ChatQA: Surpassing GPT-4 on Conversational QA and RAG

Zihan Liu [*][1]   **Wei Ping** [*][1]   **Rajarshi Roy** [1]   **Peng Xu** [1]

Chankyu Lee [1]   **Mohammad Shoeybi** [1]   **Bryan Catanzaro** [1]

[1] NVIDIA

Abstract

In this work, we introduce ChatQA, a suite of models that outperform GPT-4 on retrieval-augmented generation (RAG) and conversational question answering (QA). To enhance generation, we propose a two-stage instruction tuning method that significantly boosts the performance of RAG. For effective retrieval, we introduce a dense retriever optimized for conversational QA, which yields results comparable to the alternative state-of-the-art query rewriting models, while substantially reducing deployment costs. We also present the CHATRAG BENCH, which encompasses ten datasets covering comprehensive evaluations on RAG, table-related QA, arithmetic calculations, and scenarios involving unanswerable questions. Our ChatQA-1.0-70B (score: 54.14), built on Llama2, a weaker foundation model than GPT-4, can slightly outperform GPT-4-0613 (score: 53.90) and GPT-4-Turbo-2024-04-09 (score: 54.03) on the CHATRAG BENCH, without relying on any synthetic data from OpenAI GPT models. Notably, Llama3-ChatQA-1.5-70B model surpasses the accuracy of GPT-4-Turbo-2024-04-09 by a margin. These results demonstrate the exceptional quality of the proposed ChatQA recipe. To advance research in this field, we open-sourced the model weights, instruction tuning data, CHATRAG BENCH, and retriever for the community: https://chatqa-project.github.io/.

1   Introduction

Most recently, ChatGPT (OpenAI, 2022) and its follow ups (OpenAI, 2023; Anthropic, 2023b; Google, 2023) have led to the paradigm shift of building question answering (QA) and retrieval-augmented generation (RAG) system in production and research community. In particular, the following aspects of the models are preferred: *i)* The users can interact with the QA models in a conversational way, thus one can easily raise follow-up questions. *ii)* The models are capable of integrating retrieved chunks of evidence in both open-domain or long document settings, where the provided context is much longer than the context window of LLM (e.g., Anthropic, 2023a; Xu et al., 2023b). *iii)* The generalist models can answer any questions with respect to table, arithmetic calculation in zero-shot manner without dataset-specific fine-tuning, while matching the accuracies of fine-tuned models. To this end, we focus on building the state-of-the-art model with all these key capabilities that are essentially important for many real-world applications.

However, building such a model that can match the accuracy of the state-of-the-art proprietary models, e.g., GPT-4 (OpenAI, 2023), is still a grand challenge for the research community. In this work, we introduce ChatQA, a family of open-sourced models that can outperform GPT-4 while utilizing relatively weak open-weights foundation model. We also open-source our training data, detailing

---

[*]Correspondence to: Zihan Liu <zihanl@nvidia.com>, Wei Ping <wping@nvidia.com>

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

techniques for synthetic data generation, along with alternative human annotation processes aimed at eliminating reliance on OpenAI GPT models for open research purpose.

Specifically, we make the following contributions:

1. We propose a two-stage instruction tuning method and design a dataset curation recipe that can largely enhance LLM's capability of integrating user provided or retrieved context for conversational QA and RAG tasks. We demonstrate that the proposed instruction tuning method significantly outperforms strong alignment baselines or RLHF-based recipes (e.g., Llama2-Chat, Llama3-Instruct) on RAG and various conversational QA tasks.

2. For retrieval, we show that fine-tuning the single-turn QA retriever on human-annotated data or synthetic multi-turn QA dataset works as well as utilizing the state-of-the-art LLM-based query rewriting model, i.e., GPT-3.5-Turbo (OpenAI, 2022). Our result also highlights the promising direction of utilizing synthetic data generation for training customized retriever.

3. We introduce CHATRAG BENCH, a comprehensive benchmark with ten conversational QA datasets, including five datasets with long documents that need retrieval and three datasets with tabular data and arithmetic calculation. We apply ChatQA training recipe on different text foundation models, and show the superb generalization capability of the proposed methods. In terms of average score on CHATRAG BENCH, our ChatQA-1.0-70B (54.14) based on Llama2 can outperform GPT-4-0613 (53.90) and GPT-4-Turbo-2024-04-09 (54.03) without utilizing any synthetic data from ChatGPT models. Notably, much smaller Llama3-ChatQA-1.5-8B can perform comparable with GPT-4 models, while Llama3-ChatQA-1.5-70B outperforms GPT-4-Turbo-2024-04-09 by a margin.

4. We study the "unanswerable" scenario, where the LLM needs to generate "cannot answer" to avoid hallucination. We show that incorporating a small amount of "unanswerable" samples significantly enhances model's capability to handle it. Our ChatQA-1.0-70B outperforms GPT-3.5-Turbo in this regard, while has a slight gap compared to GPT-4-0613 (around 3.5%).

We discuss related work in § 2. We introduce the two-stage instruction tuning method and data curation for ChatQA in § 3, and study retrieval in conversational QA in § 4. We present the experimental setup in § 5, results in § 6, and conclude the paper in § 7.

2 Related Work

2.1 Conversational QA and RAG

Question answering in a conversational way naturally improves user experiences by addressing follow-up questions. The model can also raise clarification questions for users if necessary, which can reduce hallucination. Thus, it becomes the default format of deploying QA models in production (e.g. OpenAI, 2022; Google, 2023; Anthropic, 2023b). In contrast to the latest LLM-based generalist solution (e.g., OpenAI, 2022), most of the previous studies focus on fine-tuned expert models on specific domains or datasets (Feng et al., 2020; Izacard & Grave, 2021; Chen et al., 2022a; Gao et al., 2022; Nakamura et al., 2022; Adlakha et al., 2022; Wu et al., 2023).

In recent years, many conversational QA datasets have been introduced. The models are asked to answer questions based on provided context or documents, which involves retrieval-augmented generation (RAG) if the provided documents are longer than the context window of LLM. The provided context or documents can be: *i*) text-only documents from various domains (Feng et al., 2020; Anantha et al., 2021; Saeidi et al., 2018; Adlakha et al., 2022; Aliannejadi et al., 2021; Reddy et al., 2019; Qu et al., 2020; Wu et al., 2023; Deng et al., 2022; Guo et al., 2021; Choi et al., 2018; Campos et al., 2020), or *ii*) documents comprising plain text along with tables (Pasupat & Liang, 2015; Nakamura et al., 2022; Chen et al., 2022a).

2.2 Retrieval for Multi-Turn QA

RAG is critically important for conversational QA in open-domain setting, e.g., utilizing update-to-date information from search engine, or when the proprietary documents are longer than the context window of LLM. The dense retrievers are usually trained to retrieve the top-k relevant chunks given a single question (e.g., Lin et al., 2023a; Wang et al., 2022a; Izacard et al., 2022). In conversational QA, the follow-up questions (e.g., with pronouns referring to entities mentioned in the previous conversation) may have insufficient information for retrieval, while feeding them along with all of the dialogue history can be redundant, thus leading to sub-optimal results.

2
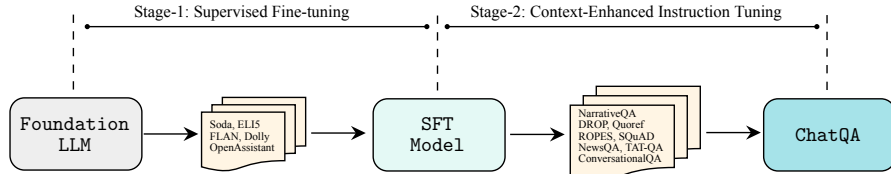
Figure 1: Two-stage instruction tuning framework for ChatQA.

Conversational Query Rewriting Most of the previous solutions are query rewriting methods. The latest turn of question is rewritten to be a standalone query without additional information from previous dialogue history (Vakulenko et al., 2021a; Ye et al., 2023; Mo et al., 2023), so it can be directly used by retrieval model to retrieve relevant context (Vakulenko et al., 2021b; Mele et al., 2021; Raposo et al., 2022; Mo et al., 2023). Many datasets have been collected to facilitate this line of research (Elgohary et al., 2019; Chu et al., 2020; Qu et al., 2020; Anantha et al., 2021; Brabant et al., 2022), alongside multiple proposed query rewriting methods (Ishii et al., 2022; Yu et al., 2020; Wu et al., 2022; Del Tredici et al., 2021; Chen et al., 2022b; Galimzhanova et al., 2023). For example, Wu et al. (2022) and Chen et al. (2022b) proposed to use reinforcement learning methods for the query rewriting. Yu et al. (2020) investigated few-shot generative models like GPT-2 for query rewriting. Galimzhanova et al. (2023) studied instruction tuned GPT-3.5-Turbo and showed that it achieved state-of-the-art results for conversational query rewriting.

Fine-tuning Retriever for multi-turn QA Some previous work fine-tune a single-turn query retriever on in-domain conversational query and context pairs (Feng et al., 2020; Gao et al., 2022; Adlakha et al., 2022; Wu et al., 2023), so it can directly take a concatenation of dialog history and current query as input. In this work, we focus on the zero-shot evaluation. We fine-tune a single-turn query retriever on a high-quality multi-turn dataset. Then, we evaluate zero-shot capability of the fine-tuned retriever on five benchmark datasets. Surprisingly, we find this simple approach can obtain comparable zero-shot results as the state-of-the-art query rewriting model, i.e., GPT-3.5-Turbo.

2.3 Instruction Tuning

The goal of instruction tuning is to equip LLMs with the capability to follow natural language instructions (Wei et al., 2022a; Sanh et al., 2022; Mishra et al., 2022; Iyer et al., 2022; Du et al., 2022; Ouyang et al., 2022; Wang et al., 2023; Zhang et al., 2023b; Gao et al., 2023; Chung et al., 2022; Muennighoff et al., 2022; Xu et al., 2023a; Wang et al., 2022c; Zhou et al., 2023; Albalak et al., 2024). There has been a surge in the development of high-quality instruction tuning datasets, including FLAN (Chung et al., 2022), Self-Instruct (Wang et al., 2022b), unnatural Instructions (Honovich et al., 2022), Dolly (Conover et al., 2023b), and OpenAssistant (Köpf et al., 2023).

Although numerous research on instruction tuning has been conducted, a few works focused on improving RAG or context awareness generation for QA. Lin et al. (2023b) introduced a retrieval-augmented instruction tuning method, which appends top-k retrieved chunks for LLM fine-tuning. Wang et al. (2024) applied instruction tuning after retrieval-augmented pretraining. In contrast, we propose a two-stage instruction tuning method to improve generation with retrieval or provided context. We find that appending top-k retrieved chunks for LLM fine-tuning does not help for a wide range of conversation QA tasks (see Appendix §A.1 for details). Similar to the latest work (Zhang et al., 2023a), we show that adding a small amount of "unanswerable" samples in instruction tuning prompts the model to respond "cannot answer" when needed, significantly reducing hallucination.

3 ChatQA

In this section, we propose a two-stage instruction tuning method for ChatQA. See Figure 1 for an illustration. Our method starts with a pretrained LLM base model. At stage-1, we apply supervised fine-tuning (SFT) as in Ouyang et al. (2022) on a blend of instruction-following and dialog datasets. After that, our model exhibits good capability to follow instructions. However, its capability for contextualized or RAG-based QA remains limited. Hence, we introduce a subsequent stage, called context-enhanced instruction tuning, which is designed specifically for enhancing our model's capability for context-aware or retrieval-augmented generation in conversational QA.

3
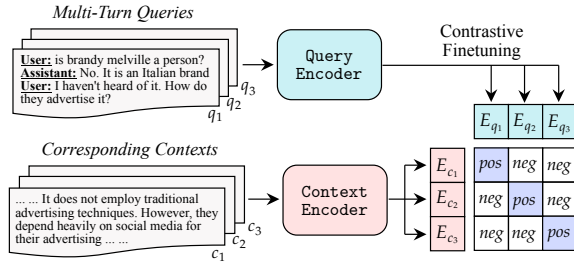
Figure 2: Illustration of fine-tuning retriever for multi-turn QA.

3.1 Stage-1: Supervised Fine-tuning

To construct a large and comprehensive supervised fine-tuning (SFT) dataset, we follow Xu et al. (2023b), Wang et al. (2024) and gather a combined set of 128K SFT samples from high-quality instruction tuning datasets. It consists of 1) a social dialogue dataset Soda (Kim et al., 2022), 2) a long-form QA dataset ELI5 containing elaborate answers (Fan et al., 2019), 3) FLAN and chain-of-thought datasets (Wei et al., 2022b; Chung et al., 2022; Longpre et al., 2023), 4) LLM synthetic instruction tuning datasets, including Self-Instruct (Wang et al., 2022b) and Unnatural Instructions (Honovich et al., 2022), and 5) a private crowd-sourced conversational dataset, as well as two public human-written conversation datasets: OpenAssistant (Köpf et al., 2023), and Dolly (Conover et al., 2023a).

We unify the structure of all the SFT data in a conversational format. We first add a "System" role at the beginning to set up a general instruction guiding LLM to provide polite and helpful answers. We also add "User" and "Assistant" roles to incorporate instruction and response pairs from the SFT dataset. We apply fine-tuning using this unified format on an LLM foundation model.

3.2 Stage-2: Context-Enhanced Instruction Tuning

To further enhance the model's conversational QA capability over a given context, we conduct a second stage instruction tuning, which integrates contextualized QA datasets into the instruction tuning blend. One of the key elements for stage-2 is to obtain a high-quality document-grounded conversational QA dataset. We leverage two strategies of collecting such dataset. One is to leverage GPT-3.5-Turbo to generate a synthetic dataset (referred to as SyntheticConvQA). Another is to collect human-annotated dataset (referred to as HumanAnnotatedConvQA), which enables us to avoid the reliance on OpenAI models. For both datasets, we also collect unanswerable data samples the answers cannot be found within the given context. More details can be found in the Appendix B.

To boost the QA capability in handling tabular documents and arithmetic calculation, we add the TAT-QA dataset (Zhu et al., 2021) which contains both elements. In addition, we integrate contextualized single-turn QA datasets to further strengthen the QA capability of our model. We also retain the stage-1 SFT dataset in the training blend to maintain the model's instruction-following capability.

Finally, the training blend for stage-2 consists of: 1) A conversational QA dataset: HumanAnno-tatedConvQA or SyntheticConvQA, [2] 2) single-turn QA datasets: DROP (Dua et al., 2019), Narra-tiveQA (Kočiský et al., 2018), Quoref (Dasigi et al., 2019), ROPES (Lin et al., 2019), SQuAD1.1 (Ra-jpurkar et al., 2016), SQuAD2.0 (Rajpurkar et al., 2018), NewsQA (Trischler et al., 2017), TAT-QA (Zhu et al., 2021), and 3) all of SFT datasets from stage-1. As for the training of Llama3-ChatQA-1.5, we additionally incorporate HybriDial (Nakamura et al., 2022) and our collected around 2K QA pairs within the financial domain to further improve our model's capability in tabular understanding and arithmetic calculations. We follow the similar template as in stage-1 to unify all single-turn QA and conversational QA datasets. The differences are in two parts: 1) Following the system role, we append relevant context for the single-turn question or the multi-turn conversation, and 2) Just before the single-turn question or multi-turn conversation, we integrate further instruction based on the answer types of different QA datasets (e.g., short answer, long answer, arithmetic calculation). We use the format for SFT dataset from stage-1. [3]

---

[2]Unless specified otherwise, the experiments use HumanAnnotatedConvQA as the default setting.

[3]Details of the templates for both stage-1 and stage-2 instruction tuning as well as the synthetic data generation can be found in the Appendix C.

4

Models	Average		Doc2Dial		QuAC		QReCC		TopiOCQA		INSCIT	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-5*	top-20*	top-5*	top-20*
Adlakha et al. (2022)	-	-	-	-	-	-	-	-	-	70.40△	-	-
Wu et al. (2023)	-	-	-	-	-	-	-	-	-	-	-	71.10△
E5-unsupervised† (Wang et al., 2022a)	31.56	59.22	23.02	55.33	43.49	77.68	44.71	84.99	26.25	37.67	20.32	40.44
E5-unsupervised + Rewrite‡	33.23	61.02	25.56	58.00	46.00	80.01	45.50	85.89	27.58	39.15	21.53	42.04
E5-unsupervised + Fine-tune†	**47.79**	**75.00**	45.28	80.96	46.52	80.74	53.37	89.91	41.01	51.07	52.79	72.31
Dragon† (Lin et al., 2023a)	46.29	73.09	43.33	75.61	56.80	82.86	46.17	81.96	57.68	78.80	27.49	46.22
Dragon + Rewrite‡	**54.46**	80.13	47.60	80.60	47.10	77.15	51.73	85.78	73.07	88.19	52.79	68.92
Dragon + Fine-tune†	52.72	**80.67**	48.94	83.01	52.64	81.95	50.73	87.17	67.86	86.28	43.43	64.94
- SyntheticConvQA◇	52.98	**81.15**	48.64	83.47	54.75	83.23	49.63	86.70	64.48	85.24	47.41	67.13

Table 1: Retrieval results across five multi-turn QA datasets with the average top-1 and top-5 recall scores. Compared to rewriting, fine-tuning performs much better on E5-unsupervised and is comparable on Dragon. *Since the average context length in TopiOCQA and INSCIT is smaller than in other datasets, we report top-5 and top-20 to roughly match the context lengths of top-1 and top-5, respectively, in those datasets. †The inputs for these two models are a concatenation of the dialogue history and the current query. ‡The input for this model is the rewritten query. ◇denotes that the HumanAnnotatedConvQA dataset is replaced with the SyntheticConvQA for fine-tuning. △The numbers are not apple-to-apple comparison (e.g., they use the training set for fine-tuning).

4 Retrieval for Multi-Turn QA

In conversational QA tasks, when a document becomes too lengthy to feed directly into LLMs, a retriever that can handle conversational queries becomes essential. This conversational retriever encodes the concatenation of the dialogue history and the current query, and then retrieve relevant context from documents. After that, only the relevant context will be used as inputs for LLMs. The state-of-the-art retrievers, e.g., Dragon (Lin et al., 2023a), are optimized for single-turn queries, resulting in a limited generalization capability for multi-turn conversational queries. In Figure 2, we depict our retriever fine-tuning method to alleviate this issue. We propose to use conversational query and context pairs for further fine-tuning a single-turn query retriever to cope with multi-turn queries.

An alternative solution is conversational query rewriting method which uses a query rewriter to rewrite the current question based on the conversational history. The rewritten query is then directly used as the input to a single-turn query retriever for retrieving relevant context. In addition to the embedding and search cost, the query rewriting model introduces a large amount of extra computational expense to generate the rewritten query.

4.1 Fine-tuning Retriever for Multi-turn QA

To build a high-quality fine-tuning dataset, we leverage the conversational QA dataset from either HumanAnnotatedConvQA or SyntheticConvQA to construct conversational query and context pairs. For the HumanAnnotatedConvQA, we directly take the annotations of the conversational query and context pairs, and use them to further fine-tune a single-turn query retriever. For the SyntheticConvQA, we first cut each document in the conversational QA dataset into different chunks. Then, we calculate the 4-gram recall score between agent's answer and each chunk. After that, we consider the chunk that has the highest recall score as the gold chunk for the current user's question. Finally, the constructed conversational query and context pairs are used to fine-tune a single-turn query retriever.

4.2 Conversational Query Rewriting

To build powerful conversational query rewriting model, we take GPT-3.5-Turbo as the rewriter given that Galimzhanova et al. (2023) demonstrated the state-of-the-art query rewriting results using GPT-3.5-Turbo. Similar to Galimzhanova et al. (2023), we not only provide GPT-3.5-Turbo with the rewriting task instruction, but also give it few-shot rewriting examples to enhance the quality of rewriting results. More details can be found in Appendix E.1.

4.3 Comparisons

In Table 1, we compare the query rewriting and fine-tuning methods across five datasets in the zero-shot setting. More details about these datasets can be found in §5.2. We conduct experiments on a state-of-the-art retriever, Dragon (Lin et al., 2023a), and a strong unsupervised retriever, E5-unsupervised (Wang et al., 2022a), which is not finetuned on MS MARCO (Nguyen et al., 2016). In terms of the experiments on Dragon, we find that fine-tuning performs marginally worse than query rewriting in average top-1 recall by 1.74%, while it achieves better results on average top-5 recall by 0.54%. It demonstrates the effectiveness of the fine-tuning approach for the conversational retrieval.

5

In addition, we observe that the results are comparable between using HumanAnnotatedConvQA and SyntheticConvQA for fine-tuning. This highlights that our human-annotated dataset is in high-quality, and we *do not rely on* ChatGPT models for building the state-of-the-art multi-turn query retriever.

Surprisingly, fine-tuning performs significantly better than rewriting on E5-unsupervised. We conjecture that since E5-unsupervised does not use human-annotated query and context pairs in the pre-training stage, it leads to weak generalization for the high-quality rewritten query. In contrast, using a high-quality dataset to fine-tune E5-unsupervised brings a giant boost, with more than a 15% improvement on both average top-1 and top-5 recall scores.

Therefore, fine-tuning a good single-turn retriever on high-quality conversational query context pairs performs on par with leveraging the state-of-the-art rewriter. However, rewriting method requires extra computational time for autoregressive generation process and probably also API cost for using powerful models like GPT-3.5-Turbo. In contrast, our proposed multi-turn fine-tuning bypasses these issues. For the QA evaluations across these five datasets, we consistently use the retrieved top-5 results from the fine-tuning approach for all the QA models. We put more results on comparisons between rewriting and fine-tuning methods in the Appendix E.2.

5   Experimental Setup

5.1   Baselines

We develop ChatQA models based on our in-house GPT-{8B, 22B} base models (pretrained with 3.5 trillion tokens), Llama2-{7B, 13B, 70B} base models (Touvron et al., 2023), and Llama3-{8B, 70B} base models (Meta, 2024). We compare ChatQA models against Llama2-Chat-{7B, 13B, 70B} and Llama3-Instruct-{8B, 70B}, which are shown to possess strong instruction following and conversational QA capabilities (Touvron et al., 2023). In addition, we compare against a powerful RAG model, Command R+, which has 104 billion parameters, as well as three very strong OpenAI models: GPT-3.5-Turbo-0613, GPT-4-0613, and GPT-4-Turbo-2024-04-09. For fair comparison, when retrieval is needed, we use the same top-k retrieved chunks from our best retriever as the context for all baselines and our ChatQA models. Note that we have carefully tuned the instructions for all the baselines to ensure they achieve as good as possible results.[4]

5.2   CHATRAG BENCH: Evaluation Benchmarks

To evaluate the model's capability on conversational QA and RAG, we construct CHATRAG BENCH, a collection of 10 datasets covering a wide range of documents and question types, which require models to generate responses from (retrieved) context, comprehend and reason over tables, conduct arithmetic calculations, and indicate when questions cannot be found within the context.

Long Document Datasets   We collect five conversational QA datasets with long documents: Doc2Dial (D2D) (Feng et al., 2020); QuAC (Choi et al., 2018), QReCC (Anantha et al., 2021), TopiOCQA (TCQA) (Adlakha et al., 2022), and INSCIT (Wu et al., 2023). One can find details of these datasets in Appendix F. Since the documents of these datasets cannot be directly fitted into LLMs with a sequence length of 4K or 8K tokens. Hence, we run our best multi-turn query retriever to get top-k relevant chunks as the inputs. For Doc2Dial, QuAC, and QReCC, we segment documents into around 300-word chunks, and we retrieve top-5 relevant chunks as context for each user question. For TopiOCQA and INSCIT, we follow their original segmentation, resulting in smaller chunks. Hence, we retrieved top-20 chunks to obtain similar context length to the first three datasets (experiments can be found in §4.3).

Short Document Datasets   To increase the diversity of document lengths, we collect five conversational QA datasets with short documents (less than 1.5K words): CoQA (Reddy et al., 2019), a dataset covering a wide range of domains; DoQA (Campos et al., 2020), a dataset covering cooking, travel and movie domains with unanswerable cases; ConvFinQA (Chen et al., 2022a), a dataset based on Financial domain with tabular document and requiring arithmetic calculation and complex numerical reasoning; SQA (Pasupat & Liang, 2015); and HybriDial (HDial) (Nakamura et al., 2022). Both SQA and HybriDial have tabular documents from Wikipedia and require reasoning for complex questions.[5]

---

[4]The prompts for these baselines can be found in Appendix F.

[5]More dataset details can be found in Appendix F.

6

Models	Avg. w/o HDial	Avg. All	D2D	QuAC	QReCC	CoQA	DoQA	CFQA	SQA	TCQA	HDial	INSCIT
Llama2-Chat-7B	37.96	38.86	33.27	25.83	46.02	72.28	33.15	36.58	26.14	36.68	47.02	31.67
Llama2-Chat-70B	44.64	45.21	36.87	32.47	49.40	80.41	38.97	46.85	37.62	44.31	50.35	34.88
Llama3-Instruct-8B	46.76	46.70	33.59	33.60	45.70	75.26	37.33	58.05	59.72	44.96	46.20	32.59
Llama3-Instruct-70B	52.95	52.52	37.88	36.96	51.34	76.98	41.24	76.60	69.61	49.72	48.59	36.23
ChatQA-1.0-7B	46.96	47.71	37.88	29.69	46.97	76.61	41.57	51.61	61.87	45.45	54.51	30.96
ChatQA-1.0-70B	53.89	54.14	38.90	41.82	48.05	78.57	51.94	73.69	69.14	50.98	56.44	31.90
Llama3-ChatQA-1.5-8B	53.99	55.17	39.33	39.73	49.03	76.46	49.60	78.46	73.28	49.96	65.76	30.10
Llama3-ChatQA-1.5-70B	57.14	58.25	41.26	38.82	51.40	78.44	50.76	81.88	83.82	55.63	68.27	32.31
Command R+	51.40	50.93	33.51	34.16	49.77	69.71	40.67	71.21	74.07	53.77	46.70	35.76
GPT-3.5-Turbo-0613	50.69	50.37	34.83	37.17	50.46	79.33	41.11	73.15	60.63	44.30	47.42	35.27
GPT-4-0613	54.35	53.90	34.16	40.29	52.01	77.42	43.39	81.28	79.21	45.09	49.81	36.34
GPT-4-Turbo-2024-04-09	54.72	54.03	35.35	40.10	51.46	77.73	41.60	84.16	79.98	48.32	47.86	33.75

Table 2: Zero-shot results on CHATRAG BENCH. ChatQA-1.0 are built on Llama2 base models, while Llama3-ChatQA-1.5 are built on Llama-3 base models. Note that Llama3-ChatQA-1.5 used training samples from the HDial dataset. Hence, we also compare average scores excluding HDial for the fair comparison.

Models	Avg.	D2D	QuAC	QReCC	CoQA	DoQA	CFQA	SQA	TCQA	HDial	INSCIT
ChatQA-1.0-70B	**54.14**	38.90	41.82	48.05	78.57	51.94	73.69	69.14	50.98	56.44	31.90
- SyntheticConvQA$^\diamond$	**54.08**	39.19	38.33	48.73	79.83	48.65	76.44	68.63	51.30	55.68	33.98
- w/o stage-1†	52.18	38.43	37.52	46.08	73.51	49.42	72.15	72.08	51.28	50.74	30.56
- w/o stage-2†	43.22	34.42	25.65	41.88	73.04	28.21	46.64	58.90	37.20	55.52	30.71
- w/o single-turn*	52.25	38.30	37.89	47.08	76.74	46.43	72.42	67.41	49.85	53.16	33.18
- w/o ConvQAData‡	48.97	35.47	28.19	41.68	75.80	31.68	73.83	68.05	45.90	58.40	30.70

Table 3: Ablation studies on ChatQA training. $\diamond$ denotes that the HumanAnnotatedConvQA is replaced with the SyntheticConvQA. $\dagger$ denotes that the stage-1 (SFT) or stage-2 tuning is removed from the ChatQA training. * denotes that the single-turn QA datasets are removed from the stage-2 training blends. $\ddagger$ denotes that the HumanAnnotatedConvQA is removed from the stage-2 training blends. In both "w/o stage-1" and "w/o single-turn" settings, the SyntheticConvQA data is used.

5.3 Evaluation Metrics

Given that F1 score is the most commonly used automatic metric to assess QA models, we use it for all datasets except for ConvFinQA. In ConvFinQA, we follow Chen et al. (2022a) to use exact match metric since the answers in ConvFinQA are about extracting numbers from documents as well as arithmetic calculations. Hence, the answer only makes sense when it is exactly the same as the answer. When models generate the arithmetic formula, we will calculate its final result based on a calculator and compare it with the gold answer. In addition, we also conduct human evaluations to assess the correctness of generated answers between our model and GPT-4-0613.

6 Results

6.1 Main Results

Overview In Table 2, we compare different model variants and OpenAI models across 10 conversational QA datasets. ChatQA-1.0-7B/70B greatly surpass Llama2-Chat counterparts, and Llama3-ChatQA-1.5-8B/70B greatly outperform Llama3-Instruct counterparts. For example, ChatQA-1.0-70B outperforms Llama2-Chat-70B by an average score of 8.93, and Llama3-ChatQA-1.5-70B surpasses Llama3-Instruct-70B by an average score of 4.19 (Avg. w/o HDial). In addition, ChatQA-1.0-70B slightly outperforms GPT-4-0613 and GPT-4-Turbo, and greatly surpasses Command R+, a powerful RAG model that has 104 billion parameters, as well as GPT-3.5-Turbo (Avg. All). Compared to state-of-the-art OpenAI models (i.e., GPT-4-0613 and GPT-4-Turbo), Llama3-ChatQA-1.5-8B achieves comparable results, and Llama3-ChatQA-1.5-70B greatly outperforms both of them. [6]

Furthermore, we compare our models and OpenAI models across different dataset types (e.g., documents with table or solely text) on CHATRAG BENCH. We observe that Llama3-ChatQA-1.5-70B surpasses GPT-4-0613 and GPT-4-Turbo in all categories. Detail results are in the Appendix A.2.

Importance of Stage-1 SFT and Stage-2 Tuning In Table 3, we find that removing the stage-1 SFT from the ChatQA training makes the average score drops 1.9 (from 54.08 to 52.18). Aside from the SQA, removing stage-1 makes the model consistently perform worse on other datasets. Results indicate that stage-1 still plays an important role. It enhances the instruction-following capability of ChatQA which is beneficial for stage-2 tuning. In addition, we find that removing ChatQA stage-2

---

[6]We place full results (e.g., models based on Llama2-13B, and our in-house 8B/22B models) in Appendix K.

7
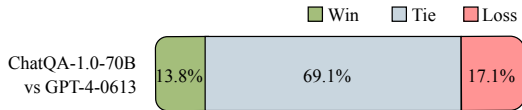
Figure 3: Human evaluation (A/B testing) comparing ChatQA-1.0-70B to GPT-4-0613 on CHATRAG BENCH.

tuning makes the average scores greatly drop by 10.92 (from 54.14 to 43.22). This is because the stage-2 tuning (i.e., context-enhanced instruction fine-tuning) enables the model to learn how to effectively leverage information from retrieved or relevant context.

Effectiveness of Single-Turn Data   To investigate how single-turn QA datasets affect model's multi-turn QA capability, we conduct an ablation study by removing them from the ChatQA-1.0-70B training blends in stage-2. As shown in Table 3, incorporating single-turn QA datasets in the stage-2 training blends generally make the scores increase across all benchmark datasets, leading to an average improvement of 1.83 score. Interestingly, we observe improvement in ConvFinQA, SQA, and HybriDial (table-based datasets), despite the added single-turn QA datasets not having tabular data in the documents. These results align with our intuitions. Adding single-turn QA datasets improves the model's capability to effectively leverage relevant context for the answers, resulting in better scores on CHATRAG BENCH.

Effectiveness of Conversational QA Data   We further explore the how conversational QA data affect the model's multi-turn QA capability by removing HumanAnnotatedConvQA data from the ChatQA stage-2 training blends. As illustrated in Table 3, "w/o ConvQAData" makes the results significantly worse than ChatQA-1.0-70B (average scores degrading from 54.08 to 48.97). We observe large degradation in datasets with text-based documents, such as QuAC, QReCC, and DoQA. However, the degradation in datasets with table-based documents (e.g., ConvFinQA, SQA) are small, thanks to having TAT-QA (Zhu et al., 2021) in the training blends.

Human Annotated Data vs. GPT-3.5-Turbo Synthetic Data   In Table 3, we also compare our ChatQA models using the 7k GPT-3.5-Turbo synthetic dataset (SyntheticConvQA) and our collected 7k human-annotated dataset (HumanAnnotatedConvQA). First, we find that both achieve comparable results in terms of average scores, which suggests that we do not need to rely on synthetic data from OpenAI models to build the state-of-the-art conversational QA models. Second, we find that using human-annotated data achieved significant improvements on QuAC and DoQA datasets. This can be attributed to the fact that the human-annotated data have higher quality on unanswerable cases which exists in QuAC and DoQA datasets. Eventually, it leads to the overall improvements on these two datasets. Detail results and analyses on unanswerable cases can be found in §6.2.

Human Evaluation   Despite F1 scores being the most commonly used metrics for evaluating the quality of QA models, there are often multiple ways to answer questions, which makes the automatic metrics less than perfect. Therefore, we use human evaluations to further compare our ChatQA-1.0-70B with GPT-4-0613. In this human evaluation, we ask annotators to verify the facts in ChatQA-1.0-70B and GPT-4's outputs and determine which model provides a more accurate response to the question. The human evaluation results averaged over the 10 datasets in CHATRAG BENCH are shown in Figure 3. We find that our ChatQA-1.0-70B and GPT-4 are tie most of the time (69.1%), and GPT-4 achieves slightly higher win rate (3.3%) than ours. This further confirms our model has powerful capability to produce correct answers. [7]

Additional Ablation Studies & Case Studies   We conduct ablation studies on ChatQA training in terms of using top-k chunks as context for stage-2 tuning (See Appendix A.1). Furthermore, we also study how different factors affect the model's generation quality on the inference stage, in terms of: 1) the number of retrieved context/chunks; 2) context ordering; and 3) different retrievers (See Appendix A.3). In addition to quantitative results, we perform detailed qualitative case studies, and compare the output of ChatQA-1.0-70B, GPT-3.5-Turbo-0613, and GPT-4-0613 in Appendix I.

6.2   Evaluation of Unanswerable Case

Evaluation Setup   In this section, we study another aspect of the model's capability, which is to discern if a question can be answered within the provided context. Generating an answer in unanswerable case will lead to hallucination. To allow this evaluation, we require the model to indicate it when no answer can be found in the given context.

---

[7]Detailed human evaluation setup and full results can be found in the Appendix G.

8

Models	Avg Both	Avg QuAC	QuAC (no*)	QuAC (yes*)	Avg DoQA	DoQA (no*)	DoQA (yes*)	ChatRAG
ChatQA-1.0-70B	**77.25**	80.76	77.66	83.85	73.74	68.81	78.67	**54.14**
Command R+	68.11	69.61	41.79	97.42	66.62	46.37	86.87	50.93
GPT-3.5-Turbo-0613	73.27	78.34	61.91	94.76	68.21	51.99	84.43	50.37
GPT-4-0613	**80.73**	87.42	83.45	91.38	74.05	74.28	73.82	**53.90**
GPT-4-Turbo-2024-04-09	80.47	88.73	80.42	97.03	72.21	72.28	72.13	**54.03**

Table 4: Accuracies on answerable and unanswerable samples across QuAC and DoQA datasets. Avg-Both is the averaged score between QuAC and DoQA. ChatRAG is the average score on the CHATRAG BENCH. * "no" and "yes" denote unanswerable and answerable samples, respectively.

Models	Average	NQ	TriviaQA	HotpotQA
Atlas (11B) (Izacard et al., 2023)	39.4	26.7	56.9	34.7
Raven (11B) (Huang et al., 2023)	-	29.6	65.7	-
RECOMP (20B) (Xu et al., 2024)	42.1	37.0	59.0	30.4
InstructRetro (43B) (Wang et al., 2024)	-	38.9	65.6	-
RePlug (65B) (Shi et al., 2023)	44.5	28.8	72.6	32.0
RA-DIT (65B) (Lin et al., 2024)	50.1	35.2	75.4	39.7
Llama3-Instruct-8B (Meta, 2024)	42.5	30.9	70.7	26.0
Llama3-Instruct-70B (Meta, 2024)	53.6	42.7	82.4	35.6
Llama3-ChatQA-1.5-8B	52.3	42.4	81.0	33.5
Llama3-ChatQA-1.5-70B	**58.7**	**47.0**	**85.6**	**42.2**

Table 5: Zero-shot exact match scores on Natural Questions (NQ), TriviaQA, and HotpotQA, which were evaluated using the data split from the KILT Benchmark (Petroni et al., 2021).

We use QuAC and DoQA datasets which have such unanswerable cases to evaluate such capability. Specifically, for unanswerable case, we consider the model indicating that the question cannot be answered as correct[8], and as for answerable cases, we consider the model not indicating the question is unanswerable as correct (i.e., the model giving an answer). Note that for answerable cases, we only select the samples where correct context is retrieved. In the end, we calculate the average accuracy score of unanswerable and answerable cases as the final metric. We consider this average accuracy as the evaluation metric since it is in the same spirit of F1 metric which measures the harmonic mean of precision and recall scores.

Results In Table 4, we compare our models with OpenAI models across QuAC and DoQA datasets. We observe that OpenAI models show powerful capability in this task, especially for GPT-4-0613. Compared to them, our best model (ChatQA-1.0-70B) achieved a significantly better average accuracy than GPT-3.5-Turbo, while we still has a slight gap compared to GPT-4 and GPT-4-Turbo (around 3.5% and 3.2%, respectively). Furthermore, we conduct ablation studies in terms of the number of unanswerable samples for training. We find that using a small amount of unanswerable samples (e.g., 1.5k) is able to achieve remarkable results on the unanswerable evaluation and incorporating more unanswerable samples does not necessarily lead to higher accuracy scores in the unanswerable evaluation (Detail results are in the Appendix H.2).

6.3 Evaluation on Single-Turn QA and RAG Benchmark

In addition to CHATRAG BENCH, we further evaluate Llama3-ChatQA-1.5 models on knowledge-intensive single-turn QA datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), Trivi-aQA (Joshi et al., 2017), and HotpotQA (Yang et al., 2018), and compare them against frontier RAG models. We use the Dragon retriever (Lin et al., 2023a) to extract the top-k most relevant contexts, which are then used as the inputs for Llama3-Instruct and Llama3-ChatQA-1.5. We report the best results obtained from using the top-15, top-20, and top-25 contexts. In Table 5, we show that, despite its significantly smaller model size, Llama3-ChatQA-1.5-8B performs better than the state-of-the-art RA-DIT (65B) model. Llama3-ChatQA-1.5-70B remarkably outperforms existing frontier RAG models.

7 Conclusion

In this paper, we build a family of models that surpass GPT-4 on conversational QA and RAG. We introduce CHATRAG BENCH, a collection of 10 conversational QA datasets covering comprehensive

---

[8]We collect a set of heuristic matching patterns from all generated samples to determine if the model suggests that the question is unanswerable. More details can be found in the Appendix H.

9

evaluations on RAG, table-based reasoning, arithmetic calculations, and unanswerable scenarios. Our ChatQA-1.0-70B model built on Llama2 can slightly outperform GPT-4-0613 and GPT-4-Turbo without using any synthetic data from OpenAI GPT models. Remarkably, Llama3-ChatQA-1.5-70B even surpasses GPT-4-Turbo in all categories of CHATRAG BENCH. In addition, we demonstrate that fine-tuning a single-turn query retriever using our curated conversational QA data performs comparably to the state-of-the-art LLM-based query rewriting model, without incurring extra computational time and potential API costs associated with rewriting. Furthermore, we show that incorporating a small amount of "unanswerable" samples can significantly enhance our model's capability to handle scenarios where answers are unavailable.

Impact Statement

In this section, we discuss potential positive and negative social impacts that could arise from our ChatQA models.

Positive Impacts   First, ChatQA enables users to interact with their data, including documents and tables, and provides accurate responses to their questions. This enhances productivity and efficient learning. Second, ChatQA can improve customer support services by quickly retrieving information about products, services, and common issues, allowing customer queries to be addressed more efficiently. Third, ChatQA can assist professionals, such as those in the medical field, by gathering relevant information and improving decision-making.

Potential Negative Impacts   First, although our ChatQA models demonstrate impressive accuracy in generating answers, they still have the potential to provide factually inaccurate responses. Second, LLMs can inadvertently learn and perpetuate biases present in their pre-training data, which may lead to biased responses in our QA models as they are further fine-tuned based on these LLMs. Third, there is a risk of malicious use of our ChatQA models, such as gathering personal information by inputting documents containing relevant personal content.

Limitation

ChatQA models are optimized for RAG and conversational QA across various types of questions and documents, making them very useful in many real-world applications. However, they may not perform well on code-related tasks or math reasoning tasks compared to GPT-4 and other versatile frontier models, because ChatQA's instruction-tuning blend does not include any code generation data. In the future, we plan to include more code-related and math reasoning data to support potential use cases.

References

Adlakha, V., Dhuliawala, S., Suleman, K., de Vries, H., and Reddy, S. Topiocqa: Open-domain conversational question answering with topic switching. *TACL*, 2022.

Albalak, A., Elazar, Y., Xie, S. M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., and Burtsev, M. Building and evaluating open-domain dialogue corpora with clarifying questions. In *EMNLP*, 2021.

Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., and Chappidi, S. Open-domain question answering goes conversational via question rewriting. In *NAACL*, 2021.

Anthropic. Introducing 100k context windows, 2023a.

Anthropic. Introducing Claude, 2023b.

Brabant, Q., Lecorvé, G., and Barahona, L. M. R. Coqar: Question rewriting on coqa. In *LREC*, 2022.

10

Campos, J. A., Otegi, A., Soroa, A., Deriu, J. M., Cieliebak, M., and Agirre, E. Doqa-accessing domain-specific faqs via conversational qa. In *ACL*, 2020.

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., and Wang, W. Y. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *EMNLP*, 2022a.

Chen, Z., Zhao, J., Fang, A., Fetahu, B., Rokhlenko, O., and Malmasi, S. Reinforced question rewriting for conversational question answering. In *EMNLP*, 2022b.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. Quac: Question answering in context. In *EMNLP*, 2018.

Chu, Z., Chen, M., Chen, J., Wang, M., Gimpel, K., Faruqui, M., and Si, X. How to ask better questions? a large-scale multi-domain dataset for rewriting ill-formed questions. In *AAAI*, 2020.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models. *arXiv preprint arXiv: 2210.11416*, 2022.

Conover, M., Hayes, M., Mathur, A., Meng, X., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., et al. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023a.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free Dolly: Introducing the world's first truly open instruction-tuned llm, 2023b.

Dai, Z., Chaganty, A. T., Zhao, V., Amini, A., Green, M., Rashid, Q., and Guu, K. Dialog inpainting: Turning documents to dialogs. In *ICML*, 2022.

Dasigi, P., Liu, N. F., Marasović, A., Smith, N. A., and Gardner, M. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP*, 2019.

Del Tredici, M., Barlacchi, G., Shen, X., Cheng, W., and de Gispert, A. Question rewriting for open-domain conversational qa: Best practices and limitations. In *CIKM*, 2021.

Deng, Y., Lei, W., Zhang, W., Lam, W., and Chua, T.-S. Pacific: Towards proactive conversational question answering over tabular and textual data in finance. In *EMNLP*, 2022.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. Glm: General language model pretraining with autoregressive blank infilling. In *ACL*, 2022.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL*, 2019.

Elgohary, A., Peskov, D., and Boyd-Graber, J. Can you unpack that? learning to rewrite questions-in-context. In *EMNLP*, 2019.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. Eli5: Long form question answering. In *ACL*, 2019.

Feng, S., Wan, H., Gunasekara, C., Patel, S., Joshi, S., and Lastras, L. doc2dial: A goal-oriented document-grounded dialogue dataset. In *EMNLP*, 2020.

Galimzhanova, E., Muntean, C. I., Nardini, F. M., Perego, R., and Rocchietti, G. Rewriting conversational utterances with instructed large language models. In *IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, 2023.

Gao, C., Zhang, W., and Lam, W. Unigdd: A unified generative framework for goal-oriented document-grounded dialogue. In *ACL*, 2022.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Google. Introducing bard, 2023.

11

Guo, M., Zhang, M., Reddy, S., and Alikhani, M. Abg-coqa: Clarifying ambiguity in conversational question answering. In *AKBC*, 2021.

Honovich, O., Scialom, T., Levy, O., and Schick, T. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022.

Huang, J., Ping, W., Xu, P., Shoeybi, M., Chang, K. C.-C., and Catanzaro, B. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*, 2023.

Ishii, E., Xu, Y., Cahyawijaya, S., and Wilie, B. Can question rewriting help conversational question answering? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, 2022.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

Izacard, G. and Grave, É. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 2023.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

Kim, H., Hessel, J., Jiang, L., Lu, X., Yu, Y., Zhou, P., Bras, R. L., Alikhani, M., Kim, G., Sap, M., et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*, 2022.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The narrativeqa reading comprehension challenge. *TACL*, 2018.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: A benchmark for question answering research. *TACL*, 2019.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. Openassistant conversations - democratizing large language model alignment. *arXiv preprint arXiv: 2304.07327*, 2023.

Lin, K., Tafjord, O., Clark, P., and Gardner, M. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019.

Lin, S.-C., Asai, A., Li, M., Oguz, B., Lin, J., Mehdad, Y., Yih, W.-t., and Chen, X. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*, 2023a.

Lin, X. V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., et al. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*, 2023b.

Lin, X. V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., Zettlemoyer, L., and tau Yih, W. RA-DIT: Retrieval-augmented dual instruction tuning. In *ICLR*, 2024. URL https://openreview.net/forum?id=22OTbutug9.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Mele, I., Muntean, C. I., Nardini, F. M., Perego, R., Tonellotto, N., and Frieder, O. Adaptive utterance rewriting for conversational search. *Information Processing & Management*, 2021.

Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

Mo, F., Mao, K., Zhu, Y., Wu, Y., Huang, K., and Nie, J.-Y. Convgqr: Generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645*, 2023.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Nakamura, K., Levy, S., Tuan, Y.-L., Chen, W., and Wang, W. Y. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of ACL*, 2022.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2016.

OpenAI. Introducing ChatGPT, 2022.

OpenAI. GPT-4, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. In *ACL*, 2015.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., et al. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL*, 2021.

Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W. B., and Iyyer, M. Open-retrieval conversational question answering. In *SIGIR*, 2020.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. In *ACL*, 2018.

Raposo, G., Ribeiro, R., Martins, B., and Coheur, L. Question rewriting? assessing its importance for conversational question answering. In *ECIR*, 2022.

Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge. *TACL*, 2019.

Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G., and Riedel, S. Interpretation of natural language rules in conversational machine reading. In *EMNLP*, 2018.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022.

13

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017.

Vakulenko, S., Longpre, S., Tu, Z., and Anantha, R. Question rewriting for conversational question answering. In *WSDM*, 2021a.

Vakulenko, S., Voskarides, N., Tu, Z., and Longpre, S. A comparison of question rewriting methods for conversational passage retrieval. In *ECIR*, 2021b.

Wang, B., Ping, W., McAfee, L., Xu, P., Li, B., Shoeybi, M., and Catanzaro, B. Instructretro: Instruction tuning post retrieval-augmented pretraining. In *ICML*, 2024.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022a.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *EMNLP*, 2022c.

Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., et al. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*, 2023.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *ICLR*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022b.

Wu, Z., Luan, Y., Rashkin, H., Reitter, D., Hajishirzi, H., Ostendorf, M., and Tomar, G. S. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. In *EMNLP*, 2022.

Wu, Z., Parish, R., Cheng, H., Min, S., Ammanabrolu, P., Ostendorf, M., and Hajishirzi, H. Inscit: Information-seeking conversations with mixed-initiative interactions. *TACL*, 2023.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023a.

Xu, F., Shi, W., and Choi, E. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *ICLR*, 2024. URL https://openreview.net/forum?id=mlJLigNHp.

Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., and Catanzaro, B. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*, 2023b.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

Ye, F., Fang, M., Li, S., and Yilmaz, E. Enhancing conversational search: Large language model-aided informative query rewriting. In *EMNLP*, 2023.

14

Yu, S., Liu, J., Yang, J., Xiong, C., Bennett, P., Gao, J., and Liu, Z. Few-shot generative conversational query rewriting. In *SIGIR*, 2020.

Zhang, H., Diao, S., Lin, Y., Fung, Y. R., Lian, Q., Wang, X., Chen, Y., Ji, H., and Zhang, T. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*, 2023a.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023b.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., and Chua, T.-S. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *ACL*, 2021.

15

A   Additional Results & Analyses

A.1   Using Top-k Chunks as Context for Stage-2 Instruction Tuning

Models	ChatRAG	Avg-ret	Avg-nonret
ChatQA-1.0-70B	**54.14**	42.33	**65.96**
- w/ "top-5" chunks	54.04	**42.91**	65.16

Table 6: Ablation study on using "top-5" retrieved chunk as the context for the stage-2 instruction tuning. We report average scores on CHATRAG BENCH, five retrieval datasets (Avg-ret) and five non-retrieval datasets (Avg-nonret).

For all the datasets used in stage-2 tuning, the context is provided as continuous paragraphs or documents containing the answer. In contrast, the model needs to handle the top-k retrieved chunks during inference for long documents. To address this train/test mismatch, we investigate whether replacing some continuous paragraphs with retrieved top-k chunks will enhance the model's robustness.

We use NarrativeQA for this study, since each question has a corresponding long document. Initially, we use a summary of the long document as the context, which contains the answer. To incorporate discontinuous contexts, we first cut the long document into 300-word chunks. Then, we use the Dragon retriever to retrieve the top-4 chunks related to the question as additional context. Finally, we consider the retrieved four chunks along with the summary of the long document as the "top-5" chunks. [9] We use this reconstructed NarrativeQA to replace the original one for the stage-2 instruction tuning.

In Table 6, we observe that using "top-5" chunks as the context for training leads to improvements on the datasets that require retrieval. But it degrades the performance on non-retrieval datasets. Overall, these two models perform comparable. It is because incorporating "top-5" retrieved chunks in stage-2 tuning aligns with the inference stage where retrieval is needed, which improves the Avg-ret score. However, mixing continuous and discontinuous documents could make the stage-2 tuning less stable, leading to sub-optimal results on non-retrieval datasets. We believe more future work can be conducted regarding the balance of incorporating continuous context and top-k retrieved chunks in stage-2 tuning.

A.2   Fine-grained Analyses

Models	Avg-text	Avg-table	Avg-table w/o HDial	Avg-ret	Avg-nonret
ChatQA-1.0-70B	48.88	66.42	71.42	42.33	65.96
Llama3-ChatQA-1.5-8B	47.74	72.50	75.87	41.63	68.71
Llama3-ChatQA-1.5-70B	**49.79**	**77.99**	**82.85**	**43.89**	**72.62**
GPT-3.5-Turbo-0613	46.07	60.40	66.89	40.41	60.33
GPT-4-0613	46.96	70.10	80.25	41.58	66.22
GPT-4-Turbo-2024-04-09	46.90	70.67	82.07	41.80	66.27

Table 7: Fine-grained studies on average scores of different dataset types. Avg-text covers datasets where the documents only have text, including Doc2Dial, QuAC, QReCC, CoQA, DoQA, TopiOCQA, and INSCIT. Avg-table covers datasets with table in the documents, including ConvFinQA, SQA, and HybriDial. Avg-ret covers datasets with long documents requiring retrieval, including Doc2Dial, QuAC, QReCC, TopiOCQA, and INSCIT. Avg-nonret covers datasets with short documents which do not require retrieval, including CoQA, DoQA, ConvFinQA, SQA, and HybriDial.

In Table 7, we further compared our models and OpenAI models across different dataset types in our conversational QA benchmarks. ChatQA-1.0-70B achieves better results in text-only documents (Avg-text), while GPT-4-0613 shows better QA capability in tabular data given (Avg-table). Regarding

---

[9]Note that, we do not directly use top-5 retrieved chunks for training, because they may not contain the answer. In such cases, fine-tuning the model to generate answer could encourage hallucination.

16

the datasets that require or do not require retrieval (Avg-ret and Avg-nonret), ChatQA-1.0-70B and GPT-4-0613 are comparable.

Furthermore, Llama3-ChatQA-1.5-70B surpasses GPT-4-0613 and GPT-4-Turbo in all categories.

A.3 Ablation Studies on Inference Stage

Models	Avg.	Doc2Dial	QuAC	QReCC	TopiOCQA	INSCIT
ChatQA-1.0-70B	42.31	39.19	38.33	48.73	51.30	33.98
- # of ctx: top-3	41.91	37.20	38.35	48.94	52.78	32.27
- # of ctx: top-10	40.71	37.06	36.95	47.61	49.40	32.53
- ctx reverse ordering	42.48	39.08	38.85	49.63	51.16	33.69
- ctx swing ordering	42.30	39.35	38.09	49.09	50.98	33.99
- ctx random ordering	42.01	39.32	38.28	48.79	50.13	33.51
- Dragon Retrieval	40.50	37.92	38.44	47.88	50.39	27.87

Table 8: Ablation studies on input context across datasets that require retrieval. All models use SyntheticConvQA. We study the number of contexts used in inputs (# of ctx), context ordering (reverse, swing, random), and the use of retrieved context from the original Dragon. In comparison, ChatQA-1.0-70B (default setting) uses "Dragon + Fine-tune" to retrieve the top-5 contexts, and arranges them sequentially from the first to the fifth context in top-5.

In Table 8, we show ablation studies on how the number of retrieved context/chunks, context ordering, and different retrievers affect the conversational QA and RAG results.

First, we find that using more contexts as inputs do not always improve the results. Utilizing top-5 contexts as input yields better results compared to using either top-3 or top-10 contexts. Intuitively, more contexts have higher probability to contain correct answers (better recall scores). As a result, using top-5 contexts achieves better results than using top-3. However, as the number of contexts further increases, the model may suffer from "lost in the middle" phenomenon (Liu et al., 2023) and the difficulty of extracting answers from the provided context could also increase, which leads to inferior results by using top-10 contexts.

Second, we study how using different orderings of top-5 contexts affects the results. We compare sequential ordering (from 1st context to 5th context) to reverse ordering (from 5th to 1st context), swing ordering (given the "lost in the middle" phenomenon, we arrange the most relevant context to appear at the beginning and the end of the input context. Hence, the ordering becomes {1st, 3rd, 5th, 4th, 2nd}), and random ordering (random shuffle the top-5 contexts). We find using sequential ordering is comparable to using reverse and swing orderings, and random shuffling is slightly worse. Results indicate that our model excels in extracting the correct answer from lengthy contexts, regardless of the answer's location. This is because, during the ChatQA fine-tuning, the answer's location is randomly distributed within the context.

Third, we observe that when we replace "Dragon + Fine-tune" with the original non-finetuned Dragon retriever, the average score drops by 1.81 (from 42.31 to 40.50). In addition, the score drops significantly in INSCIT dataset (from 33.98 to 27.87) due to the large performance gap between the two retrievers (as shown in Table 1). Basically, when the quality of retrieval improves, it directly enhances the performance of question answering.

17

B Conversational QA Data Collection for Stage-2 Instruction Tuning

B.1 Human Annotated Data

In addition to the public available datasets, one of the key elements for stage-2 is to obtain a high-quality document-grounded conversational QA dataset. We create a human-annotated conversational QA (called HumanAnnotatedConvQA) dataset only consisting of 7k dialogues. To build this dataset, we first collected 7k documents covering diverse topics from the Internet. Then, we instruct the annotator to act both as a curious user asking questions (and follow-up questions) about the document and as an agent to give the answers. We create a multi-turn conversation for each document, resulting in a total of 7k conversational QA dialogues with an average of 5 user-agent turns per dialogue. Details of data collection guideline can be found in the Appendix J.2.

To reduce hallucinated answers in unanswerable cases, we aim to empower our model to explicitly indicate it when the answer cannot be found within the given context. To obtain these unanswerable data samples, we requested annotators to identify all parts of the context locations that are relevant to the user's question. Hence, it enabled us to construct unanswerable scenarios by deleting the text from the corresponding locations in the context. After deleting the relevant text to the question, we use a sentence, "`Sorry. I cannot find the answer based on the context`", as the response for the unanswerable question. Finally, we construct another 1.5k user-agent turns with unanswerable annotations, which provides a good trade-off for answerable and unanswerable cases. Detailed results can be found in the Appendix H.2.

B.2 Synthetic Data Generation

To validate the quality of the HumanAnnotatedConvQA, we leverage GPT-3.5-Turbo to generate a synthetic conversational QA dataset given its powerful instruction-following and text generation capability. Note that large-scale synthetic data for conversational QA has also been explored in Dai et al. (2022). In this work, we focus on midsize high-quality synthetic data for LLM fine-tuning. One can find the details of synthetic data generation at Appedix D.

Similar to the HumanAnnotatedConvQA, we construct another 1.5k user-agent turns with unanswerable annotations in this synthetic dataset. Since there are no annotations of the context location for the agent's answer, we construct synthetic unanswerable samples from SyntheticConvQA. Specifically, we first cut the document (for each dialogue) into different chunks. Then, we consider it as a valid unanswerable sample only when there are chunks that have "high overlaps" with the agent's answer to be removed and the rest of the chunks show "low overlaps" with the agent's answer. We use the 4-gram recall score between each chunk and the agent's answer (which measures the ratios of the answer's 4-gram phrases are within each chunk) as the metric to measure their overlaps, and consider it higher than 0.5 as "high overlaps" and lower than 0.1 as "low overlaps".

18

C ChatQA Instruction Tuning Templates

C.1 Stage-1: Supervised Fine-tuning

The format template of LLM inputs in stage-1 is as follows:

```
System: This is a chat between a user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the user's questions
based on the context. The assistant should also indicate when the answer cannot be
found in the context.

User: {Question 1}

Assistant: {Answer 1}

...

User: {Latest Question}

Assistant:
```

We use the {Latest Answer} from Assistant as the supervision for model outputs.

C.2 Stage-2: Context-Enhanced Instruction Tuning

Based on the stage-1 format template, the LLM inputs in stage-2 adds {Context for Latest Question} and {Instruction} from User, as depicted below:

```
System: This is a chat between a user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the user's questions
based on the context. The assistant should also indicate when the answer cannot be
found in the context.

{Context for Latest Question}

User: {Instruction} + {Question 1}

Assistant: {Answer 1}

...

User: {Latest Question}

Assistant:
```

Same as Stage-1, we use the {Latest Answer} from Assistant as the supervision for model outputs.

As for the {Instruction}, we use different instructions for different datasets based on the answer types. Details can be found below:

"`Please give a full and complete answer for the question.`" This is for datasets with long answers. We use it for the HumanAnnotatedConvQA or SyntheticConvQA.

"`Answer the following question with a short span. The answer needs to be just in a few words.`" This is for datasets with short answers. We use it for SQuAD1.1, SQuAD2.0, NarrativeQA, DROP, ROPES, NewsQA, and Quoref.

"`Answer the following question with a number from context or the math arithmetic using +, -, *, or /.`" This is for datasets using arithmetic calculation or extracting numbers from the context. We use it for the TAT-QA dataset where questions need the arithmetic calculation or extracting numbers from the context.

19

"Answer the following question with a short span, or a full and complete answer." This is for datasets with both short and long answers. The model needs to tell whether to generate short or long answers based on the questions. We use it for the TAT-QA dataset where questions do not need the arithmetic calculation, since we observe the lengths of the corresponding answers can be both short and long.

For the stage-1 SFT data used in stage-2 instruction tuning, we keep the same format template as in stage-1. In other words, the {Context for Latest Question} and {Instruction} are empty.

D   Synthetic Data Generation

The instruction for GPT-3.5-Turbo comprises three parts: 1) system role guiding the model to provide helpful answers, 2) examples of conversational QA indicating the required data types, and 3) a document that directs the model to generate conversational QA based on its content. We collect 7k documents (average ∼1k words per document) from common crawl, which cover a wide range of domains. Each document is used for generation of a single conversational QA sample, which leads to a total of 7k multi-turn QA dialogues with an average of 4.4 user-agent turns per dialogue (called SyntheticConvQA).

We use a two-step prompting on GPT-3.5-Turbo for obtaining a high-quality multi-turn QA conversation. In step-1, we prompt GPT-3.5-Turbo to generate a multi-turn QA conversation based on a given document. However, we find that the generated conversation lacks pronouns or common nouns to refer back to words or phrases in dialogue history. Hence, in step-2, we prompt GPT-3.5-Turbo to revise the generated conversation by adding pronouns or common nouns. For both stages, we give GPT-3.5-Turbo few-shot examples to ensure a high-quality generation.

The in-context few-shot prompt in step-1 is as follows (aims to generate multi-turn QA conversation):

```
You are a helpful, pattern-following assistant.

role-user:
PersonA has access to information in the following passage:
###There are certain important aspects of college preparation. The student preparing
for college should define their goals and interests. The student should also plan
financially and academically. We have a checklist that helps students get ready
financially and academically.###
Generate a conversation of around four turns between PersonA and PersonU where
PersonU needs information from PersonA that is in the passage. Use only information
found in the passage.
PersonU starts the conversation and only asks questions. Most of PersonU's questions
are follow-up or continuations of previous questions.
Conversation:

role-assistant:
PersonU: I need to know how to prepare for college.
PersonA: You should first define your goals and interest and be sure to know the
costs and how to plan financially and academically for college.
PersonU: Is there something I can use as a guide to get ready for it?
PersonA: you can use our checklist to get ready financially and academically.
PersonU: Any other tip?
PersonA: Be sure to check out the reasons to attend a College or Career schools.
PersonU: Ok, that sounds good. Can you give me more tips?
PersonA: Another thing to take into consideration is to explore your career options.

role-user:
PersonA has access to information in the following passage:
###There are several disability compensation benefits. One of the benefits is the
annual clothing allowance. This allowance covers the cost of purchasing new clothes
up to a limit of 200 dollars annually.###
Generate a conversation of around two turns between PersonA and PersonU where
```

20

PersonU needs information from PersonA that is in the passage. Use only information found in the passage.
PersonU starts the conversation and only asks questions. Most of PersonU's questions are follow-up or continuations of previous questions.
Conversation:

role-assistant:
PersonU: I heard about a disability benefit for clothing. What is it about?
PersonA: This is a disability compensation benefit known as an annual clothing allowance.
PersonU: I am quite interested about such benefits.
PersonA: Would you like to find out if you are eligible?

role-user:
PersonA has access to information in the following passage:
###Vascepa (icosapent ethyl) is a prescription fish oil medication approved for treating adults with extremely high triglycerides, defined as a triglyceride level of 500 mg/dL or higher. While using this medicine, you may need frequent blood tests at your doctor's office. Vascepa is only part of a treatment program that may also include diet, exercise, and weight control. Follow your doctor's instructions very closely. You should not use Vascepa if you are allergic to icosapent. To make sure Vascepa is safe for you, tell your doctor if you have: 1 liver disease; 2 diabetes; 3 a thyroid disorder; 4 problems with your pancreas; 5 a bleeding or blood-clotting disorder; 6 if you are allergic to fish or shellfish; or. 7 if you drink large amounts of alcohol, Vascepa (icosapent) works in the liver and bloodstream to reduce very low-density triglycerides.###
Generate a conversation of around three turns between PersonA and PersonU where PersonU needs information from PersonA that is in the passage. Use only information found in the passage.
PersonU starts the conversation and only asks questions. Most of PersonU's questions are follow-up or continuations of previous questions.
Conversation:

role-assistant:
PersonU: I heard about a medication called Vascepa. What is it used for?
PersonA: Vascepa is a prescription fish oil medication used for treating adults with extremely high triglyceride levels.
PersonU: Are there any side effects or precautions I should be aware of?
PersonA: While using Vascepa, you may need frequent blood tests, and it's important to follow your doctor's instructions closely. It's important to talk to your doctor about any allergies or existing conditions you have.
PersonU: Is it effective in lowering cholesterol levels?
PersonA: Vascepa is used together with a low-fat diet and other treatments to lower total cholesterol in people with severe high cholesterol.

role-user:
PersonA has access to information in the following passage:

{Document}

Generate a conversation of around three turns between PersonA and PersonU where PersonU needs information from PersonA that is in the passage. Use only information found in the passage.
PersonU starts the conversation and only asks questions. Most of PersonU's questions are follow-up or continuations of previous questions.
Conversation:

role-assistant:

21

The in-context few-shot prompt in step-2 is as follows (aims to revise the multi-turn QA conversation by adding pronouns and common nouns):

```
You are a helpful, pattern-following assistant.

role-user:
Given the following conversation between PersonU and PersonA:
###PersonU: I've heard about adobo sauce in Mexican cuisine. Can you tell me more
about it?
PersonA: Adobo sauce is a dark red sauce made from ground chiles, herbs, and
vinegar. It's used as a marinade and cooking sauce in Mexican and Southwestern
cuisines.
PersonU: What are the main ingredients in adobo sauce?
PersonA: The main ingredients are typically chile peppers, garlic, and vinegar.
However, there can be variations in the recipe.
PersonU: How can I make adobo sauce at home?
PersonA: One way to make adobo sauce at home is by using smoked jalapeno peppers
called chipotles. You can stew the chipotles in a sauce with tomatoes, garlic,
vinegar, salt, and spices. It's also possible to find canned adobo sauce in the
Hispanic foods section of many grocery stores.
PersonU: Are there any other uses for adobo sauce?
PersonA: Adobo sauce can be used as a condiment or as a marinade for various dishes.
It enhances the flavor of meats and can be used as a cooking and serving sauce.
Additionally, adobo is also used in a stew-like presentation of meat prepared in
this sauce.
PersonU: Is adobo sauce easy to preserve?
PersonA: Yes, you can preserve adobo sauce by freezing or canning it when you're
done making it. This allows you to store it for up to a year and enjoy its flavors
whenever you want.###
Modify PersonU's questions to use pronouns or common nouns to refer back to words
or phrases in PersonU's previous questions or PersonA's previous responses.

role-assistant:
PersonU: I've heard about adobo sauce in Mexican cuisine. Can you provide more
information about it?
PersonA: Adobo sauce is a dark red sauce made from ground chiles, herbs, and vinegar.
It's used as a marinade and cooking sauce in Mexican and Southwestern cuisines.
PersonU: What are the main ingredients in the sauce?
PersonA: The main ingredients are typically chile peppers, garlic, and vinegar.
However, there can be variations in the recipe.
PersonU: How can I make the sauce at home?
PersonA: One way to make it at home is by using smoked jalapeno peppers called
chipotles. You can stew them in a sauce with tomatoes, garlic, vinegar, salt, and
spices. It's also possible to find canned adobo sauce in the Hispanic foods section
of many grocery stores.
PersonU: Are there any other uses for it?
PersonA: It can be used as a condiment or as a marinade for various dishes. It
enhances the flavor of meats and can be used as a cooking and serving sauce.
Additionally, it is also used in a stew-like presentation of meat prepared in
this sauce.
PersonU: Is it easy to preserve?
PersonA: Yes, you can preserve it by freezing or canning it when you're done
making it. This allows you to store it for up to a year and enjoy its flavors
whenever you want.

role-user:
Given the following conversation between PersonU and PersonA:
###PersonU: I've heard about a test called urinalysis. What is it used for?
PersonA: Urinalysis is a test that evaluates a sample of your urine. It is used to
```

22

detect and assess a wide range of disorders, such as urinary tract infection, kidney disease, and diabetes. It examines the appearance, concentration, and content of urine.
PersonU: Can it help diagnose bladder cancer?
PersonA: Yes, urinalysis can be used as a screening and diagnostic tool for bladder cancer. One way to test for bladder cancer is to check for blood in the urine, which is called hematuria. This can be done during a urinalysis by examining the urine cells under a special microscope.
PersonU: What other conditions or diseases can be detected through urinalysis?
PersonA: Urinalysis is commonly used to diagnose urinary tract or kidney infections, evaluate causes of kidney failure, screen for the progression of chronic conditions like diabetes mellitus and high blood pressure, and it can be used in combination with other tests to diagnose certain diseases. It is a versatile test that helps detect abnormalities in the urinary tract.
PersonU: Where can I get a urinalysis done?
PersonA: Urinalysis can be performed in various healthcare settings, including doctors' offices, urgent care facilities, laboratories, and hospitals. You can provide a urine sample in a specimen cup, and the sample can be analyzed on-site or sent to a laboratory for testing.
PersonU: How much urine is typically needed for a urinalysis?
PersonA: Usually, only small amounts of urine, ranging from 30 to 60 ml, are required for urinalysis testing. It's a non-invasive and relatively simple test that provides valuable information about your health.###
Modify PersonU's questions to use pronouns or common nouns to refer back to words or phrases in PersonU's previous questions or PersonA's previous responses.

role-assistant:
PersonU: I've heard about a test called urinalysis. What is it used for?
PersonA: Urinalysis is a test that evaluates a sample of your urine. It's used to detect and assess a wide range of disorders, such as urinary tract infection, kidney disease, and diabetes. It examines the appearance, concentration, and content of urine.
PersonU: Can it help diagnose bladder cancer?
PersonA: Yes, urinalysis can be used as a screening and diagnostic tool for bladder cancer. One way to test for bladder cancer is to check for blood in the urine, which is called hematuria. This can be done during a urinalysis by examining the urine cells under a special microscope.
PersonU: What other conditions or diseases can be detected through this test?
PersonA: Urinalysis is commonly used to diagnose urinary tract or kidney infections, evaluate causes of kidney failure, screen for the progression of chronic conditions like diabetes mellitus and high blood pressure, and it can be used in combination with other tests to diagnose certain diseases. It is a versatile test that helps detect abnormalities in the urinary tract.
PersonU: Where can I go to get this test done?
PersonA: Urinalysis can be performed in various healthcare settings, including doctors' offices, urgent care facilities, laboratories, and hospitals. You can provide a urine sample in a specimen cup, and the sample can be analyzed on-site or sent to a laboratory for testing.
PersonU: How much urine is typically needed for the test?
PersonA: Usually, only small amounts of urine, ranging from 30 to 60 ml, are required for urinalysis testing. It's a non-invasive and relatively simple test that provides valuable information about your health.

role-user:
Given the following conversation between PersonU and PersonA:

{multi-turn QA conversation}

Modify PersonU's questions to use pronouns or common nouns to refer back to words

23

or phrases in PersonU's previous questions or PersonA's previous responses.

```
role-assistant:
```

E   More Details and Results for Retrieval in Conversational QA

E.1   Query Rewriting Prompts for GPT-3.5-Turbo

The in-context few-shot query rewriting prompt we use for GPT-3.5-Turbo is as follows:

```
You are a helpful, pattern-following assistant.

role-user:
Given the following conversation between PersonU and PersonA:
PersonU: Hello, I would like to know what to do if I do not agree with any decision.
PersonA: disagree with our decision about your monthly income adjustment amounts?
PersonU: no. Where can I find my SHIP contact information?
PersonA: You can find your local SHIP contact information in the back of your
Medicare & You 2020 Handbook online.
PersonU: and how do they calculate the adjustments?
Instead of having this entire conversation, how can PersonU get what he or she is
looking for using a single question? Respond with that question.

role-assistant:
How is the calculation for adjustments made by SHIP determined?

role-user:
Given the following conversation between PersonU and PersonA:
PersonU: I need to know how to prepare for college.
PersonA: You should first define your goals and interest and be sure to know the
costs and how to plan financially and academically for college.
PersonU: Is there something I can use as a guide to get ready for it?
Instead of having this entire conversation, how can PersonU get what he or she is
looking for using a single question? Respond with that question.

role-assistant:
What resources or guides can I use to help me prepare for college?

role-user:
Given the following conversation between PersonU and PersonA:

{Dialogue History + Latest Question}

Instead of having this entire conversation, how can PersonU get what he or she is
looking for using a single question? Respond with that question.

role-assistant:
```

E.2   More Results for Retrieval in Conversational QA

In Table 9, we show comprehensive comparisons between fine-tuning and rewriting methods.

Interestingly, we find that compared to only using the rewritten query as input (Dragon + Rewrite (w/ single query only)), giving additional dialog history (Dragon + Rewrite (w/ dialog history)) makes the average scores significantly drop. This is because Dragon is originally pretrained on single-turn queries, it will naturally have better generalization ability when a single-turn rewritten

24

Models	Average		Doc2Dial		QuAC		QReCC		TopiOCQA		INSCIT	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-5*	top-20*	top-5*	top-20*
Dragon (w/ dialog history)	46.29	73.09	43.33	75.61	56.8	82.86	46.17	81.96	57.68	78.80	27.49	46.22
Dragon + Rewrite (w/ dialog history)	47.57	74.12	44.54	76.98	57.23	83.04	46.45	82.60	60.94	81.74	28.69	46.22
Dragon + Rewrite (w/ single query only)	**54.46**	**80.13**	47.60	80.60	47.10	77.15	51.73	85.78	73.07	88.19	52.79	68.92
Dragon + Fine-tune (w/ dialog history)	52.72	80.67	48.94	83.01	52.64	81.95	50.73	87.17	67.86	86.28	43.43	64.94
Dragon + Fine-tune + Rewrite (w/ dialog hisotry)	**53.17**	**80.84**	49.30	84.64	55.04	83.23	51.23	87.99	60.50	81.03	49.80	67.33

Table 9: Comprehensive multi-turn retrieval results across five datasets.

query instead of a multi-turn conversation is provided. And the rewritten query already contains sufficient information from the dialog history.

In addition, we observe that "Dragon + Fine-tune" performs on par with "Dragon + Fine-tune + Rewrite". In other words, for the multi-turn fine-tuning method, replacing original query with rewritten query as inputs yields comparable results. This is because rewritten query will not provide much additional information for the model since the dialog history has already been provided, and it might even cause negative effects (e.g., results on the TopiOCQA dataset) since it makes the whole conversation input not natural. This once again demonstrates how effective the fine-tuning method is in equipping the model with the capability to comprehend the multi-turn context.

F Details of CHATRAG BENCH

F.1 Dataset Statistics and Details

Doc2Dial   (D2D)(Feng et al., 2020), a dataset covering DMV, SSA, VA, and Student Aid domains with an average document length of 101K words. We use the test set of Doc2Dial for the evaluation. It consists of 719 dialogs with 3939 user-agent turns. License of this dataset is "Apache License 2.0".

QuAC   (Choi et al., 2018), an open-domain dataset with an average document length of 15K words and unanswerable cases. We use the validation set of QuAC for the evaluation since its test set cannot be directly obtained. Its validation set consists of 1000 dialogs with 7354 user-agent turns. Among these 7354 user-agent turns, there are 1486 (around 20.2%) unanswerable questions. For the combined evaluation of answerable and unanswerable questions, we set the ground truth response for the unanswerable question as "`Sorry. I cannot find the answer based on the context.`", same as the one we set in our stage-2 tuning. For the fair comparison, we replace the unanswerable responses in the baseline models (i.e., Llama2-SFT/Chat, GPT-3.5-Turbo, GPT-4) with the same sentence (Details can be found in Appendix H). License of this dataset is "CC BY-SA 4.0".

QReCC   (Anantha et al., 2021), an open-domain dataset with an average document length of 5K words and maximum document length of 20K words. We use the test set of QReCC for the evaluation. Its test set includes some dialogue samples source from the QuAC dataset. To avoid overlap with other benchmark datasets, we remove those QuAC-sourced samples, resulting in 2805 user-agent turns. License of this dataset is "Apache License 2.0".

TopiOCQA   (TCQA)(Adlakha et al., 2022) uses the whole Wikipedia as the documents. We use the validation set of TopiOCQA since its test set is not available yet. Its validation set consists of 205 dialogs with 2514 user-agent turns. Each question has a total of four human-written answers. It has unanswerable cases. However, we do not include this dataset in unanswerable case evaluation because we only find 46 questions with no answer being found, which is not sufficient enough for the unanswerable case evaluation. Given that this dataset use the whole Wikipedia as the retrieval corpus, we leverage its topic information for multi-turn retrieval since we focused more on conversational QA over documents which typically narrow down to a specific topic or domain. License of this dataset is "Attribution-NonCommercial-ShareAlike 4.0 International".

INSCIT   (Wu et al., 2023) use the whole Wikipedia as the documents. We use the validation set of TopiOCQA since its test set is not available yet. Its validation set consists of 86 dialogues with 502 user-agent turns. Each question has an average of 1.9 human-written answers. It has unanswerable response for a few questions. However, we do not include this dataset in unanswerable case evaluation

25

because every question always has at least one answerable response, which means every question is answerable. INSCIT also use the whole Wikipedia as the retrieval corpus. Hence, we also leverage its topic information for multi-turn retrieval for the same reason mentioned in TopiOCQA. License of this dataset is "arXiv.org perpetual, non-exclusive license".

CoQA   We use the validation set of CoQA since its test set cannot be directly obtained. Its validation set consists of 500 dialogues with 7983 user-agent turns. Each question has a total of four human-written answers. It has unanswerable cases. However, we do not include this dataset in unanswerable case evaluation for the same reason as in the TopiOCQA dataset. We only find 13 questions with no answer being found, which is not sufficient enough for the unanswerable case evaluation. CoQA contains passages from seven domains. Five of these public ones are under the following licenses: Literature and Wikipedia passages are shared under "CC BY-SA 4.0 license"; Children's stories are collected from MCTest which comes with "MSR-LA license"; Middle/High school exam passages are collected from RACE which comes with its own license; and News passages are collected from the DeepMind CNN dataset which comes with "Apache license".

DoQA   We use the test set of DoQA for the evaluation. Its test set consists of 1200 dialogues with 5394 user-agent turns across Cooking, Travel and Movie domains. Among 5394 user-agent turns, there are 1479 (around 27.4%) unanswerable questions. We use the same strategy as the one mentioned in QuAC dataset to incorporate unanswerable samples into the evaluation. License of this dataset is "CC BY-SA 4.0".

ConvFinQA   We use the validation set of ConvFinQA since its test set cannot be directly obtained. Its validation set consists of 421 dialogues with 1490 user-agent turns. License of this dataset is "MIT License".

SQA   We use the test set of SQA for the evaluation. Its test set consists of 1025 dialogues with 3100 user-agent turns. License of this dataset is "Microsoft Research Data License Agreement".

HybriDial   We use the test set of HybriDial for the evaluation. Its test set consists of 243 dialogues with 1111 user-agent turns. License of this dataset is "arXiv.org - Non-exclusive license".

F.2   Details of Incorporating Table as Context

We use the markdown format to process the table in the document (i.e., using "|" to separate each cell in the table). However, as the number of rows in a table increases, it is harder for the model to link the values of cells in the last few rows to their corresponding headers in the table. To address this issue, we append the headers to the corresponding cell values in each row, so as to make it easier for the model to capture the connections between the cells and their respective headers. For example, a original markdown table

```
| Boat | Crew | Nation | Date | Meet | Location |
| M1x | Mahe | FR | 2009 | | Poznan |
| M2- | Hamish | US | 2012 | Olympics | Lucerne |
| M2+ | Igor | DE | 1994 | | Indianapolis |
```

will be converted to

```
| M1x (Boat) | Crew: Mahe | Nation: FR | Date: 2009 | Meet: | Location: Poznan |
| M2- (Boat) | Crew: Hamish | Nation: US | Date: 2012 | Meet: Olympics | Location: Paris |
| M2+ (Boat) | Crew: Igor | Nation: DE | Date: 1994 | Meet: | Location: Indianapolis |
```

We apply this table processing strategy to all tabular datasets.

F.3   Prompts for the Benchmarks

We adopt the prompt format mentioned in §C.2 for benchmarking ChatQA, Llama2-Chat, Llama3-Instruct, GPT-3.5-Turbo, GPT-4, and GPT-4-Turbo. For Llama2-Chat and Llama3-Instruct, we modify the prompt based on their own templates, and we convert the prompt into the chat completion

26

template for OpenAI models. We keep the {Context for Latest Question} the same for all models. We adjust the {Instruction} for the baselines (i.e., Llama2-Chat, GPT-3.5-Turbo and GPT-4) to ensure they do not get sub-optimal results. We use different instructions for test benchmarks with different answer types (e.g., long answer, short answer, arithmetic calculation). The {Instruction} for all models on the benchmarks are as follows:

F.3.1 ChatQA

We keep the {Instruction} consistent between the data blends in stage-2 and test benchmarks based on different answer types.

We use "`Please give a full and complete answer for the question.`" for DoQA, IN-SCIT, HybriDial, Doc2Dial, QuAC, and QReCC, since these datasets generally have long answers for the questions.

We use "`Answer the following question with a short span, or a full and complete answer.`" for SQA and TopiOCQA, since these datasets have both short and long answers based on the questions.

We use "`Answer the following question with a short span. The answer needs to be just in a few words.`" for CoQA, since it generally has short answers for the questions.

We use "`Answer the following question with a number from context or the math arithmetic using +, -, *, or /.`" for ConvFinQA, since this dataset requires the model to either extract numbers from the context or do arithmetic calculation. We will calculate the number based on the arithmetic formula the model generates and compare it with the gold answer.

F.3.2 Llama2-Chat & Llama3-Instruct

Note that we tried original Llama2-Chat prompt template[10], which gives slightly worse results compared to using the one in §C.2. We have tried several {Instruction} for Llama2-Chat. We find the ones below works the best.

We use "`Please give an answer in just one sentence.`" for DoQA, INSCIT, HybriDial, Doc2Dial, QuAC, and QReCC, since these datasets generally have long answers but within one sentence. We notice that it is important to give Llama2-Chat specific instruction like "one sentence" instead of "full and complete answer" to prevent the model from generating very long answers.

We use "`Answer the following question with a short span, or one sentence.`" for TopiOCQA, since this dataset has both short and long answers based on the questions, and the long answers are generally within one sentence.

We use "`Answer the following questions with one or a list of entities.`" for SQA, since the answer for this dataset always consists of one or a list of entities from the context.

We use "`Answer the following question with a short span. The answer needs to be just in a few words.`" for CoQA, since it generally has short answers for the questions.

We use "`Answer the following question with just a number from context or just the math arithmetic using +, -, *, or /.`" for ConvFinQA, since this dataset requires the model to either extract numbers from the context or do arithmetic calculation. We extract the arithmetic formula generated by the model and use a calculator to get its final result.

F.3.3 GPT-3.5-Turbo & GPT-4 & GPT-4-Turbo

We have tried several {Instruction} for GPT-3.5-Turbo, GPT-4, and GPT-4-Turbo, we find the ones below works the best.

We use "`Please give an answer in just one sentence.`" for DoQA, INSCIT, HybriDial, Doc2Dial, QuAC, and QReCC, since these datasets generally have long answers but within one sentence. Similar to Llama2-Chat, we also find OpenAI models tend to generate quite long answers

---

[10]https://gpus.llm-utils.org/llama-2-prompt-template/

27

given the instruction of "full and complete answer". Hence, we make the instruction more specific (i.e., "one sentence") to prevent the model from generating very long answers.

We use "`Answer the following questions in JUST a few words or one sentence.`" for TopiOCQA, since this dataset has both short and long answers based on the questions, and the long answers are generally within one sentence.

We use "`Answer the following questions with one or a list of entities. Do not give a detailed explanation. Answer needs to be as short as possible.`" for SQA. We find that OpenAI models often give detailed explanation for SQA dataset unless specifically instructed not to.

We use "`Answer the following question with a short span. The answer needs to be just in a few words.`" for CoQA, since it generally has short answers for the questions.

We use "`Answer the following questions with just a number from context or just the math arithmetic using +, -, *, or /.`" for ConvFinQA, since this dataset requires the model to either extract numbers from the context or do arithmetic calculation. We extract the arithmetic formula generated by the model and use a calculator to get its final result.

G Human Evaluation

G.1 Evaluation Setup

You will be provided some text paragraphs (context), a dialogue history (between User and Assistant), a final question from User that could be answered based on the given text paragraphs, and two respones to the final question.
You will be given a human rewritten reference answer for the final question. **Note that the reference answer would indicate it when the answer cannot be found.**
Note that the dialogue history could be empty.

What you need to do:
1. After reading the text paragraphs (context), the dialog history, the question, and the reference answer, you have to select the response that is more factually correct.
2. If both are equally factually correct, select both. If neither is factually correct, select neither.

======= **Attention** =======
--> Consider both responses are **Equally Good**, **as long as both of them answer the question**, and do **NOT** have unverifiable or factually incorrect content.
--> Please penalize **verbosity** if the response has some content that are **NOT directly relevant to the question**
--> Please penalize it if some content in the response can **NOT be verifiable.**.

Context:
${context}

Dialogue History:
${dialogue_history}

Final Question:
${question}

Reference Answer:
${reference}

Response1: ${response1}

Response2: ${response2}

Select an option

Response1	1
Response2	2
Both	3
Neither	4

Submit

Figure 4: Human evaluation instructions and the interface for annotators.

We conduct human evaluation across the ten test benchmark datasets. We randomly select 60 samples for each dataset, and each sample is labelled by three annotators, which results in a total of 1800 annotations.

We ask the annotators to verify the facts in models' outputs and determine which model provides a more accurate response to the question. We use Amazon Mechanical Turk platform to conduct the human evaluation. The details of our human evaluation instructions and the interface shown to annotators are illustrated in Figure 4.

28

	Ours Win	Tie	GPT-4 Win
Average	13.81%	69.09%	17.10%
Doc2Dial	14.29%	68.00%	17.71%
QuAC	11.67%	73.33%	15.00%
QReCC	11.11%	77.22%	11.67%
CoQA	7.78%	80.00%	12.22%
DoQA	22.78%	57.78%	19.44%
ConvFinQA	16.67%	67.78%	15.55%
SQA	11.11%	61.67%	27.22%
TopiOCQA	19.31%	60.69%	20.00%
HybriDial	7.78%	78.33%	13.89%
INSCIT	15.56%	66.11%	18.33%

Table 10: Human evaluation (A/B testing) comparing our ChatQA-1.0-70B to GPT-4-0613 on CHATRAG BENCH. In terms of average scores, our model and GPT-4 are tie most of the time (∼69.1%), and GPT-4-0613 achieves slightly higher win rate (∼3.3%) than ours.

G.2 Full Human Evaluation Results

The full human evaluation results over the 10 datasets on CHATRAG BENCH are shown in Table 10. We first find that our ChatQA-1.0-70B and GPT-4 are tie most of the time (∼69.1%), and GPT-4 achieves slightly higher win rate (∼3.3%) than ours. This further confirms our model has powerful capability to produce correct answers. Second, we find that our model achieves slightly better win rate than GPT-4 in ConvFinQA, which indicates the strong arithmetic calculation ability of our model. Third, we find that GPT-4 achieves significantly better win rate in SQA task, which suggests that there is still a gap between our model and GPT-4 on tabular reasoning task.

H Unanswerable Case Evaluation

H.1 Evaluation Setup

For unanswerable case, we consider the model indicating that the question cannot be answered as correct. For answerable cases, we consider the model not indicating the question is unanswerable as correct (i.e., the model giving an answer). We find that ChatQA always generates "`Sorry. I cannot find the answer based on the context`" when it finds the question is unanswerable, since we use this sentence as a response for unanswerable questions in the stage-2 tuning. As for Llama2-Chat, Llama3-Instruct, GPT-3.5-Turbo, GPT-4, and GPT-4-Turbo, their outputs for unanswerable questions usually follow some specific patterns. Below, we list the heuristic matching patterns derived from all generated samples in DoQA and QuAC datasets to determine if the model suggests that the question is unanswerable. When we find the generated output contains the pattern listed below, we interpret it as an indication that the question is unanswerable, and we will replace its answer with "`Sorry. I cannot find the answer based on the context.`" for the F1 evaluation.

```
cannot find, can't find, not able to, unable to, does not provide, cannot
provide, cannot answer, couldnot answer, can't answer, couldn't answer,
cannot be found, cannot be determined, do not have, couldn't find, no
information, does not mention, doesn't mention, not explicitly mentioned,
not explicitly explain, can not find, could not find, does not specify,
doesn't provide, doesn't specify, there is no, not mentioned, don't have,
don't know, does not include, doesn't include, does not contain, doesn't
contain, not provided, does not indicate, doesn't indicate, does not
disclose, doesn't disclose
```

H.2 Ablation Studies on the Number of Unanswerable Samples for Training

In Table 11, we find that compared to using SyntheticConvQA, using HumanAnnotatedConvQA significantly increase the average accuracies on both QuAC and DoQA. It is because the unanswerable annotations are in higher quality for human-annotated data, which significantly improve the accuracy

29

Models		Avg Both	Avg QuAC	QuAC (no*)	QuAC (yes*)	Avg DoQA	DoQA (no*)	DoQA (yes*)	ChatRAG
ChatQA-1.0-70B	- 1k unanswerable[†]	76.88	80.89	75.10	86.67	72.88	64.49	81.26	54.16
	- **1.5k unanswerable**[†]	**77.25**	80.76	77.66	83.85	73.74	68.81	78.67	54.14
	- 2k unanswerable[†]	77.10	80.82	77.59	84.05	73.38	67.95	78.80	53.86
	- 2.5k unanswerable[†]	75.87	78.81	73.76	83.85	72.93	66.54	79.31	53.78
	- SyntheticConvQA[◇]	69.84	72.92	55.38	90.42	66.77	45.09	88.45	54.08

Table 11: Accuracies on answerable and unanswerable samples across QuAC and DoQA datasets. Avg-Both is the averaged score between QuAC and DoQA, and ChatRAG is the average score on the CHATRAG BENCH. [◇] denotes that the HumanAnnotatedConvQA is replaced with the Synthetic-ConvQA. * "no" denotes the unanswerable samples, while "yes" denotes the answerable samples. [†]We conduct ablation studies in terms of the number of unanswerable samples in HumanAnnotatedConvQA. We use 1.5k unanswerable samples for our final ChatQA models, as it produces both high-quality generation and less hallucination.

for unanswerable cases. Furthermore, we conduct ablation studies in terms of the number of unanswerable samples in HumanAnnotatedConvQA. We find that using a small amount of unanswerable samples (e.g., 1.5k) is able to achieve remarkable results on the unanswerable evaluation, as well as the overall CHATRAG BENCH score. Interestingly, incorporating more unanswerable samples does not necessarily lead to higher accuracy scores in the unanswerable evaluation.

30

I Case Study

Context	Conversation	Assistant Responses
The process for Board hearings has changed. If you're appealing a VA decision under the new process, keep reading below to learn how to request and prepare for a hearing. You'll follow the new process if you're : Appealing a VA decision you received on or after February 19 , 2019. Participating in the Rapid Appeals Modernization Program Opting into the new process using VA Form 10182 If you're using VA Form 9 to appeal a VA decision you received before February 19, 2019, go to the Board of Veterans' Appeals BVA Hearing page ... The Board of Veterans Appeals schedules hearings in docket order the order in which they receive requests ...	User: What can I learn at my Board hearing? Assistant: Do you want to learn exactly what to expect? ... User: Is it necessary for me to consider requesting a Board hearing? Assistant: You can request a video conference hearing if you want to provide testimony to a judge about your appeal. User: when will I follow the new process for a Board hearing? Assistant:	ChatQA-1.0-70B: If you're appealing a VA decision you received on or after February 19, 2019, you'll follow the new process. GPT-4-0613: You'll follow the new process for a Board hearing if you're appealing a VA decision you received on or after February 19, 2019, participating in the Rapid Appeals Modernization Program, or opting into the new process using VA Form 10182. Ground Truth: You will follow the new process if appealing a VA decision if you received it on or after February 19, 2019.
title: How did Arya survive this confrontation unscathed? There was no time or intention to turn her into White Walker. Let us look at the scene from S04E04. A white walker brings a baby holding in his lap and puts it on an ice table (or whatever it is) for the arrival of the Night King. ... When Arya drops the dagger, Night King looks surprised. This shows he doesn't know what would happen next. There was no time to think and turn her into a white walker by touch. There was no planning in this scene like S04E04. ...	User: How did Arya survive this confrontation unscathed? Assistant: Arya was quick enough to react, drop the dagger and do the rest. User: How old is the child that they lay on the altar of ice? Assistant:	ChatQA-1.0-70B: The child appeared to be a baby. GPT-4-0613: The text does not provide information on the exact age of the child that was laid on the ice altar. Ground Truth: a baby

Massif	Type of nature reserve	Preserved area
Chornohora	Carpathian Biosphere	2476.8 ha
Uholka / Wide Meadow	Carpathian Biosphere	11860 ha
Svydovets	Carpathian Biosphere	3030.5 ha
Maramoros	Carpathian Biosphere	2243.6 ha
Rozok	Presov Preserved	67.1 ha
Jasmund	Jasmund National Park	492.5 ha
...		

User: which of the forests is a carpathian biosphere reserve?

Assistant: Chornohora, Uholka / Wide Meadow, Svydovets, Maramoros, Kuziy / Trybushany

User: which of those covers more than 3100 ha?

Assistant:

ChatQA-1.0-70B: Uholka / Wide Meadow

GPT-4-0613: Chornohora, Uholka / Wide Meadow, Svydovets, Maramoros

Ground Truth: Uholka / Wide Meadow

title: Summary of enemy aircraft damaged/destroyed (John B. England)

Date	Location	No.	Type	Status
24 April 1944	Munich, Germany	3	Me-110	Destroyed
27 May 1944	Strasbourg, France	1	Me-109	Damaged
18 August 1944	Paris, France	0.5	Me-109	Destroyed
13 September 1944	Nordhausen, Germany	2.5	Me-109	Destroyed
6 October 1944	Berlin, Germany	1	Me-109	Damaged
6 October 1944	Berlin, Germany	2	Me-109	Destroyed
27 November 1944	Magdeburg, Germany	4	FW-190	Destroyed
14 January 1945	Berlin, Germany	1	Me-109	Destroyed
...				

User: what were all of the dates in which john b. england destroyed or damaged enemy planes?

Assistant: 24 April 1944, 27 May 1944, 18 August 1944, 13 September 1944, 6 October 1944, 27 November 1944, 14 January 1945 ...

User: on which of those dates did john b. england destroy or damage more than one enemy plane?

Assistant:

ChatQA-1.0-70B: 6 October 1944, 13 September 1944, 27 November 1944 (missing 24 April 1944)

GPT-4-0613: 24 April 1944, 13 September 1944, 6 October 1944, 27 November 1944

Ground Truth: 24 April 1944, 13 September 1944, 6 October 1944, 27 November 1944

Table 12: Case studies on ChatQA-1.0-70B and GPT-4-0613. We use blue color to highlight the relevant context for the last user question. And we use green and red color to highlight the correct and incorrect parts in the assistant responses.

In Table 12, we show four examples of ChatQA-1.0-70B and GPT-4-0613 outputs. More examples can be found in Appendix I. The first example (about *Board hearing*) is a simple information seeking problem, and both ChatQA-1.0-70B and GPT-4 answer correctly. In the second example (about *Arya*), the model needs to find implicit information (highlighted in blue) for giving the answer. GPT-4 tends to be safe on providing an answer, and it replies that the context does not provide exact information about the age which is also correct.

Both the third and fourth examples require the model to have good tabular understanding and reasoning ability. In the third example (about *Massif*), ChatQA-1.0-70B gives correct answer by comparing the size of preserved area against 3100 ha, while GPT-4 fails to do so. In the fourth example (about *John B. England*), ChatQA-1.0-70B correctly lists three dates but misses one date, while GPT-4 correctly answers the question.

More examples on ChatQA-1.0-13B, ChatQA-1.0-70B, GPT-3.5-Turbo-0613, and GPT-4-0613 can be found in Table 13. We use blue color to highlight the relevant context for the last user question. And we use green and red color to highlight the correct and incorrect parts in the assistant responses.

31

Context			Conversation	Assistant Responses

Example 1

Pos	Rider	Time/Retired
1	Loris Capirossi	38:04.730
2	Valentino Rossi	+0.180
3	Jeremy McWilliams	+0.534
4	Tohru Ukawa	+0.537
5	Shinya Nakano	+0.742
6	Ralf Waldmann	+7.019
7	Franco Battaini	+20.889
8	Stefano Perugini	+20.891
9	Jason Vincent	+21.310
10	Anthony West	+26.816
11	Alex Hofmann	+26.933
...		
23	Arno Visscher	+1:40.635
24	Henk Van De Lagemaat	+1 Lap
Ret	Maurice Bolwerk	Retirement
Ret	Andre Romein	Retirement
Ret	Roberto Rolfo	Retirement
Ret	Marcellino Lucchi	Retirement

Conversation:

User: what racers are shown?

Assistant: Loris Capirossi, Valentino Rossi, Jeremy McWilliams, Tohru Ukawa, ... , Henk Van De Lagemaat, Maurice Bolwerk, Roberto Rolfo, Marcellino Lucchi.

User: of those, what times were recorded?

Assistant: 38:04.730, +0.180, +0.534, +0.537, ... , +1:40.635, +1 Lap, Retirement.

User: of those times, which were not actual times?

Assistant: +1 Lap, Retirement

User: which of was an actual finish?

Assistant:

Assistant Responses:

ChatQA-1.0-13B: Loris Capirossi

ChatQA-1.0-70B: +1 Lap

GPT-3.5-Turbo-0613: 38:04.730

GPT-4-0613: 38:04.730

Ground Truth: +1 Lap

Example 2

Context: 2014 compared to 2013 mst 2019s net sales decreased $ 305 million, or 3% in 2014 as compared to 2013 ... space systems 2019 operating results included the following (in millions):

	net sales	backlog at year-end
2015	$ 9105	$ 17400
2014	$ 9202	$ 20300
2013	$ 9288	$ 21400

2015 compared to 2014 space systems 2019 net sales in 2015 decreased $ 97 million, or 1%, compared to 2014. these decreases were partially offset by higher net sales of approximately $ 235 million for businesses acquired in 2014; and approximately $ 75 million for the orion program due to increased volume.

Conversation:

User: what was the average backlog in 2015?

Assistant: 17400

User: what was it in 2014?

Assistant: 20300

User: what is the sum of those years?

Assistant: 17400 + 20300

User: what was the average backlog in 2013?

Assistant: 21400

User: what is the total sum for all 3 years?

Assistant:

Assistant Responses:

ChatQA-1.0-13B: 17400 + 20300 + 21400

ChatQA-1.0-70B: 17400 + 20300 + 21400

GPT-3.5-Turbo-0613: 9288 + 9202 + 9105

GPT-4-0613: 17400 + 20300 + 21400

Ground Truth: 17400 + 20300 + 21400

Example 3

Context: The XVI World Rhythmic Gymnastics Championships were held in Brussels, Belgium, on November 20-22, 1992.

Rank	Gymnast	Country	Point
1	Oxana Kostina	Russia	9.775
2	Maria Petrova	Bulgaria	9.700
3	Diana Popova	Bulgaria	9.625
4	Carmen Acedo	Spain	9.625
5	Irina Deleanu	Romania	9.550
6	Ekaterina S.	Ukraine	9.525
7	Viktoriya Yani	Ukraine	9.475
8	Maria Sansaridou	Greece	9.425

Conversation:

User: What is the XVI World Rhythmic Gymnastics Championships?

Assistant: The XVI World Rhythmic Gymnastics Championships were held in Brussels, Belgium, on November 20-22, 1992

User: How many Gymnastics participated this championship?

Assistant:

Assistant Responses:

ChatQA-1.0-13B: There are 8 Gymnastics participated in the championship.

ChatQA-1.0-70B: There are 8 gymnasts* participated in the championship.

GPT-3.5-Turbo-0613: There were a total of 8 gymnasts* who participated in this championship.

GPT-4-0613: The context does not provide the total number of gymnasts who participated in the XVI World Rhythmic Gymnastics Championships.

Ground Truth: There are 8 gymnasts participated this championship.

*Both models correct the User's typo of "Gymnastics" to "gymnasts"

32

Context	Conversation	Assistant Responses
The 3000 metres steeplechase or 3000-meter steeplechase is the most common distance for the steeplechase in track and field. It is an obstacle race over the distance of the 3000 metres, which derives its name from the horse racing steeplechase. | Year	Athlete	Location
1968	Jouko Kuha (FIN)	Stockholm
1969	Vladimiras Dudinas (URS)	Kiev
1970	Kerry O'Brien (AUS)	Berlin
1971	Kerry O'Brien (AUS)	Adelaide
1972	Anders Gärderud (SWE)	Helsinki
1973	Ben Jipcho (KEN)	Helsinki
1974	Anders Gärderud (SWE)	Helsinki
1975	Anders Gärderud (SWE)	Stockholm
1976	Anders Gärderud (SWE)	Montréal
1977	Michael Karst (FRG)	Stockholm | ...	User: Hello, can you tell me about the 3000 metres steeplechase? Assistant: The 3000 metres steeplechase is the most common distance for the steeplechase in track and field that goes over the distance of the 3000 metres, which derives its name from the horse racing steeplechase. User: Can you tell me the men that are part of it? Assistant: Some of the men are Henry Rono and Michael Karst. User: Which locations did Kerry O'Brien run for? Assistant:	ChatQA-1.0-13B: He ran in Berlin, Adelaide, and Stockholm. ChatQA-1.0-70B: Kerry O'Brien ran in Adelaide and Berlin. GPT-3.5-Turbo-0613: Kerry O'Brien ran in Berlin, Adelaide, and Champaign. GPT-4-0613: Kerry O'Brien ran in Berlin and Adelaide. Ground Truth: He ran for both Berlin and Adelaide.
title: Non-EU family member traveling to Schengen countries. source: You can travel visa-free only if you have a "residence" card of a family member of a union citizen" issued under article 10 of the free movement directive (or if you are a citizen of a country whose nationals don't require visas in the Schengen area). A residence card issued by Romania to the spouse of a Romanian citizen is not necessarily an article 10 card. If you do not have an article 10 card, EU countries other than Romania must issue you a visa free of charge and on the basis of an expedited procedure. The grounds for refusal of that visa are strictly limited. These provisions are also dependent on your traveling with or joining your spouse. In case you are not, you'll need a regular Schengen visa. As Relaxed pointed out in a comment, it appears that your concern stems in part from the fact that Romania is not yet a full participant in the Schengen agreement. These rules apply to all EU countries, not just Schengen countries, because they are based on the EU right of freedom of movement.	User: Can I travel from Romania to the Schengen area visa-free? Assistant: You can travel visa-free only if you have a "residence card of a family member of a union citizen" issued under article 10 of the free movement directive User: I have a temporary residence card also, can we travel together visa-free? Assistant: If you do not have an article 10 card, EU countries other than Romania must issue you a visa free of charge and on the basis of an expedited procedure. User: How do I get a article 10 card? Assistant:	ChatQA-1.0-13B: The grounds for refusal of that visa are strictly limited. These provisions are also dependent on your traveling with or joining your spouse. In case you are not, you'll need a regular Schengen visa. ChatQA-1.0-70B: Sorry. I cannot find the answer based on the context. GPT-3.5-Turbo-0613: To obtain an article 10 card, you would need to meet the requirements set out in the free movement directive and apply for it through the appropriate channels in Romania. GPT-4-0613: The text does not provide information on how to obtain an article 10 card.* Ground Truth: Sorry. I cannot find the answer based on the context. *In the evaluation, we will detect this unanswerable pattern and replace it with "Sorry. I cannot find the answer based on the context" to match with the ground truth.
title: Bengal fox. "Vulpes bengalensis" is a relatively small fox with an elongated muzzle, long, pointed ears, and a bushy tail about 50 to 60% of the length of the head and body. Its dorsal pelage is very variable, but mostly grayish and paler ventrally; its legs tend to be brownish or rufous. It is more daintily built than "Vulpes vulpes". The tail is bushy with a prominent black tip which distinguishes it from "V. vulpes". Back of ears are dark brown with black margin. Its rhinarium is naked and the lips are black, with small black hair patches on upper part of nuzzle (shaft) in front of eyes. title: Red fox. The red fox is considered a more specialised form of "Vulpes" than the Afghan, corsac and Bengal foxes in the direction of size and adaptation to carnivory; the skull displays far fewer neotenous traits than in other species ... title: Bengal fox. The ears have the same colour as the nape or maybe darker, but not having a dark patch as in "V. vulpes". Extensive variation in coat colour exists across populations and seasonally within populations, but generally varies from grey to pale brown ...	User: what is red fox? Assistant: It is the largest of the true foxes and one of the most widely distributed members of the order Carnivora. User: where did it originate? Assistant: The species is Eurasian in origin. User: how did it evolve? Assistant: It is considered a more specialised form of "Vulpes" than the Afghan, corsac and Bengal foxes in the direction of size and adaptation to carnivory. User: what is the appearance of this third type of fox? Assistant:	ChatQA-1.0-13B: The ears have the same colour as the nape or maybe darker, but not having a dark patch as in "V. vulpes". Extensive variation in coat colour exists across populations and seasonally within populations, but generally varies from grey to pale brown. ChatQA-1.0-70B: "Vulpes bengalensis" is a relatively small fox with an elongated muzzle, long, pointed ears, and a bushy tail about 50 to 60% of the length of the head and body. GPT-3.5-Turbo-0613: I'm sorry, but I couldn't find any information about a third type of fox in the given context. GPT-4-0613: The context does not provide information on a third type of fox. Ground Truth: It is a relatively small fox with an elongated muzzle, long, pointed ears, and a bushy tail about 50 to 60% of the length of the head and body.

33

Context	Conversation	Assistant Responses
Zimmer Holdings, Inc. Notes to consolidated financial statements (continued) the unaudited pro forma results for 2003 include events or changes in circumstances indicate that the carrying $ 90.4 million of expense related to centerpulse hip and knee value of an asset may not be recoverable... inventories payments that are contingent on the growth of implex inventories at december 31, 2003 and 2002, consist of product sales through 2006. the net value transferred at the following (in millions) : closing will be approximately $ 89 million, which includes.	User: what was the total of inventories in 2003? Assistant: 527.7 User: and what was it in 2002? Assistant: 257.6 User: how much, then, did the 2003 amount represent in relation to this 2002 one? Assistant: 527.7 / 257.6 User: and what is that excluding the portion equivalent to the 2002 amount? Assistant: 527.7 / 257.6 - 1 User: between those same two years, what was the change in the total of finished goods, in millions? Assistant:	ChatQA-1.0-13B: 384.3 - 206.7 ChatQA-1.0-70B: 384.3 - 206.7 GPT-3.5-turbo-0613: 384.3 - 206.7 GPT-4-0613: 384.3 - 206.7 Ground Truth: 384.3 - 206.7

	finished goods	progress	inventories
2003	$ 384.3	90.8	$ 527.7
2002	$ 206.7	50.9	$ 257.6

made by zimmer to implex pursuant to their existing alliance raw materials and work in progress 90.8 50.9 arrangement, escrow and other items.
the acquisition will be inventory step-up 52.6 2013 accounted for under the purchase method of accounting.

Context	Conversation	Assistant Responses
market for the registrant 2019s common equity, related stockholder matters and issuer purchases of equity securities the following graph compares annual total return of our common stock, the standard & poor 2019s 500 composite stock index (201cs&p 500 index 201d) and our peer group (201cloews peer group 201d) for the five years ended december 31, 2015. the graph assumes that the value of the investment in our common stock , the s&p 500 index and the loews peer group was $ 100 on december 31, 2010 and that all dividends were reinvested.	User: from 2010 to 2011, what was the change in the value of the loews common stock? Assistant: 97.37 - 100 User: and what was this change as a portion of the 2010 value of that stock? Assistant: (97.37 - 100) / 100 User: and from 2010 to 2012, what was the change in the value of that stock? Assistant: 106.04 - 100 User: how much did this change represent in relation to that 2010 value? Assistant:	ChatQA-1.0-13B: (106.04 - 100) / 100 ChatQA-1.0-70B: (106.04 - 100) / 100 GPT-3.5-turbo-0613: (106.04 - 100) / 100 GPT-4-0613: (106.04 - 100) / 100 Ground Truth: (106.04 - 100) / 100

	common stock	s&p 500	peer group
2010	100.0	100.0	100.0
2011	97.37	102.11	101.59
2012	106.04	118.45	115.19
2013	126.23	156.82	145.12
2014	110.59	178.29	152.84
2015	101.72	180.75	144.70

the loews peer group consists of the following companies that are industry competitors of our principal operating subsidiaries...
dividend information we have paid quarterly cash dividends on loews common stock in each year since 1967.

Table 13: Case study for ChatQA-1.0-13B, ChatQA-1.0-70B, GPT-3.5-turbo-0613, and GPT-4-0613. We use blue color to highlight the relevant context for the last user question. And we use green and red color to highlight the correct and incorrect parts in the assistant responses.

34

J   Human Annotated Conversational QA Data

J.1   Statistics of User-Agent Turns
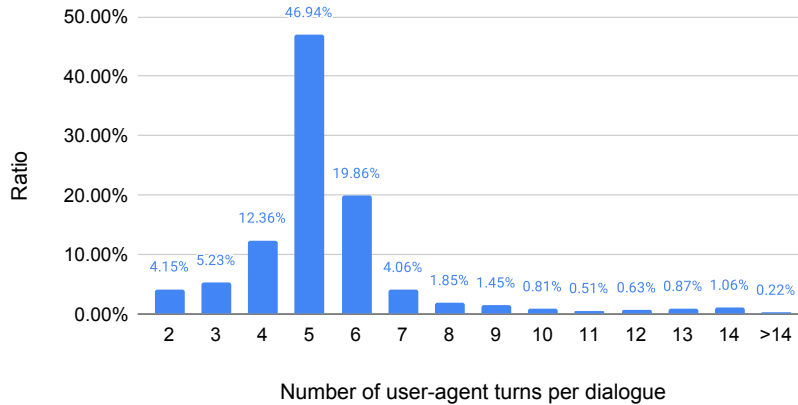
Figure 5: Statistics regarding the number of user-agent turns for each dialogue

Figure 5 illustrates the statistics regarding the number of user-agent turns per dialogue over the 7k human annotated conversational QA data.

J.2   Guidelines for Conversational QA Data Collection

In this section, we demonstrate the guidelines we provide for the conversational QA data collection. We request the annotators to act as both user and agent to construct the conversational QA sample based on the given document, which makes the annotation more efficient. The guidelines consists of three parts: 1) What does conversational QA samples look like; 2) What kinds of conversational QA samples we need; and 3) What we need to annotate.

J.2.1   What does conversational QA samples look like

One conversational QA sample is based on a document that we provide. The user and agent behaviors are as follows:

- User behavior: 1) Ask questions to agent based on the given document; 2) Answer questions from agent when agent wants to clarify something.
- Agent behavior: 1) Answer questions from user based on the document; 2) Ask questions to the user when the user's question is not clear, or too general/broad.

J.2.2   What kinds of multi-turn QA samples we need

We list the requirements for user's questions and agent's responses below.

User's Questions

- User's questions can refer to his/her previous (or multiple turns before) question.
- User's questions can also refer to the previous (or multiple turns before) answer from the agent.
- Try to use pronouns or common nouns to replace the entities mentioned before.
- Try to make the user's questions diverse. For the same type of questions, try to use different ways to express it in different dialogue annotations.
- Other user question types we need

35

- Given the answer from agent, ask for extra information (e.g., What else ...; any other ...; is there more ...).
- Switch the topic and starting a new thread in conversation.
- Ask two questions at the same time.

Agent's Response

- Try to make the agent's answer within 1-2 sentences. If the answer has to be long, try to make it concise.
- Try not to directly copy the whole relevant context from the document as the answer. Instead, try to construct the answer by paraphrasing the chosen context.
- Try to come up with a small proportion of cases where agent asks questions to user to clarify something. Specifically, when the user's question is too broad or not clear, the agent needs to narrow down the scope by asking clarification questions to see which specific aspects that the user cares more about.

J.2.3   What we need to annotate

We list what we need to annotate for each conversation below.

- For each document, we need to annotate the user's questions and corresponding agent's responses. The average number of user-agent turns per conversation needs to be around five.
- For each user's question, we need to annotate all the relevant context within the document.

36

K   Results of All Models on CHATRAG BENCH

Models	Avg. w/o HDial	Avg. All	D2D	QuAC	QReCC	CoQA	DoQA	CFQA	SQA	TCQA	HDial	INSCIT
Llama2-SFT-7B*	33.24	34.81	30.26	19.21	37.55	62.75	21.76	34.43	32.18	32.88	48.96	28.16
Llama2-Chat-7B	37.96	38.86	33.27	25.83	46.02	72.28	33.15	36.58	26.14	36.68	47.02	31.67
ChatQA-1.0-7B	46.96	47.71	37.88	29.69	46.97	76.61	41.57	51.61	61.87	45.45	54.51	30.96
GPT-SFT-8B	32.48	34.46	31.03	20.07	37.69	59.24	21.72	15.44	40.06	38.17	52.29	28.86
ChatQA-1.0-8B	48.86	49.36	36.76	33.95	45.54	77.90	44.65	61.68	60.74	47.03	53.81	31.50
Llama2-SFT-13B*	36.08	37.69	30.68	21.59	38.25	69.52	21.70	41.14	37.85	35.26	52.22	28.73
Llama2-Chat-13B	39.29	40.34	34.74	27.89	47.19	72.50	32.60	41.54	25.39	39.25	49.82	32.52
ChatQA-1.0-13B	50.27	50.86	38.05	34.28	48.06	77.23	43.31	65.44	66.41	48.88	56.19	30.79
GPT-SFT-22B*	37.11	38.83	32.98	22.83	39.92	68.19	22.83	30.54	47.01	38.07	54.22	31.66
ChatQA-1.0-22B	53.01	53.53	39.71	36.89	50.47	80.32	44.88	72.08	69.08	48.83	58.20	34.11
Llama2-SFT-70B*	41.85	43.22	34.42	25.65	41.88	73.04	28.21	46.64	58.90	37.20	55.52	30.71
Llama2-Chat-70B	44.64	45.21	36.87	32.47	49.40	80.41	38.97	46.85	37.62	44.31	50.35	34.88
ChatQA-1.0-70B	53.89	54.14	38.90	41.82	48.05	78.57	51.94	73.69	69.14	50.98	56.44	31.90
Llama3-Instruct-8B	46.76	46.70	33.59	33.60	45.70	75.26	37.33	58.05	59.72	44.96	46.20	32.59
Llama3-ChatQA-1.5-8B	53.99	55.17	39.33	39.73	49.03	76.46	49.60	78.46	73.28	49.96	65.76	30.10
Llama3-Instruct-70B	52.95	52.52	37.88	36.96	51.34	76.98	41.24	76.60	69.61	49.72	48.59	36.23
Llama3-ChatQA-1.5-70B	**57.14**	58.25	41.26	38.82	51.40	78.44	50.76	81.88	83.82	55.63	68.27	32.31
Command R+	51.40	50.93	33.51	34.16	49.77	69.71	40.67	71.21	74.07	53.77	46.70	35.76
GPT-3.5-Turbo-0613	50.69	50.37	34.83	37.17	50.46	79.33	41.11	73.15	60.63	44.30	47.42	35.27
GPT-4-0613	54.35	53.90	34.16	40.29	52.01	77.42	43.39	81.28	79.21	45.09	49.81	36.34
GPT-4-Turbo-2024-04-09	**54.72**	54.03	35.35	40.10	51.46	77.73	41.60	84.16	79.98	48.32	47.86	33.75

Table 14: Full zero-shot results on CHATRAG BENCH. * SFT denotes that only stage-1 supervised fine-tuning is applied. We add Llama2-13B and our in-house 8B/22B models.

Models	Avg. w/o HDial	Avg. All	D2D	QuAC	QReCC	CoQA	DoQA	CFQA	SQA	TCQA	HDial	INSCIT
Llama3-Instruct-8B	46.76	46.70	33.59	33.60	45.70	75.26	37.33	58.05	59.72	44.96	46.20	32.59
Llama3-Instruct-8B + Stage-2	**54.41**	55.59	39.18	36.54	49.13	82.43	48.91	77.99	72.59	50.88	66.21	32.04
Llama3-ChatQA-1.5-8B	**53.99**	55.17	39.33	39.73	49.03	76.46	49.60	78.46	73.28	49.96	65.76	30.10
Llama3-Instruct-70B	52.95	52.52	37.88	36.96	51.34	76.98	41.24	76.60	69.61	49.72	48.59	36.23
Llama3-Instruct-70B + Stage-2	**56.80**	57.88	40.22	40.58	50.66	76.44	51.48	82.75	80.30	56.21	67.57	32.58
Llama3-ChatQA-1.5-70B	**57.14**	58.25	41.26	38.82	51.40	78.44	50.76	81.88	83.82	55.63	68.27	32.31

Table 15: Studies on directly applying stage-2 instruction tuning to the powerful Llama3-Instruct models.

In Table 14, we report zero-shot results of all ChatQA models on CHATRAG BENCH. ChatQA models generally give great improvements compared to the corresponding SFT/Chat/Instruct counterparts.

In Table 15, we study the effectiveness of directly applying stage-2 instruction tuning to the powerful Llama3-Instruct models. Llama3-Instruct models can be considered as those that have already finished the Stage-1 SFT training. They use a much larger dataset than the stage-1 training for Llama3-ChatQA-1.5. However, we find that "Llama3-Instruct-8B/70B + Stage-2" performs comparable to Llama3-ChatQA-1.5-8B/70B. This further suggests that the stage-2 instruction tuning plays a key role in improving model's capability in conversational QA and RAG.

L   Training Details

For all ChatQA models, in stage-1 SFT, we use a learning rate of 5e-6, and train 1000 iterations with a global batch size of 128; and in stage-2 instruction tuning, we use a learning rate of 3e-7, and train 3300 iterations with a global batch size of 64. We use 256 NVIDIA A100 GPUs for training ChatQA-1.0-70B and Llama3-ChatQA-1.5-70B models, and it takes around three hours for stage-1 training and around six hours for the stage-2 training. We use 64 NVIDIA A100 GPUs for training ChatQA-1.0-7B and Llama3-ChatQA-1.5-8B models, and it takes around one and half hours for stage-1 training and around three hours for stage-2 training.

37

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have comprehensive experimental results in § 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our models at the beginning of the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

38

Justification: This work does not propose any theory assumptions and does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the details of the data collection for training and evaluations, which can be found in section 3, 4, 5 and appendix C, D, E and H. We also open-source model weights, training data, evaluation data, and scripts for reproducing our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

39

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We open-source model weights, training data, evaluation data, and scripts for reproducing our results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Dataset details can be found in section 5 and Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: All results are zero-shot and deterministic (generations are using greedy search, retrieval scores are also deterministic). Hence, the results will not have any variances.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

40

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the compute resources information in the training details section of the Appendix.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research is conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics

   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential societal impacts and negative impacts at the beginning of the Appendix.

    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

41

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We include licenses of the datasets in the Appendix F.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

42

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide details explanations of how we collect human-annotated data and synthetic data in the Appendix D and Appendix J.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

43

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

44