some relevant information about elephants being highly intelligent and social animals, but doesn't fully satisfy the user's need for a definite answer.

Conversation History:
What was the relation between Dimebag and Darrell?
Dimebag Darrell was an American musician and songwriter.
He was the guitarist of the heavy metal bands Pantera and Damageplan.
When was the band Pantera formed?
Pantera was formed in 1981.
What was the purpose of the band?
Response: Vinnie Paul was asked to join a band alongside his high school classmates Terry Glaze (guitar), Tommy Bradford (bass) and Donny Hart (vocals).
GPT-4-Rating: 2

Table 12: GPT-4 prompt for collecting training data for Utility critic task.

Conversation summarization
Given a conversation history, your task is to summarise the conversation history in 40-50 words and ask a question so that the summary and the question can be used without the conversation history to generate a meaningful response.
Converation History: What was the first job John Sherman Cooper held? He was admitted to the bar by examination in 1928 and opened a legal practice in Somerset. What was the first office John Sherman Cooper ran for? After being urged into politics by his uncle, Judge Roscoe Tartar, Cooper ran unopposed for a seat in the Kentucky House of Representatives as a Republican in 1927. How long was John Sherman Cooper in office in the Kentucky House of Representatives? Member of the Kentucky House of Representatives from the 41st district. In office, 1928–1930 Did he run for another political office after that? **Summary:** John Sherman Cooper started his career as a lawyer in Somerset after being admitted to the bar in 1928. He was later encouraged by his uncle, Judge Roscoe Tartar, to join politics and subsequently ran for a seat in the Kentucky House of Representatives as a Republican candidate in 1927. He went unopposed and served in office from 1928-1930. **Question**: Did John Sherman Cooper pursue any other political offices after his term in the Kentucky House of Representatives? **Converation History**: When did Sachin Tendulkar first join a team? On 14 November 1987, Sachin Tendulkar was selected to represent Bombay in the Ranji Trophy, India's premier domestic First-class cricket tournament, for the 1987–88 season. Was he successful with that team? **Summary**: Sachin Tendulkar first joined the Bombay team in the Ranji Trophy, India's premier domestic First-class cricket tournament, on November 14, 1987. **Question**: Was Sachin Tendulkar's performance successful with this team?

Table 13: GPT-4 prompt for collecting generator training data for conversation summarization.

	QReCC	Ultrachat	MT-Eval
Avg. # words in one question	7	19	32
Avg. # words in response	67	208	144
Example conversation	**User:** which is the best jekyll and hyde film?	**User:** Provide a detailed analysis of the role of artificial intelligence in enhancing customer experience within the hospitality industry.	**User:** What is the central dogma of molecular biology? What processes are involved? Who named this?
	Assistant: The 1931 film adaptation of Robert Louis Stevenson's horror novella is considered to be the best Jekyll an Hyde film of all time.	**Assistant:** Artificial intelligence (AI) has been playing an increasingly important role in enhancing customer experience in various industries, including hospitality. Hospitality businesses such as hotels, restaurants, and travel companies are using AI to improve their customer service, streamline processes and gain a competitive edge. Below are some of the ways AI is transforming customer experience in the hospitality industry: 1. Personalized recommendations and experiences AI-powered systems can analyze guest data such as booking history, preferences, demographics, and behavior to provide personalized recommendations and experiences. For instance, a hotel can use AI to suggest personalized room upgrades, recommend nearby attractions, or offer personalized dining recommendations based on guests' dietary requirements. 2. Chatbots and virtual assistants .... In summary, AI is transforming the hospitality industry by enhancing customer experience, increasing efficiency, and reducing costs. With the implementation of AI, businesses can provide personalized experiences, streamline processes, and gain a competitive edge. As AI technology continues to improve, it is likely to become an essential aspect of the hospitality industry.	**Assistant:** The central dogma of molecular biology is a fundamental principle that describes the flow of genetic information within a biological system. It was first proposed by Francis Crick in 1958. The central dogma states that genetic information flows from DNA to RNA to proteins, which are responsible for the structure and function of cells. ... In summary, the central dogma of molecular biology, proposed by Francis Crick, describes the flow of genetic information from DNA to RNA to proteins, involving the processes of DNA replication, transcription, and translation.
	User: when was strange case of dr jekyll and mr hyde published	**User:** I love the idea of personalized recommendations and experiences. Can you give me an example of a hotel that effectively uses AI for this purpose?	**User:** Identify and fix one incorrect fact in your previous response.
	Assistant: Strange Case of Dr Jekyll and Mr Hyde is a gothic novella by Scottish author Robert Louis Stevenson, first published in 1886.	**Assistant:** Sure! One example of a hotel that effectively uses AI for personalized recommendations and experiences is the Henn-na Hotel in Japan. This hotel uses robot receptionists and facial recognition technology to provide personalized experiences for their guests. ... Overall, the use of AI in Henn-na Hotel provides a seamless and personalized hospitality experience that's entirely automated while maximizing customer service.	**Assistant:** Apologies for the mistake in my previous response. The central dogma of molecular biology was first proposed by Francis Crick in 1957, not 1958. The rest of the information provided remains accurate.
	User: what other books did the author write"	**User:** Wow, that sounds really futuristic. I'd love to stay at that hotel and experience all of their AI-powered amenities. Do you know if more hotels are adopting these technologies?	**User:** Can you elaborate on how DNA replication in stem cells differs from that in other types of cells, particularly in the context of its accuracy and regulation?

Table 14: Statistics and examples for the three datasets considered in this study.

D GPT-4 Evaluation Form

We employ GPT-4 to evaluate the responses generated from the different SELF-RAG models. We set temperature to be 1.0 and maximum number of generated tokens (`max_tokens`) to be 512. The prompt for GPT-4 evaluation is detailed in Table 15.

E Mechanical Turk Setup

We conduct human evaluations using Amazon Mechanical Turk to determine the quality of the generated answers based on coherence, engagingness and understandability. We give annotators fair compensation. We also use a bonus incentive structure. Every worker who passes the automatic quality checks receives a bonus at the end. In addition, we only consider workers from a country whose main language is English, who has completed 100 or more HITs so far with an acceptance rate of 95% or higher. Figure 5 shows the template for evaluating the generated response with respect to the question and conversational history. In Table 16, we provide the scores of generated response along the dimensions of coherence, engagingness and understandability as rate by the human annotators.

F Examples of Generated Responses

In Table 17, we provide an example where SELF-*multi*-RAG comprehends longer conversational contexts better than its single-turn counterpart.

Prompt for GPT-4 Evaluation

Conversation History:
`conversation`
The Last User Question:
`question`
The Start of Assistant's Answer
`generated_response`
The End of Assistant's Answer
System:
Given the conversation history given above, we would like to request your feedback on the performance of the assistant in response to the **Last User Question** as displayed above. Please rate the quality, helpfulness, level of details, and relevance of the assistant's answer to the conversation.
Use your judgement considering factors such as informativeness, satisfaction, readability, ease of understanding etc. You can also base your judgement on the naturalness, factuality, correctness, usefulness and objectiveness of the answer.
The assistant receives an overall score on a scale of 0 to 5, where a higher score indicates better overall performance. Please output "Score: an integer number between 0 and 5". In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias.

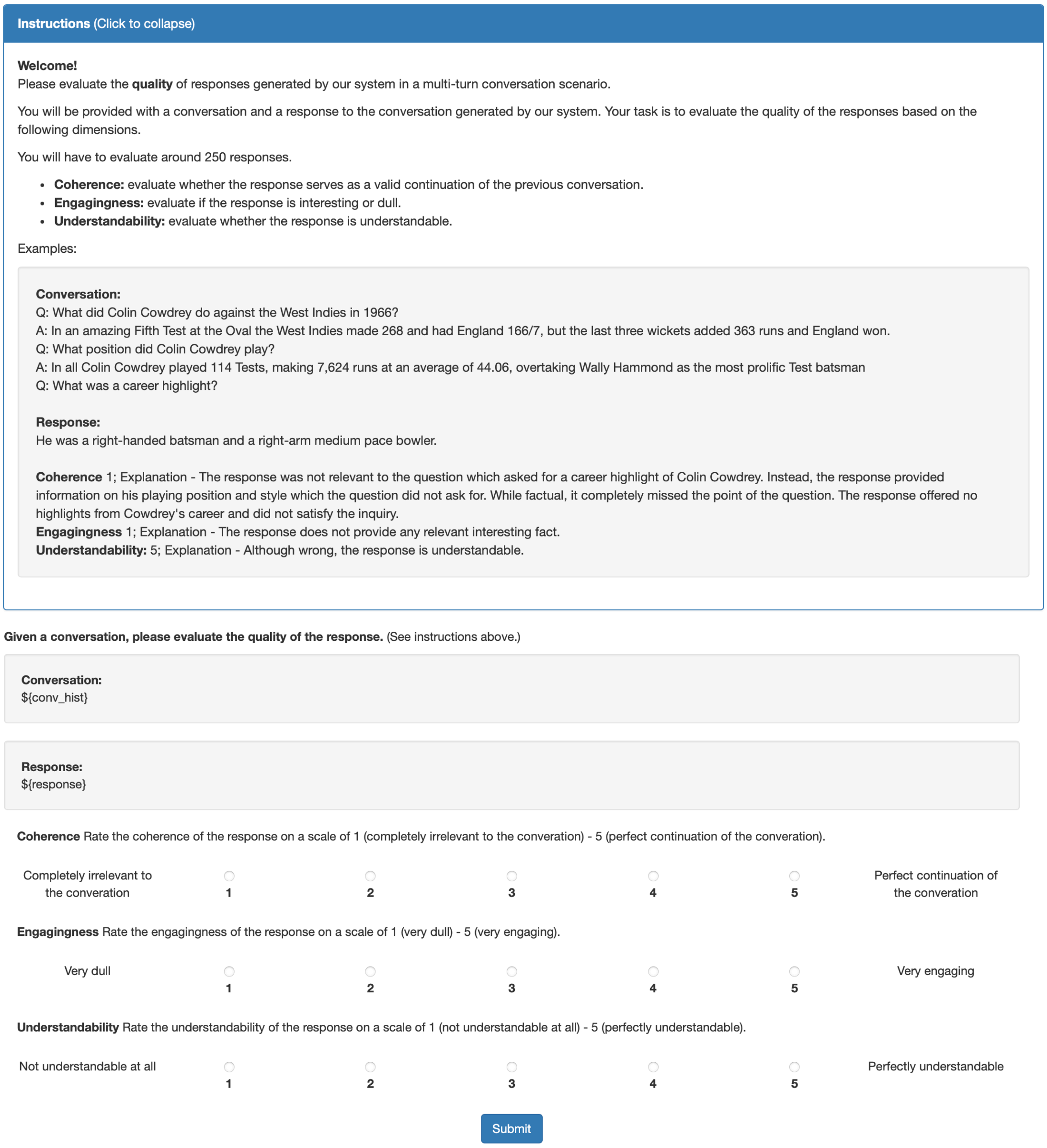
Table 15: Prompt for GPT-4 evaluation of generate responses.

Benchmark	Model	Coherence	Engagingness	Understandability
QReCC	`Mistral-7B-Instruct-v0.2`$^{-}$	4.26	**3.17**	4.78
	SELF-RAG†	3.91	2.21	4.52
	SELF-RAG$_{s}$†	4.26	2.3	4.72
	SELF-*multi*-RAG†	**4.75**	2.88	**4.89**
Ultrachat	`Mistral-7B-Instruct-v0.2`$^{-}$	4.20	3.20	4.70
	SELF-RAG†	4.03	2.63	4.59
	SELF-RAG$_{s}$†	4.39	2.55	4.72
	SELF-*multi*-RAG†	**4.81**	**3.20**	**4.85**
MT-Eval	`Mistral-7B-Instruct-v0.2`$^{-}$	4.13	3.08	4.62
	SELF-RAG†	4.08	2.64	4.58
	SELF-RAG$_{s}$†	4.38	2.63	4.71
	SELF-*multi*-RAG†	**4.77**	**3.23**	**4.81**

Table 16: Detailed result of human annotation scores along different dimensions of generated responses by different models.

Type	Conversation History	SELF-RAG$_s$	SELF-*multi*-RAG
I. Better processing of conversation history	Q: Who was the father of programming? A: Ada Lovelace is often credited as the 'father (or rather, Mother) of programming' or the first computer programmer. Ada was an English mathematician and writer who worked with Charles Babbage on the design of his proposed mechanical general-purpose computer, the Analytical Engine. She is known for writing the first algorithm intended to be processed by a machine, specifically the Analytical Engine. Her work on the engine, including her notes on the method of using the machine to calculate Bernoulli numbers, is considered the first published algorithm and has led her to be recognized as the world's first computer programmer. Q: When was that algorithm written?	She was born in 1815. Her father was the poet Lord Byron. She was a gifted mathematician and wrote an algorithm for the first mechanical computer, the Analytical Engine	Ada Lovelace's algorithm was written between 1842 and 1843.
	Q: Where did the i have a dream speech happen? A: I Have a Dream was delivered from the steps of the Lincoln Memorial in Washington, D.C. Q: Who made the i have a dream speech? A: I Have a Dream is a public speech that was delivered by American civil rights activist Martin Luther King Jr. Q: how big was the audience for the i have a dream speech? A: I Have a Dream is a public speech that was delivered to over 250,000 civil rights supporters. Q: Qass the speech important?	Martin Luther King Jr. made the I Have a Dream speech from the steps of the Lincoln Memorial in Washington, D. C	Yes, the I Have a Dream speech was a pivotal moment in American history and is considered to be one of the most important speeches in American history
II. Passage provided in question	Q: **Here is a piece of text:** ShadowParty – a supergroup consisting of Devo, New Order and Verve members – have announced their first ever gig. The band is made up of Josh Hager, currently playing guitar and keyboards in Devo, New Order bassist Tom Chapman and guitarist Phil Cunningham and Devo drummer Jeff Friedl. Speaking of the band's origins, Chapman says it was due to him feeling a bit "restless" after moving to Boston due to family commitments. ***The four piece have recorded an album, with guests including Denise Johnson (Primal Scream) and The Verve's guitarist Nick McCabe.*** It's said to have ***a synth-pop vibe, with strings and guitars***. They'll be making their live debut on May 18th at 1930s cinema, The Plaza, as part of The Charlatans North By Northwich. The gig is free and the show will start around 8pm. You can see the poster below. Last year, Devo were given their own set of emojis. Devo formed in 1973 and released debut album 'Q: Are We Not Men? A: We Are Devo!' in 1978. Their last full-length LP 'Something for Everybody' was released in 2010. Many of the new emojis reference memorable moments from videos for Devo hits like 'Whip It' and '(I Can't Get No) Satisfaction' as well as other imagery unique to the band. Based on the text material above, generate the response to the following quesion or instruction: *Can you describe the musical style of ShadowParty and their upcoming debut album?*	I'm not sure.	ShadowParty's upcoming debut album is said to have ***a synth-pop vibe with strings and guitars***. The band has ***recorded an album with guests including Denise Johnson (Primal Scream) and Nick McCabe (The Verve)***.
	Q: MILAN (Reuters) - Italian luxury yacht builder Ferretti is preparing for a return to the Milan stock exchange it abandoned more than 15 years ago, a source close to the matter said on Thursday. The group was delisted in 2003 and acquired by private equity firms before being bought by Chinese industrial conglomerate Weichai Group in 2012. *Under the Chinese ownership, the maker of Riva and Pershing yachts has paid off its debt and returned to profit in 2016. According to the most recent results available on the company's website, Ferretti said it had a "production value" of 623 million euros ($709 million) in 2017. The website said 2017 core profit was up 11 percent at 59 million euros, with net profit jumping 71 percent to 24 million euros.* Owner Weichai Group believes that a new listing would improve the company's reputation and give it the visibility a luxury group deserves, the source said, adding that timing has yet to be decided and will depend on market conditions. The source said the process is at an early stage, with the company having yet to appoint a bank as adviser and no detail available on how much of the company might be floated or at what price. A Ferretti Yachts representative was not immediately available for comment. The group considered a return to the stock market in 2008 but halted the process as the global financial crisis unfolded. The news of Ferretti potentially planning a market comeback were first reported by Italian daily MF. **Based on the passage above answer:** Has Ferretti made any previous attempts to return to the stock market, and what factors led to the company halting those attempts? A: Yes, Ferretti considered a return to the stock market in 2008, but halted the process as the global financial crisis unfolded. Q: Can you provide more information on Ferretti's financial performance in 2016 and 2017, and how it has improved under Chinese ownership?	Ferretti is preparing for a return to the Milan share market it abandoned more than 15 years ago, a source close to the matter said on Thursday. The group was delisted in 2003 and acquired by private equity firms before being bought by Chinese industrial conglomerate Weichai Group in 2012. *Under the Chinese ownership, the maker of Riva and Pershing yachts has paid off its debt and returned to profit in 2016*	Absolutely. *According to the most recent financial reports available on Ferretti's website, the company had a 'production value' of 623 million euros ($709 million) in 2017. Core profit for that year was up 11 percent at 59 million euros, with net profit jumping 71 percent to 24 million euros. The company has improved under Chinese ownership, having paid off its debt.*

Table 17: Examples of SELF-*multi*-RAG improvements in conversational QA over SELF-RAG$_s$.



Instructions (Click to collapse)

Welcome!
Please evaluate the **quality** of responses generated by our system in a multi-turn conversation scenario.

You will be provided with a conversation and a response to the conversation generated by our system. Your task is to evaluate the quality of the responses based on the following dimensions.

You will have to evaluate around 250 responses.

- **Coherence:** evaluate whether the response serves as a valid continuation of the previous conversation.
- **Engagingness:** evaluate if the response is interesting or dull.
- **Understandability:** evaluate whether the response is understandable.

Examples:

Conversation:
Q: What did Colin Cowdrey do against the West Indies in 1966?
A: In an amazing Fifth Test at the Oval the West Indies made 268 and had England 166/7, but the last three wickets added 363 runs and England won.
Q: What position did Colin Cowdrey play?
A: In all Colin Cowdrey played 114 Tests, making 7,624 runs at an average of 44.06, overtaking Wally Hammond as the most prolific Test batsman
Q: What was a career highlight?

Response:
He was a right-handed batsman and a right-arm medium pace bowler.

Coherence 1; Explanation - The response was not relevant to the question which asked for a career highlight of Colin Cowdrey. Instead, the response provided information on his playing position and style which the question did not ask for. While factual, it completely missed the point of the question. The response offered no highlights from Cowdrey's career and did not satisfy the inquiry.
Engagingness 1; Explanation - The response does not provide any relevant interesting fact.
Understandability: 5; Explanation - Although wrong, the response is understandable.

Given a conversation, please evaluate the quality of the response. (See instructions above.)

Conversation:
${conv_hist}

Response:
${response}

Coherence Rate the coherence of the response on a scale of 1 (completely irrelevant to the converation) - 5 (perfect continuation of the converation).

Completely irrelevant to the converation	1	2	3	4	5	Perfect continuation of the converation

Engagingness Rate the engagingness of the response on a scale of 1 (very dull) - 5 (very engaging).

Very dull	1	2	3	4	5	Very engaging

Understandability Rate the understandability of the response on a scale of 1 (not understandable at all) - 5 (perfectly understandable).

Not understandable at all	1	2	3	4	5	Perfectly understandable

Submit

Figure 5: Screenshot of the human annotation template for the response quality measurement.