RAG-QA Arena: Evaluating Domain Robustness for Long-Form Retrieval-Augmented Question Answering

Rujun Han[2†]   Yuhao Zhang[3†]   Peng Qi[4†]   Yumo Xu[1]   Jenyuan Wang[1]
Lan Liu[1]   William Yang Wang [5†]   Bonan Min[1]   Vittorio Castelli[1]

[1]AWS AI Labs [2]Google [3]Samaya.ai [4]Orby.ai [5]University of California, Santa Barbara
rujunh@google.com; yuhao@samaya.ai; peng@orby.ai; william@cs.ucsb.edu
{liuall,yumomxu,jenwan,bonanmin,vittorca}@amazon.com

Abstract

Question answering based on retrieval-augmented generation (RAG-QA) is an important research topic in NLP and has a wide range of real-world applications. However, most existing datasets for this task are either constructed using a single source corpus or consist of short extractive answers, which fall short of evaluating large language model (LLM) based RAG-QA systems on cross-domain generalization. To address these limitations, we create Long-form RobustQA (LFRQA), a new dataset comprising human-written long-form answers that integrate short extractive answers from multiple documents into a single, coherent narrative, covering 26K queries and large corpora across seven different domains. We further propose RAG-QA ARENA by directly comparing model-generated answers against LFRQA's answers using LLMs as evaluators. We show via extensive experiments that RAG-QA ARENA evaluation and human judgments on answer quality are highly correlated. Moreover, only 41.3% of the most competitive LLM's answers are preferred to LFRQA's answers, demonstrating RAG-QA ARENA as a challenging evaluation platform for future research.[1]

1 Introduction

Traditional reading comprehension task is constrained to fixed contexts (Rajpurkar et al., 2016; Kočiský et al., 2018; Huang et al., 2019). It is inadequate at addressing real-world questions, where no context is readily provided for a system to find answers. Such open-ended questions require a system to identify answers in an enormous knowledge base (e.g., Wikipedia) that is computationally prohibitive to feed into question answering systems

Figure 1: LFRQA annotation example. There are three documents (some text removed for brevity) relevant to the query. We instruct annotators to combine RO-BUSTQA's answers into a coherent long-form answer with added text if necessary. Citations [1], [2] and [3] indicate the supporting documents of each sentence.

such as large language models (LLMs). Retrieval-augmented generative question answering (RAG-QA) becomes an effective tool to filter out massive amounts of noise and select only a few highly relevant passages for LLM-based QA models.

The wide applications of RAG-QA (Gao et al., 2023) necessitate the evaluation of systems' out-of-domain (OOD) performances, because a real-world system often confronts new data unseen during training. Existing popular benchmark datasets such as Natural Questions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) consist solely of Wikipedia or Web documents, which fall short at measuring OOD performances.

ROBUSTQA (Han et al., 2023) was the first dataset created to benchmark cross-domain robustness for RAG-QA. However, as illustrated by the yellow highlights in Figure 1, ROBUSTQA follows NQ's annotation format with short answer spans extracted from the documents. Such data format is not the most suitable reference answer to evaluate the current leading LLMs that typically gen-

---

[1]Code: https://github.com/awslabs/rag-qa-arena

Dataset Name	Answers grounded in corpus	Long-form answers	Multiple documents	Coherent answers	Multiple domains	Human annotated	# Test queries
LFRQA	✓	✓	✓	✓	✓	✓	16.1K
ROBUSTQA (Han et al., 2023)	✓	✗	✓	✗	✓	✓	16.1K
NQ (Kwiatkowski et al., 2019)	✓	✗	✗	✗	✗	✓	3.6K
MULTIHOP-RAG (Tang and Yang, 2024)	✓	✗	✓	✓	✓	✓	2.5K
ASQA (Stelmakh et al., 2022)	✓	✓	✓	✓	✗	✓	1.0K
LONGFACT (Wei et al., 2024)	✗	✓	✗	✓	✓	✗	2.3K
ELI5 (Fan et al., 2019)	✗	✓	✓	✓	✗	✓	25.0K

Table 1: Comparison of datasets. LFRQA distinguishes from previous work by uniquely encompassing seven features: 1) RAG-QA dataset with answers annotated based on underlying corpus; 2) Long-form answers of paragraph length; 3) Multiple documents that provide different facts/views; 4) Coherent answers that handle conflicting information; 5) Multiple-domain corpus to benchmark domain robustness; 6) Human annotated high-quality answers; 7) Large-scale evaluation set.

erate long-form responses with multiple pieces of information combined in one coherent narrative (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2024). Consequently, token overlap metrics used in the extractive QA era (Karpukhin et al., 2020; Izacard et al., 2024) will penalize unfairly on the additional supporting tokens generated by LLMs, resulting in extremely low overlap scores. As an example, Fig. 1's extractive answers have poor Exact-Match or F_1 scores with the final long-form answer. To create a long reference answer, one could simply concatenate these short answers, but the synthesized answers are either incoherent or ill-formatted, as illustrated by examples in Sec. 3.1.

To address these drawbacks, we propose long-form RobustQA (LFRQA) that integrates multiple short extractive answers into a coherent long-form answer. Figure 1 shows an annotation where three extractive answers are combined by annotators to create a comprehensive answer. Table 1 summarizes seven features in LFRQA that make it uniquely beneficial for RAG-QA evaluations. ASQA (Stelmakh et al., 2022) and ELI5 (Fan et al., 2019) are the most similar datasets to LFRQA. However, they are either not directly annotated against the underlying corpus (thus, not RAG-QA), or rely on single-domain data, which is insufficient to benchmark systems' cross-domain performances.

With LFRQA annotations, we propose RAG-QA ARENA that leverages model-based evaluators to directly compare LLMs' answers with LFRQA without the necessity to examine long and potentially noisy retrieved passages. By demonstrating the high correlation with human judges following the same instruction and rubrics, we show that RAG-QA ARENA is an efficient and accurate framework to benchmark the RAG-QA system's cross-domain performances. In this work, we primarily focus on the LLMs used for the QA component, but RAG-

QA ARENA can be easily extended to study retrieval's impact on answer generation quality.

We summarize our contributions: 1) We present LFRQA, the first high-quality and large-scale multi-domain human annotations with coherent long-form answers for RAG-QA. 2) We propose an efficient model-based evaluation framework, RAG-QA ARENA that enables users to directly compare LLMs' answers with ground-truth answers in LFRQA. 3) We build a dashboard incorporating a wide range of leading LLMs and conduct in-depth analysis to show that LFRQA's answers are preferred significantly more to the best LLMs with long context. Therefore, we believe RAG-QA ARENA will serve as a challenging and robust evaluation benchmark for future RAG-QA research.

2 RAG-QA Task Formulation

We briefly introduce the RAG-QA task in this section. Passage retrieval is the first step of a RAG-QA pipeline. Following the passage retrieval set-up in DPR (Karpukhin et al., 2020) and RO-BUSTQA (Han et al., 2023), we denote a collection of documents as $\mathcal{D}$. We split each document $d^i \in \mathcal{D}$ with a fixed length N tokens and obtain a collection of M ($\geq |\mathcal{D}|$) passages denoted as $\mathcal{C} = \{p_1, p_2, ...p_m, ...p_M\}$, where p_m is a passage. Given a question q, the passage retrieval task is to select K most relevant passages for q with a retriever $\mathcal{R}$ from $\mathcal{C}$. Formally, $\mathcal{R}(q, \mathcal{C}) \rightarrow \mathcal{C}_q$.

Upon receiving top K passages or $\mathcal{C}_q$, a QA model reads them as context to generate an answer for the query. Unlike the extractive QA setting in ROBUSTQA and NQ, we adopt *generative QA* as it is most compatible with the generative nature of the leading LLMs with the flexibility to produce free-form answers. The answer generation task can be modeled as $\Sigma_1^T \mathcal{P}_q(w_t|w_{0:t-1}; \mathcal{C}_q)$, where $\mathcal{P}$ is an LLM. We focus on the variations of $\mathcal{P}$ and fix

Queries	RobustQA answers	LFRQA answers
Will the word schadenfreude be understood in an English text?	Yes it would not be understood by the majority of English-speaking adults No mostly anecdotal evidence here suggests not	The term "schadenfreude" has seen an increased use in the English language and is understood to mean deriving pleasure from others' misfortunes. However, it would not be widely recognized by the majority of English-speaking adults, as its current usage is still negligible. Furthermore, there is a consensus that the term isn't generally understood, and anecdotes reflect that only a minor portion of certain professional groups, such as IT personnel, might be familiar with it.
How much can I trust Tor?	can not say Tor can solve all your problems many ways to compromise your identity give you considerably more protection than browsing directly Tor is probably what I would recommend don't let it be the only type of thread in your safety net use Tor everywhere except on an Internet connection	You cannot say Tor can solve all your problems, as there can be many ways to compromise your identity, so while Tor will be useful as one thread in a safety net, don't let it be the only type of thread in your safety net. Tor is probably to be recommended as it gives considerably more protection than browsing directly, but be aware that your ISP is in a position to see that your IP address is using Tor, even though it can't tell what you're using Tor for. You should use Tor everywhere, except on an Internet connection which can be strongly associated with you.

Figure 2: LFRQA v.s. ROBUSTQA. Citations are removed in LFRQA's answers, and a few answer spans are removed for clarity. Green and orange texts represent positive and negative opinions, respectively.

$\mathcal{R}$ in this work.

In real-world applications, we deploy RAG-QA systems into various domains such as healthcare, finance, and technology whose corpus and query types may not be well covered in a trained retriever and LLM. Lewis et al. (2021) and Han et al. (2023) show significant performance gaps between in-domain and out-of-domain data for RAG-QA systems, further verifying domain adaptation problems. Therefore, it is crucial to gauge the domain robustness of RAG-QA based on LLMs, and LFRQA helps achieve this evaluation goal.

3 Data Creation

LFRQA consists of two types of QA samples: 1) new annotations in Finance ([FI]), Lifestyle ([LI]), Recreation ([RE]), Technology ([TE]), Science ([SI]), and Writing ([WR]) domains; 2) adapted long-form BioASQ ([BI]). We describe the details of both QA samples in the following sections.

3.1 Annotated Data

Following ROBUSTQA (Han et al., 2023), LFRQA's new annotations are also based on the LoTTE and FiQA queries and corpus. LoTTE was proposed in the ColBERTv2 paper (Santhanam et al., 2022) and consists of information retrieval (IR) datasets across five domains: lifestyle, recreation, technology, writing, and science, each can have relevant answers coming from either web search or on-line forum. FiQA (Maia et al., 2018) proposes a task, "Opinion-based QA over financial data" that answers finance-related questions from financial corpora such as microblogs, reports, and news. It is important to note that both FiQA and LoTTE are IR datasets with answers as long documents, which may include a large amount of irrelevant information to the query.

As IR datasets, both FiQA and LoTTE could only provide relevant documents to users, as there are no precise answer annotations. ROBUSTQA addresses this short-coming by extracting short answer spans from the long documents in the similar format of NQ, which serves as a high-quality benchmark for extractive RAG-QA. Figure 1 shows an example where the yellow highlights in Documents 1-3 are the extracted answers to the question.

Limitations of extractive RAG-QA. In the era of LLMs, models' responses to user queries are often long and comprehensive (OpenAI, 2024; Anthropic, 2024; MetaAI, 2024; Jiang et al., 2024), which the short, extractive reference answers in ROBUSTQA are no longer the most compatible format to evaluate against. First, in ROBUSTQA, annotators are limited to only taking 3 answer spans per relevant document, each with no more than 16 words. This process could result in a loss of useful information to help answer the query. Second, to reconcile multiple extractive references for model evaluation, prior work in extractive RAG-QA (Karpukhin et al., 2020; Han et al., 2023; Izacard et al., 2024) adopt the maximum of token overlaps between a model prediction and a list of references to compute EM or F_1 metrics, which penalizes unfairly the long-form responses from modern LLMs. Finally, if we naively concatenate or list all short answer spans as shown by examples in Figure 2, the combined answers are often too ill-formatted or incoherent as ground-truth answers.

LFRQA addresses all of these drawbacks by instructing the annotators to integrate all short answers in ROBUSTQA into a coherent long-form answer. Below, we show a summary of our annotation instruction and quality control mechanism.

Annotation instruction. As Fig. 1 shows, a query, all relevant documents, and original short answers (highlighted in the documents) are presented to annotators on a single annotation page. Annotators need to combine all highlighted answers into a single complete and coherent answer. All highlighted

				Test Set		ROBUSTQA		LFRQA	
Domain	**Source**	**Label**	$\|Q\|$	$\|D\|$	$\|P\|$	A/Q	W/A	A/Q	W/A
Biomedical	BioASQ	[BI]	1,956	15,559,026	37,406,880	2.6	2.4	1.0	30.0
Finance	FiQA	[FI]	3,612	57,638	105,777	3.0	9.4	1.0	69.1
Lifestyle	LoTTE	[LI]	2,208	119,461	241,780	5.7	8.7	1.0	99.5
Recreation	LoTTE	[RE]	2,094	166,975	315,203	3.2	7.2	1.0	60.3
Technology	LoTTE	[TE]	2,111	638,509	1,252,402	6.0	8.7	1.0	99.7
Science	LoTTE	[SC]	1,423	1,694,164	3,063,916	5.3	7.8	1.0	92.0
Writing	LoTTE	[WR]	2,695	199,994	347,322	6.2	6.6	1.0	88.0

Table 2: Data (test set) summary: LFRQA v.s. ROBUSTQA. $\|Q\|$, $\|D\|$, $\|P\|$, A/Q, and W/A represent numbers of questions, documents, passages, answers per question, and words per answer, respectively. Each passage consist of 100 words at most. LFRQA has only one answer per query as we integrate multiple answers from ROBUSTQA, which results in more words in (long-form) answers. Dev set statistics can be found in Appendix Table 7.

answers MUST be included; otherwise, the annotation is considered as a failure. Annotators are encouraged to include more information in the documents if it helps to answer the queries. To ensure annotators faithfully use the document information, we request annotators to provide citations after each answer sentence. For example, the first sentence in Fig. 1 is composed using information from Documents 2 and 3. Annotators should add "[2, 3]" after that sentence. We use these citations primarily for data quality control and remove them during the answer evaluation. The actual annotation UI can be found in Appendix Fig. 5.

Quality control. The data annotations are performed by contracted data professionals. We also have a dedicated team of data linguists to validate the annotation quality. Specifically, our data linguists randomly audit 10% of each batch of the annotations, and if the valid answer ratio is $< 90\%$, we send the batch back to the annotators for rework. The process iterates until the valid answer ratio exceeds 90%. Here is a list of failure cases: **1. Incompleteness:** Answers do not include all highlighted answers, or there is clear relevant information in the documents, but not included in the answer. **2. Redundancy:** Clear irrelevant information is included in the answer. **3. Incoherence:** Answers are not coherent or not written in natural English. **4. Citation Error:** Wrong/missing citations, which indicate annotators do not use correct information from the right documents.

3.2 Adapted Data

For the biomedical domain, LFRQA leverages the same set of test queries as in ROBUSTQA, but uses the complete rather than span answers in the BioASQ dataset. The original BioASQ annotations provide two types of answer formats: 1) exact answer, which is the short extractive answers

used in ROBUSTQA; 2) ideal answer, which is a long-form abstractive answer to be consistent with other datasets in this work. We did not perform further annotations. We notice in Table 2 that BioASQ's answers are shorter compared with other datasets. This is due to its dominant amount of factoid queries, which do not require elaborated explanations as in other datasets with more open-ended reasoning questions (Han et al., 2023).

We drop SearchQA in ROBUSTQA as this dataset only has short-form extractive answers, and its documents contain a significant amount of text omission ("...") that prevents us from re-constructing long-form answers.

3.3 Data Statistics and Analysis

Table 2 summarizes the statistics for the test set, which consists of 16K queries across 7 domains. We filter out queries with more than 80 ground-truth documents, resulting in 73 fewer queries compared with ROBUSTQA. Since LFRQA combines multiple short answers, the answer per query ratio (A/Q) is always 1, and the word per answer ratio (W/A) is substantially higher compared with ROBUSTQA. We also annotate a dev set with 10K queries for future model development purposes, and the statistics can be found in Appendix Table 7. We conduct further analysis below to demonstrate the unique contributions of LFRQA in Table 1.

Answers over multiple documents. Figure 3 illustrates the distribution of number of documents used by LFRQA's answers. Specifically, Figure 3a shows that around 65% of the answers use ≥ 2 documents' information. 4.9% of the answers consist of information from 10 or more documents (maximum = 80). In Figure 3b, we divide long-form answers into sentences and show the distribution of the number of documents used per answer sentence. Nearly 22% of the answer sentences combine in-
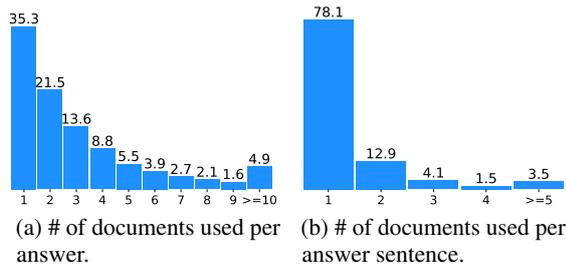
(a) # of documents used per answer.

(b) # of documents used per answer sentence.

Figure 3: Distribution of number (#) of documents used in LFRQA's answers. All numbers are %.

formation from multiple documents. Both show that LFRQA's answers effectively combine information across multiple ground-truth documents. This makes LFRQA challenging for RAG-QA, as it requires identification and aggregation of information across sources.

Coherent answers. LFRQA's answers further organize facts and views across multiple documents in a coherent paragraph. Answers with multiple views are common in the original ROBUSTQA's answers. Conducting a string match of both "yes" and "no" as a leading word in an answer list, we found more than 200 examples with such conflicting information. This does not account for more subtle cases where answers semantically contradict each other. In Figure 2, we show 2 examples with conflicting views. ROBUSTQA's annotations simply list them (separated by new-lines), whereas LFRQA's answers organize them as coherent narratives, with conflicting information reconciled in helpful context.

Fluency. ROBUSTQA's answers are extracted from documents and often cut off unnaturally to satisfy the limit of 16 words. In Figure 2, ROBUSTQA's answers, such as "many ways to compromise your identity" and "mostly anecdotal evidence here suggest no" are incomprehensible without further context, whereas LFRQA's answers are all well written in complete sentences.

All of these features show that LFRQA provides both challenging and high-quality annotations for evaluating RAG-QA systems.

4 RAG-QA Arena

In this section, we propose our evaluation framework RAG-QA ARENA. Inspired by the pairwise human preference evaluation framework such as Chatbot Arena (Chiang et al., 2024), we calculate win-rate and win+tie rate against LFRQA as ground-truth as a metric to gauge systems' RAG-QA quality. Figure 4 illustrates the evaluation framework.
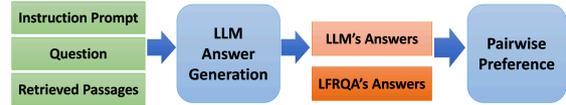
The choice of LFRQA as the target to compare

Figure 4: RAG-QA ARENA framework. Green blocks are LLM's inputs to generate answers. Orange blocks are LLM and LFRQA's answers presented to both human and LLM judges to determine pairwise preferences.

has been partly justified in Sec. 3 for 1) Completeness: the annotation process encourages the inclusion of as much relevant information as possible. 2) Coherence: its answers are written more coherently and naturally than ROBUST-QA as references for LLM generations. Complete and coherent answers can be considered as a comprehensive summary of all relevant information in the entire corpus. This allows us to evaluate generated answers against LFRQA answers only, which is much more informative and concise than using retrieved passages, potentially with a large amount of noise.

We implement human and model-based evaluations with the same instructions and report their correlations. We will show results in Sec. 6.1 that further justify using LFRQA as evaluation targets.

4.1 Human Evaluation

We present a query and a pair of answers (one from LFRQA and one from an LLM), to human annotators. We instruct them to rate their preferences based on three aspects. **1. Helpfulness:** information that is helpful/relevant to answer the query. (Touvron et al., 2023; Bai et al., 2022). **2. Truthfulness:** information that is correct to answer the query. By our definition, truthful information should also be helpful information (Stephanie Lin, 2021; Aisha Khatun, 2024). **3. Completeness:** include as much truthful and helpful information as possible. We further instruct annotators to use Truthfulness (being both truthful and helpful) as the primary criterion since it is stricter than Helpfulness. Helpfulness is used when a decision cannot be made by Truthfulness alone. More details including the definition of aspects, rating categories, and step-by-step guidelines can be found in Appendix A.6 and Fig. 6-7 (annotation interface).

4.2 Model-Based Evaluation

Since human evaluation is too costly, we adopt model-based evaluators for scalable evaluation of LLMs on the entire LFRQA test set.

As for the evaluation approach, we provide LLM-based evaluators with a query and a pair of answers

(including one from LFRQA). Similar to human evaluation, we prompt LLMs to rate their preferences based on the same three aspects above. We only modify the human instruction slightly to be compatible with LLM readable input text, but the majority of the prompt, especially the input data, and rubric, stay the same (Appendix Table 13-15).

For both human and model-based evaluations, we allow "tie" (no preference) as an option. For human evaluation, we take the majority votes from 3 annotators to mitigate biases. If there is no majority vote, we default the label to "tie."

5 Experimental Setup

In this section, we discuss our retriever, LLMs experimented for both answer generation and pairwise evaluations and their prompts in more detail.
Retrieval setting. We employ COLBERTV2 (Santhanam et al., 2022) as our passage retriever, considering its superior performance on the underlying corpus for both ROBUSTQA and LFRQA as shown in Han et al. (2023). We follow the same retrieval setting and split passages into text chunks with 100 consecutive words. We use the top 5 retrieved passages for our main results in Table 3 and experiment with the top 10 passages for further analysis.
Answer generation. We consider LLMs ranked top 25[2] in the Chatbot Arena (Chiang et al., 2024) and their smaller version to show the impact of model sizes. Due to resource and legal constraints, for proprietary LLMs, we only use OpenAI models: a) GPT-4-TURBO (2024-04-09), b) GPT-4O and c) GPT-4-0125-PREVIEW). For public models, we experiment with 1) MIXTRAL-8X22B-INSTRUCT and MIXTRAL-8X7B-INSTRUCT (Jiang et al., 2024); 2) LLAMA-3-70B-INSTRUCT and LLAMA-3-8B-INSTRUCT (MetaAI, 2024); 3) COMMAND R+ and COMMAND R (Gomez, 2024); 4) QWEN1.5-110B-CHAT and QWEN1.5-32B-CHAT (Bai et al., 2023). Answer generation prompt can be found in Appendix Table 11.
Pairwise evaluation. For LLM-based evaluators, we focus on a few larger models with strong context understanding capability, such as GPT-4-TURBO, GPT-4-0125-PREVIEW, GPT-4O, MIXTRAL-8X22B-INSTRUCT and LLAMA-3-70B-INSTRUCT. Appendix Table 13-15 show the details of the pairwise evaluation prompts, including instruction, example prompt template,

and in-context-learning examples. We shuffle the order of the answer pairs so that both human and model judges are not biased by the position of an answer. We select the LLM with the highest correlation with human judgments as the evaluator (Appendix Table 8). We use OpenAI API to run GPT-4 models. We download public models from HuggingFace Hub (HuggingFace, 2024) and run them on up to 8 Nvidia A100 GPUs with PyTorch (1.13.0) and Transformers (4.41.0) whose `tokenizer.apply_chat_template()` function can help adapt the generic prompts to different LLMs' input formats.

We follow OpenAI's recommendation[3] to design prompts with chain-of-thoughts (CoT) (Wei et al., 2022), in-context learning (Dong et al., 2023), and HTML tags as delimiters. We remove the thinking process in model outputs as final answers.

6 Results and Analysis

Leveraging the evaluation framework described in Sec. 4, we first show that LFRQA's ground truth answers are dominantly preferred as answers than ROBUSTQA. Then, we use the same evaluation framework to establish a new leaderboard, RAG-QA ARENA, aiming to reliably measure RAG-QA systems' performances across diverse domains.

6.1 LFRQA v.s. RobustQA

In Sec. 3.3, we demonstrate LFRQA's advantages via data statistics. Here, we show a more rigorous study to highlight the benefits. We subsample 700 queries (100 from each of the 7 domains) and conduct pairwise preference comparisons using both human and model-based evaluations. We compare three types of answers: 1) ROBUSTQA: concatenation of its extractive answers, separated by "\n"; 2) LFRQA: long-form answers in this work; 3) GPT-4's answers based on the top 5 retrieved passages.

Table 4 shows that when compared directly in Row (1), LFRQA dominates ROBUSTQA. When comparing with GPT-4 in Row (2)-(3), LFRQA significantly out-performs GPT-4, but ROBUSTQA significantly under-performs. These results show strong evidence that LFRQA's answers can serve as better ground-truth than ROBUSTQA.

6.2 Quality of Model-based Evaluator

To build RAG-QA ARENA on the LFRQA test set, we need a scalable evaluation method to benchmark

Compared Models	Overall		[BI]		[FI]		[LI]		[RE]		[TE]		[SC]		[WR]	
	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T
GPT-4O* [#1]	**36.9**	**41.0**	**52.9**	**59.3**	38.4	42.3	**25.1**	**27.9**	40.4	46.4	35.6	38.8	<u>42.8</u>	47.6	**28.4**	**31.1**
GPT-4-TURBO [#2]	34.4	<u>39.1</u>	36.0	43.9	40.6	45.1	<u>23.2</u>	<u>26.1</u>	<u>36.7</u>	<u>44.1</u>	**36.6**	**39.6**	42.6	<u>47.9</u>	<u>26.2</u>	<u>29.6</u>
GPT-4-0125-PREVIEW [#6]	28.9	33.7	31.4	40.1	36.8	40.8	18.1	21.3	31.5	38.6	30.4	34.0	34.7	40.5	19.2	22.3
MIXTRAL-8X22B [#3]	<u>34.5</u>	38.8	37.0	46.0	**44.1**	**47.6**	21.3	24.4	34.4	41.0	33.9	36.8	**45.0**	**49.5**	25.9	28.1
MIXTRAL-8X7B [#7]	27.5	31.0	31.9	39.1	35.3	38.5	15.9	18.4	24.8	29.5	30.3	32.1	33.9	38.0	20.0	21.9
LLAMA-3-70B [#8]	21.7	25.2	30.3	37.2	24.6	27.7	12.9	15.1	22.3	27.3	22.4	24.4	25.6	30.0	15.5	18.2
LLAMA-3-8B [#10]	20.4	23.5	34.7	39.6	24.0	27.0	11.2	13.2	19.4	24.7	20.5	22.5	22.3	26.1	12.5	14.4
COMMAND R+ [#9]	21.1	25.8	26.0	33.5	25.8	30.3	13.5	16.4	22.6	30.0	22.4	25.4	24.9	31.2	13.6	16.0
COMMAND R [#11]	11.1	15.2	18.6	26.1	13.0	17.1	5.2	7.4	10.4	17.0	10.3	12.3	14.9	20.2	7.3	9.4
QWEN1.5-110B-CHAT [#4]	33.4	37.8	<u>36.2</u>	<u>44.0</u>	42.6	46.9	22.3	25.1	34.1	40.7	<u>34.8</u>	<u>37.5</u>	40.8	46.1	22.5	25.2
QWEN1.5-32B-CHAT [#5]	32.8	37.1	34.9	42.8	<u>43.2</u>	<u>47.3</u>	20.7	23.7	32.3	38.3	34.0	37.1	40.8	44.8	22.6	25.2

Table 3: Evaluation results on LFRQA test set. W and W+T indicate win and win+tie rate against LFRQA's answers. LLM's answers are generated based on the top 5 passages. **bold** and <u>underline</u> indicate the best and runner-up results. * means using the answer generation prompt w/o CoT. [#n] indicates the Elo ranking in Appendix Table 5.

	Answer Pairs	Human	MBE	Pearson Corr.	Cohen's Kappa
(1)	ROBUSTQA	6.1	1.0	0.82***	0.71
	LFRQA	**83.9**	**95.6**		
	Tie	10.0	3.4		
(2)	GPT-4	**78.1**	**77.1**	0.60***	0.54
	ROBUSTQA	13.7	19.9		
	Tie	8.1	3.0		
(3)	GPT-4	29.9	32.0	0.54***	0.44
	LFRQA	**59.1**	**63.1**		
	Tie	11.0	4.9		
(4)	MIXTRAL	31.7	36.9	0.54***	0.43
	LFRQA	**54.4**	**59.7**		
	Tie	13.9	4.4		
(5)	LLAMA-3	24.0	24.7	0.52***	0.45
	LFRQA	**65.0**	**71.3**		
	Tie	11.0	4.0		

Table 4: Pairwise comparisons between ROBUSTQA, LFRQA, and LLMs. GPT-4: GPT-4-0125-PREVIEW; MIXTRAL: MIXTRAL-8X22B-INSTRUCT; LLAMA-3: LLAMA-3-70B-INSTRUCT. Answer generated based on top 5 passages using ColBERT-v2. **MBE** stands for the model-based evaluator. All numbers are % except for Pearson Corr. and Cohen's Kappa. *** indicates strong correlation with p-values ≪ 0.001.

various LLMs. We rely on model-based evaluation to achieve this goal. Before showing the final dashboard results, we check the quality of our selected evaluator (GPT-4-0125-PREVIEW) in Table 4.

To alleviate model bias, we use three LLMs' answers as benchmark data, and the query set is the same 700 subsample above. Row (3)-(5) use GPT-4-0125-PREVIEW, MIXTRAL-8X22B-INSTRUCT and LLAMA-3-70B-INSTRUCT, respectively. All answers are generated based on the top 5 passages.

We observe that LLM evaluators' numbers align well with the average human scores, except that LLMs tend to predict less "tie." Most importantly, all Pearson Correlation (Freedman et al., 2007) are above 0.52 (with p-values ≪ 0.01), and all Cohen's

Kappa (Cohen, 1960) are above 0.43, both showing strong agreement between model and human judgments. In Appendix Table 8, we show correlation numbers using alternative LLMs as evaluators, but none of them works better than a single GPT-4-0125-PREVIEW model, which we select as our best quality evaluator for RAG-QA ARENA.

6.3 RAG-QA Arena

Finally, we show RAG-QA ARENA's benchmark results. In Table 3 we report each model's win and win+tie rate against LFRQA.

Dashboard leaders. GPT-4O leads the dashboard, with GPT-4-TURBO and MIXTRAL-8X22B-INSTRUCT as close runners-up. GPT-4O performs the best for [BI], [LI], [RE] and [WR] domains, MIXTRAL-8X22B-INSTRUCT leads in [FI] and [SC], and GPT-4-TURBO champions in [TE].

Impact of "no answer found." In RAG-QA, we rely on a passage retriever to provide context, which could be irrelevant. Our prompt (Appendix Table 11) asks an LLM to refrain from answering if it "couldn't find an answer." When we use this answer generation prompt with CoT (the last two lines in the prompt), GPT-4O produces 48.3% "I couldn't find an answer" responses (Appendix Table 9). We randomly sample 20 such examples, and surprisingly found that in 16 cases, GPT-4O puts an answer in its **<thinking>** process, but continues to generate "I couldn't find an answer." Fig. 8-9 show four such examples in comparison with other LLMs' answers with the same prompt, and GPT-4O's new answers without CoT. As the answer generation prompt with CoT only fails for GPT-4O, we remove CoT for GPT-4O, which improves its answer format and reduces the "no-answer" ratio to the level similar to other competitive models.

| | (A) LFRQA only pairs | | | (B) LFRQA + 700 complete pairs | | | (C) LFRQA + 1400 complete pairs | | |
RAG-QA ARENA Ranking	Rating	95% CI	Votes	Rating	95% CI	Votes	Rating	95% CI	Votes
LFRQA	1144	+1/-1	176.7K	1145	+1/-1	176.7K	1146	+1/-1	176.7K
GPT-4O	1066	+5/-5	16.1K	1081	+4/-4	23.1K	1085	+3/-3	30.1K
GPT-4-TURBO	1050	+5/-4	16.1K	1058	+4/-3	23.1K	1065	+3/-2	30.1K
MIXTRAL-8X22B	1049	+4/-4	16.1K	1059	+3/-3	23.1K	1063	+3/-3	30.1K
QWEN1.5-110B-CHAT	1041	+4/-4	16.1K	1047	+4/-3	23.1K	1052	+3/-3	30.1K
QWEN1.5-32B-CHAT	1036	+6/-4	16.1K	1034	+4/-3	23.1K	1037	+3/-3	30.1K
GPT-4-0125-PREVIEW	1008	+6/-5	16.1K	1005	+4/-4	23.1K	1008	+3/-3	30.1K
MIXTRAL-8X7B	991	+4/-4	16.1K	991	+3/-4	23.1K	987	+3/-3	30.1K
LLAMA-3-70B	939	+4/-5	16.1K	931	+4/-4	23.1K	930	+2/-3	30.1K
COMMAND R+	938	+5/-5	16.1K	931	+3/-4	23.1K	924	+3/-3	30.1K
LLAMA-3-8B	924	+6/-6	16.1K	910	+4/-4	23.1K	903	+4/-3	30.1K
COMMAND R	816	+8/-6	16.1K	802	+5/-5	23.1K	796	+4/-4	30.1K

Table 5: Elo rating **including** "couldn't find answer" responses. "LFRQA only" indicates the pairs that always include an LFRQA answer. "LFRQA + N complete pairs" means we subsample N additional pairs evenly across 7 domains and conduct comparison for all pairs of models. We have 11 models, so the total new pairs are $\frac{N \times 11 \times 10}{2}$.

These results raise a research question about the impact of prompt engineering. We emphasize that the goal of RAG-QA ARENA is to propose a reliable evaluation framework, not to conduct extensive prompt engineering or model training to pursue the best RAG-QA system. We provide a dev set of LFRQA, which can be leveraged in future research for model development purposes.

Elo rating. Table 3 shows dashboard results of win and win+tie ratio against LFRQA. We can further convert these pairwise comparisons into Elo ratings similar to Chatbot Arena (Chiang et al., 2024). Table 5 reports our Elo ranking. The leftmost column shows RAG-QA ARENA's ranking based on win ratio: LFRQA is ranked on the top followed by GPT-4O, GPT-4-TURBO, etc. Column (A) uses the same data in Table 3, but here we only have comparisons between LFRQA and other LLMs' responses, i.e. there are no direct comparisons between LLMs' responses. Thus, the total number of votes for are 176.7K for LFRQA, and 16.1K for other LLMs. The new ranking based on Elo rating aligns with RAG-QA ARENA, but the 95% confidence interval (CI) is not yet able to separate all different model pairs.

In Column (B), we add pairwise comparisons for all unique model pairs on 700 randomly sampled queries across seven domains (100 for each domain). This increases the votes for all LLMs compared to 23.1K. We continue to add 700 more queries in Column (C), which further increases the votes for all LLMs to 30.1K. We rank these pairs with the same LLM evaluator as in the main result table, and found that the ranking in (C) based on Elo rating aligns perfectly with RAG-QA ARENA,

and as we increase the pairs, the 95% CI can finally separate different models.

In general, the total added preference pairs are $\frac{N \times K \times (K-1)}{2}$, where K is the number of models in Table 3, and N=700 and 1400 for Column (B) and (C), respectively. With 43.6% increase of total pairs (and thus the compute), the final ranking is identical with RAG-QA ARENA based on win ratio, and only differs only slightly with the win+tie ranking in Table 3 by flipping the order of LLAMA-3-70B and COMMAND R+. These results present additional evidence that our approach of using LFRQA only for pairwise comparisons is reliable. Furthermore, it reduces the computational costs from $\mathcal{O}(K^2)$ to $\mathcal{O}(K)$ as we now only need to compare each LLM response once with the ground-truth in LFRQA.

Impact of the number of passages. In Appendix Table 10, we compare the top 3 LLMs by increasing the number of retrieved passages from 5 to 10. Doubling the number of passages (with extra costs) increases RAG-QA performances significantly. We also find that both GPT-4 models' improvements are greater than MIXTRAL-8X22B-INSTRUCT, showing their superior capability to understand long context and identify useful information from noise. The best win rate of GPT-4O against LFRQA is 41.3%, which is 13.7% points lower than LFRQA answers' win rate against GPT-4O. This result shows that LFRQA's answer quality is difficult to surpass, further justifying using it as an evaluation target.

Impact of model sizes. For the non-GPT LLM family, more parameters lead to better performances, but a larger increase in model sizes does not always indicate greater performance gains in

	Completeness	Citation Accuracy	Helpfulness
LFRQA	90.8	88.9	48.1
GPT-4	75.3	65.5	35.2

Table 6: Comparison between human and GPT-4 annotations. All numbers are in %.

our study. For example, the two QWEN1.5 models have the second-largest difference, but the lift from the smaller to the larger model is marginal. We leave more rigorous investigations to future research.

7 Discussion

7.1 LLM as Annotators

Using large language models to provide annotations has been explored in previous works (Tan et al., 2024). It could provide a more scalable solution than human annotations but can suffer from hallucination and accuracy issues that require human validations (Huang et al., 2023). We also experimented with LLM as annotators before we start human annotations. We subsample 100 queries from LFRQA and prompt GPT-4-0125-PREVIEW to follow the similar procedure in Sec. 3.1 to combine answers (Appendix Table 12). Then we request our data linguists to compare LFRQA and GPT-4 annotations based on 1) Completeness: whether all ROBUSTQA answers are integrated into the final answers; 2) Citation Accuracy: whether citations in answers pointing to the right documents; 3) Helpfulness: defined the same as in Sec. 4. Table 6 shows LFRQA out-performs GPT-4 annotations by 15.5%, 23.4% and 12.9% for the three dimensions, respectively, suggesting human annotations are both valuable and necessary for our task.

7.2 Alternative Evaluation Approaches

Using retrieved passages. RAG-QA ARENA leverages only LFRQA's annotations as ground-truth to directly evaluate LLM responses, and we explain this design choice in Sec. 4 that LFRQA consists of complete and coherent answers that can be viewed as high-quality summary of all available answers in the entire corpus. This enables us to not show retrieved passages as they 1) increase the input length and thus the latency of an evaluator; and 2) they could contain incorrect information due to retrieval error, which mislead evaluators.

Using LFRQA as references. We can also use LFRQA's annotations as references when constructing the prompt for pairwise evaluation. That is, we can potentially compare a pair of LLMs' responses by comparing them both against the references in a single trial. However, this approach would still require the similar $\mathcal{O}(K^2)$ pairs as in the Elo rating, which is not as efficient as our proposed RAG-QA ARENA framework.

For these reasons, we do not adopt the above two evaluation approaches. It is conceivable that prompt engineering, in-context example selections and even task specific evaluator training could further enhance alignments with human judges. We leave them for future research efforts.

8 Related Work

RAG-QA has been widely studied. Prior datasets are limited in the evaluation as their corpus relies heavily on Wikipedia and the answers are mostly short and extractive (Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Amouyal et al., 2019). ROBUSTQA and MULTIHOP-RAG (Tang and Yang, 2024) address the single domain issue, but still adopt short, extractive answers, which is not as suitable as LFRQA to evaluate modern LLMs that generate long-form answers.

Longform QA datasets have been proposed in prior work. ELI5 (Fan et al., 2019) and LONG-FACT (Wei et al., 2024) contain answers that are either not annotated directly on the corpus, and or not created by humans. Krishna et al. (2021) also points out that ELI5's small validation set has significant leakage from its train set. ASQA (Stelmakh et al., 2022) is the most similar data to our work, but its corpus is in the single Wikipedia domain. LFRQA is by far the RAG-QA dataset with the most comprehensive long-form answers.

Pairwise preference is now a standard way to evaluate LLMs. It allows direct comparison between two responses (Chiang et al., 2024; Lin et al., 2024). RAG-QA ARENA is unique by always including a high-quality human annotated **LFRQA** answer, thereby making the evaluation more trustworthy.

9 Conclusion

We create LFRQA, the first multi-domain dataset with coherent long-form answers to reliably benchmark RAG-QA. We propose a reliable LLM-based evaluation framework, RAG-QA ARENA, that enables direct comparisons between LLMs' answers and LFRQA, which we believe will facilitate the evaluation RAG-QA in the era of LLMs.

Limitations

We discuss some limitations of this work for future research efforts. RAG-QA ARENA can potentially cover more models. We didn't include some leading LLMs, such as Claude (Anthropic, 2024) and Gemini (Google, 2023) models, due to legal and resource constraints, but we plan to add them to the leaderboard in the future. Evaluation using GPT-4-0125-PREVIEW is not cheap. It costs on average 300 U.S. dollars per model on the full LFRQA's test set. We plan to subsample 10-20% of the queries for the final public leaderboard, which will be more cost-friendly for future users. Future research can also study training smaller but equally accurate models as evaluators. Finally, we mainly focus on different LLMs for RAG-QA in this work, but future research can study the impact of different retrievers or joint retrievers and LLM training using RAG-QA ARENA.

Ethics Statement

The authors of this paper are committed to conducting research ethically. We are leveraging existing LLMs to generate answers for LFRQA, which include many open-ended questions. LLM-generated answers could be incorrect or unfaithful, as retrievers could find irrelevant passages and LLM can hallucinate (Huang et al., 2023). These are known issues in the AI research community, and that is the reason we created LFRQA to better evaluate RAG-QA systems. The additional risks and potential harms are discussed in numerous previous works (Bender et al., 2021; Weidinger et al., 2021). The authors strive to ensure that the research and its results do not cause harm.

Data used in this work have been collected from public sources and used in accordance with all applicable laws and regulations. We use contracted data professionals for LFRQA annotations, and Appen platform[4] for human pairwise preference annotations. In both cases, we ensure our hourly rate is higher than 15 U.S. dollars per local minimum wage standard. The intended usage of LFRQA is compatible with the underlying data's access conditions (Appendix A.2)

References

Daniel G. Brown Aisha Khatun. 2024. Trutheval: A dataset to evaluate llm truthfulness and reliability.

---

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

ArXiv, arXiv:2406.01855.

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2019. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs. *ArXiv*, abs/1910.10683.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, and et al. 2023. Qwen technical report. arXiv:2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *ArXiv*, abs/2403.04132.

Jocab Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1).

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Baobao Chang Zhiyong Wu, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *ArXiv*, arXiv:2301.00234.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Aidan Gomez. 2024. Introducing command r+: A scalable llm built for business. [Online; posted 04-April-2024].

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. arXiv:2312.11805.

Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. RobustQA: Benchmarking the robustness of domain adaptation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4294–4311, Toronto, Canada. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

HuggingFace. 2024. Hugging face hub documentation. Https://huggingface.co/docs/hub/en/index.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: few-shot learning with retrieval augmented language models. 24(1).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, and et al. 2024. Mixtral of experts. arXiv:2401.04088.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Khyathi Chandu, Faeze Brahman, Yuntian Deng, Abhilasha Ravichander, Valentina Pyatkin, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking language models with challenging tasks from real users in the wild.

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

MetaAI. 2024. Introducing Meta Llama 3: The most capable openly available llm to date.

OpenAI. 2024. GPT-4 technical report. arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Owain Evans Stephanie Lin, Jacob Hilton. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *ArXiv*, arXiv:2109.07958.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *ArXiv*, arXiv:2402.13446.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multihop queries. *ArXiv*, https://arxiv.org/abs/2401.15391.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *ArXiv*, arXiv:2403.18802.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

A Appendix

A.1 Dev Set of LFRQA

We do not use dev set to fit the scope of the paper, but we provide an additional 10K queries for future model developments, including prompting engineering and training. Details are shown in Table 7. All data are collected using the same process and quality control described in Sec. 3.

A.2 Data License

LFRQA is created based on the following datasets. We make sure to follow the data distribution license for our usage of the data.

- **FiQA:** no license provided, but all data information can be found one the official website[5].
- **LoTTE:** MIT license[6].
- **BioQAS:** CC BY 2.5 license[7].
- **ROBUSTQA:** Apache-2.0 license[8].

A.3 Annotation Interface

Our annotation interface can be found in Fig. 5. Annotators use this UI to write long-form answers. Blue highlights are the original answers from ROBUSTQA.

A.4 Answer Generation Prompts

Table 11 shows our answer generation prompt. The last two lines "First, think step-by-step..." are what we refer to as the CoT prompt, which we remove for GPT-4O.

A.5 Evaluation Interface

We collect human pairwise preference data to benchmark our LLM-based evaluators. Fig. 6-7 show the details.

---

[5] https://sites.google.com/view/fiqa/home
[6] https://github.com/stanford-futuredata/ColBERT
[7] http://participants-area.bioasq.org/datasets/
[8] https://github.com/awslabs/robustqa-acl23

				Dev Set			ROBUSTQA		LFRQA	
Domain	**Source**	**Label**	$\|Q\|$	$\|D\|$	$\|P\|$	A/Q	W/A	A/Q	W/A
Lifestyle	LoTTE	[LI]	2,151	268,893	597,729	5.9	5.9	1.0	102.2
Recreation	LoTTE	[RE]	2,325	263,025	731,124	6.3	7.2	1.0	112.8
Technology	LoTTE	[TE]	2,223	1,000,000	1,707,346	5.4	8.9	1.0	83.3
Science	LoTTE	[SC]	2,137	343,642	854,756	4.8	6.4	1.0	100.6
Writing	LoTTE	[WR]	1,972	277,072	713,692	7.0	7.7	1.0	109.6

Table 7: Data summary for the dev set. Based on ROBUSTQA, there is no dev split for BioASQ and FiQA data.

A.6 Human Evaluation Instructions

Helpfulness: Information that is helpful/relevant to answer the query. An ideal answer consists of only information that is helpful/relevant to answer the query (Touvron et al., 2023; Bai et al., 2022).

Truthfulness: Information that is correct to answer the query. By our definition, truthful information should also be helpful information. If it is difficult to determine the truthfulness of some information, we consider it untruthful. Sometimes, this is due to not enough context provided in the answer. Another source of untruthfulness is when conflicting information is presented, and the answer does not coherently reconcile them (Stephanie Lin, 2021; Aisha Khatun, 2024).

Completeness: include as many helpful and truthful information.

Here are the details of our instructions.
1. If one answer has all truthful information while the other has some untruthful information, prefer the all-truthful one. **2.** If both have some untruthful information, prefer the one with less untruthful information. **3.** If both have all truthful information, prefer the one with more truthful or helpful information. **4.** If two answers look equally good, or it is too hard to differentiate, choose "Not sure."

As the annotation UI shows, the actual ratings are "Better," "Slightly Better," "Tie," "Slightly Worse" and "Worse". We merge "Better" and "Slightly Better," and "Slightly Worse" and "Worse" when computing correlation with model-based evaluators.

A.7 Evaluation Model Selection

Table 8 shows Pearson Corr. between human and LLM-based evaluators. We find using GPT-4-0125-PREVIEW alone achieves the best outcome, and thus, use it for RAG-QA ARENA evaluations.

A.8 No Answer Ratio

Table 9 shows the ratio of "I couldn't find an answer" in each LLM's answers.

Evaluator Models	LFRQA v.s. GPT-4	LFRQA v.s. MIXTRAL	LFRQA v.s. LLAMA-3
Llama-3-70b	0.52	0.51	0.48
Mixtral-8x22b	0.53	0.49	0.50
GPT-4-turbo	0.47	0.49	0.48
GPT-4o	0.43	0.46	0.45
GPT-4-0125-preview	**0.54**	**0.54**	**0.52**
Ensemble	0.53	0.53	**0.52**

Table 8: Correlation between human judges and leading LLMs as evaluators. Column titles are answer pairs. GPT-4: GPT-4-0125-preview; MIXTRAL: Mixtral-8x22b-Instruct-v0.1; LLAMA-3: Llama-3-70b-Instruct. The ensemble takes the majority vote out of the best three models.

A.9 Pairwise Evaluation Prompts

Table 13-15 show the details of the pairwise evaluation prompts including instruction, example prompt template, and in-context-learning examples.

A.10 Error Analysis

Fig. 8-9 show examples where the original GPT-4O puts an answer in its **<thinking>**, but generate "I couldn't find an answer." It is fixed by removing CoT prompt. We also compare them with MIXTRAL-8X22B, GPT-4-TURBO and LFRQA's answers.

	Overall	[BI]	[FI]	[LI]	[RE]	[TE]	[SC]	[WR]
GPT-4-TURBO	14.1	10.9	19.8	12.2	15.9	11.2	12.3	12.1
GPT-4O	48.3	32.4	55.2	50.3	52.8	43.4	49.4	48.4
GPT-4O *	16.9	10.4	26.8	13.9	19.8	11.2	14.7	13.9
GPT-4-0125-PREVIEW	15.8	11.0	21.9	13.5	19.1	12.6	13.5	14.1
MIXTRAL-8X22B	9.9	10.1	13.6	9.1	13.7	7.1	6.6	6.0
MIXTRAL-8X7B	18.5	15.4	24.6	17.8	23.7	13.3	15.3	15.0
LLAMA-3-70B	25.7	17.7	35.9	24.4	29.2	20.5	23.3	21.2
LLAMA-3-8B	12.6	10.0	15.9	11.5	16.2	8.4	12.9	11.0
COMMAND R+	14.4	10.8	21.7	12.5	16.1	10.1	14.7	10.5
COMMAND R	6.5	5.0	10.8	6.1	6.8	3.4	4.6	5.0
QWEN1.5-110B-CHAT	11.5	10.8	15.9	10.7	13.8	6.8	9.0	10.0
QWEN1.5-32B-CHAT	8.9	7.4	13.6	7.5	11.8	5.0	7.0	6.5

Table 9: No answer ratio for the entire LFRQA's test set. * means using answer generation prompt without CoT.

	Overall		[BI]		[FI]		[LI]		[RE]		[TE]		[SC]		[WR]	
	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T
GPT-4O* w/ **5 psgs**	36.9	41.0	52.9	59.3	38.4	42.3	25.1	27.9	40.4	46.4	35.6	38.8	42.8	47.6	28.4	31.1
GPT-4O* w/ **10 psgs**	**41.3**	<u>45.0</u>	**59.1**	**64.4**	45.7	49.4	<u>27.4</u>	<u>30.0</u>	**43.7**	**49.9**	<u>39.6</u>	<u>42.2</u>	46.0	<u>50.0</u>	30.7	<u>33.1</u>
GPT-4-TURBO w/ **5 psgs**	34.4	39.1	36.0	43.9	40.6	45.1	23.2	26.1	36.7	44.1	36.6	39.6	42.6	47.9	26.2	29.6
GPT-4-TURBO w/ **10 psgs**	<u>40.6</u>	**45.4**	40.8	49.5	**49.3**	**53.7**	**27.5**	**30.7**	<u>41.9</u>	<u>49.5</u>	**44.2**	**47.8**	**48.3**	**53.1**	**31.6**	**34.4**
MIXTRAL-8X22B w/ **5 psgs**	34.5	38.8	37.0	46.0	44.1	47.6	21.3	24.4	34.4	41.0	33.9	36.8	45.0	49.5	25.9	28.1
MIXTRAL-8X22B w/ **10 psgs**	38.1	42.5	<u>41.8</u>	<u>50.0</u>	<u>47.4</u>	<u>51.3</u>	24.9	28.0	39.1	46.0	38.9	41.1	46.7	51.5	28.1	30.7

Table 10: Impact of the number of passages. Evaluation results on the entire LFRQA test set based on top 5 or 10 passages. W and W+T indicate win and win+tie rate against LFRQA. **bold** and <u>underline</u> indicate the best and runner-up results. * means using the answer generation prompt w/o CoT.

Answer Generation Prompt

Based on the passages, provide a helpful answer to the query. Your answer must be faithful to the content in the passages. Do not use your own knowledge to answer the query. If you couldn't find any helpful information in the passages, respond "I couldn't find an answer."

Passages are inside <passage></passage> tags. Query is in the <query></query> tags.
{x.passages}

<query>
{x.question}
</query>

First, think step by step, and put your thinking in <thinking> tags. Your thinking must be shorter than 50 words.
Then, provide your answer.

Table 11: Prompt for answer generation. {x.*} indicates a component input that can be replaced by actual data. We modify this prompt slightly to be compatible with different LLMs' input formats, but the majority of the prompt, particularly instructions, remain the same. * The last section starting with "First, think step by step..." is what we refer to as CoT prompt. We remove it for GPT-4O only.

Query: Closing a futures position

Passage 1:
Futures exchanges are essentially auction houses facilitating a two-way auction. While they provide a venue for buyers and sellers to come together and transact (be that a physical venue such as a pit at the CME or an electronic network such as Globex), they don't actively seek out or find buyers and sellers to pair them together. The exchanges enable this process through an order book. As a futures trader you may submit one of two types of order to an exchange: Market Order - this is sent to the exchange and is filled immediately by being paired with a limit order. Limit Order - this is placed on the books of the exchange at the price you specify. If other participants enter opposing market orders at this price, then their market order will be paired with your limit order. In your example, trader B wishes to close his long position. To do this he may **enter a market sell order, which will immediately close his position at the lowest possible buy limit price**, or he may **enter a limit sell order, specifying the price at or above which he is willing to sell**. In the case of the limit order, **he will only sell and successfully close his position if his order becomes the lowest sell order** on the book. All this may be a lot easier to understand by looking at a visual image of an order book such as the one given in the explanation that I have published here: Stop Orders for Futures Finally, not that as far as the exchange is concerned, there is no difference between an order to open and an order to close a position. They're all just 'buy' or 'sell' orders. Whether they cause you to reduce/exit a position or increase/establish a position is relative to the position you currently hold; if you're flat a buy order establishes a new position, if you're short it closes your position and leaves you flat.

Passage 2:
For exchange contracts, yes. **A trader can close a position by taking an offsetting position**. CME's introduction to Futures explains it quite well (on page 22). Exiting the Market Jack entered the market on the buy side, speculating that the S&P 500 futures price would move higher. He has three choices for exiting the market:

Passage 3:
Assuming these are standardized and regulated contracts, the short answer is yes. In your example, Trader A is short while Trader B is long. If Trader B wants **to exit his long position, he merely enters a "sell to close" order** with his broker. Trader B never goes short as you state. **He was long while he held the contract, then he "sold to close"**. As to who finds the buyer of Trader B's contract, I believe that would be the exchange or a market maker. Therefore, Trader C ends up the counterparty to Trader A's short position after buying from Trader B. Assuming the contract is held until expiration, Trader A is responsible for delivering contracted product to Trader C for contracted price. In reality this is generally settled up in cash, and Trader A and Trader C never even know each other's identity.

Task 1: Write an answer.

Please write your new combined, long and well-formed answer below (required)

Figure 5: Annotation Interface

Annotation Generation Prompt

Provide a response around 100 words to the query in the <query></query> tags based on the passages. Passages are inside <passage></passage> tags. The response must incorporate all candidate answers in the <ans></ans>, and you are allowed to rephrase these answers in order to make your final response natural. The response should not include any information outside passages.
You should cite the passage number (indices) in the format of [1], [2], [3, 4], etc. at the end of each sentence.

{x.passages}

<query>
{x.question}
</query>

Table 12: Prompt for GPT-4 annotations. {x.*} indicates a component input that can be replaced by actual data.

Instruction

In this task, we will show you a query and a pair of answers to the query. You need to provide your preference over this pair of answers.

Try your best to determine whether the information in an answer can **help truthfully answer** the query. Rate your preference based on two criteria.

- **Helpfulness**: information that is helpful/relevant to answer the query.
  - An ideal answer consists of only information that is helpful/relevant to answer the query.
- **Truthfulness**: information that you believe is correct to answer the query. By our definition, truthful information should be helpful information by default.
  - If you find it difficult to determine the truthfulness of some information, consider it untruthful. Often time, this is due to *not enough context* provided in the answer.
  - Another source of untruthfulness is when *conflicting information* presented, and the answer does not reconcile them in a coherent way.

Here are a few cases to help you rate the preference (in the order of importance),

- **Case 1**: If one answer has all truthful information while the other has some untruthful information, prefer the all truthful one.
- **Case 2**: If both have some untruthful information, prefer the one with less untruthful information.
- **Case 3**: If both have all truthful information, prefer the one with more truthful or helpful information.
- **Case 4**: If two answers look equally good, or it is too hard to judge using the 3 cases above, choose "Not sure."

Please rate your preference of the pair of the answers to the query considering the above criteria. Here are the possible options.

- *Better*: Answer 1 is *noticeably better* than Answer 2.
- *Slightly better*: Answer 1 is *slightly better* than Answer 2.
- *Not sure*: It is difficult to determine which system is better.
- *Slightly worse*: Answer 1 is *slightly worse* than Answer 2.
- *Worse*: Answer 1 is *noticeably worse* than Answer 2.

Example 1

Query: should utf-16 be considered harmful?

Answer 1	The question of whether UTF-16 should be considered harmful is subject to differing opinions. One perspective suggests that UTF-16 is harmful due to a historical misconception about character encoding and recommends UTF-8 as the superior choice for various text interfaces. Another viewpoint argues that UTF-16 is not inherently harmful, emphasizing its utility in specific scenarios where it serves as a compromise between simplicity and compactness. The choice between UTF-16 and other encodings like UTF-8 depends on the specific requirements of the application, such as compatibility with ASCII or the need to efficiently encode certain character sets.
Answer 2	definitely not Yes harmful

Correct Answer: Better. Answer 1 is *noticeably better* than Answer 2. We can see Answer 2 simply lists three conflicting responses to the query without further explanation; whereas Answer 1 integrate them together with full explanation for people to understand the context. Therefore, Answer 1 is better.

Example 2

Query: should utf-16 be considered harmful?

Answer 1	The question of whether UTF-16 should be considered harmful is subject to differing opinions. One perspective suggests that UTF-16 is harmful due to a historical misconception about character encoding and recommends UTF-8 as the superior choice for various text interfaces. Another viewpoint argues that UTF-16 is not inherently harmful, emphasizing its utility in specific scenarios where it serves as a compromise between simplicity and compactness. The choice between UTF-16 and other encodings like UTF-8 depends on the specific requirements of the application, such as compatibility with ASCII or the need to efficiently encode certain character sets.
Answer 2	UTF-16 should not be considered harmful. However, contrasting views argue that UTF-16 should indeed be considered harmful. Some argue that the very reason UTF-16 exists is because some time ago there used to be a misguided belief that WideChar is going to be what UCS-4 now is. Additionally, the harmfulness of UTF-16 is tied to issues with exercising code.

Correct Answer: Not sure/Slightly better. Both Answer 1 and 2 provide the two conflicting views with explanation. One could argue that Answer 1 is more detailed for the "not harmful" aspect, so both Not sure and Slightly better is acceptable.

Figure 6: Pairwise preference: instruction.

Example 3

Query: what did European/American historical cooks do with the egg whites?

Answer 1	Historical European and American cooks used egg whites for making egg white omelettes and egg white pasta, as well as for stiffening clothing, similar to how starch is used today.
Answer 2	Clearing beer and wine (using egg white as finings As an adhesive in bookbinding and gilding, and as a size ancient nail polish stiffen clothing make egg white omelettes and egg white pasta

Correct Answer: Slightly Worse/Not sure. Both Answer 1 and 2 provide several usages of egg whites for European/American cooks. Answer 1 provides fewer options compared to 2, but is more contextualized, and thus more trustworthy. Answer 2, despite having more information, contains words/phrases that are hard to comprehend without further context. For example, "As an adhesive in bookbinding and gilding, and as a size", the second half of the phrase is meaningless and not helpful. So Slightly Worse is the most proper option here. Not sure is an acceptable option if you don't believe the information in both Answer 1 and 2.

Example 4

Query: what did European/American historical cooks do with the egg whites?

Answer 1	Historical European and American cooks used egg whites for making egg white omelettes and egg white pasta, as well as for stiffening clothing, similar to how starch is used today.
Answer 2	Egg whites have found their place in various non-culinary applications throughout history, such as in the clarification of beer and wine, in the conservation of books through bookbinding and gilding, and in makeup as an ancient form of nail polish. They were also utilized historically as a stiffening agent for clothing, similar to how starch is used today. The culinary landscape was not left untouched, with egg whites making their way into recipes for omelettes and pastas as early as the 15th century.

Correct Answer: Worse. Both Answer 1 and 2 provide several usages of egg whites for European/American cooks. Answer 2 clearly provides more options with fully explained details. Therefore, Answer 1 is *noticeably worse* than Answer 2.

Compare the pair of answers to the query below.

Query: what is the mode of administration of ubrogepant?

Answer 1
oral

Answer 2
Ubrogepant is administered orally.

How does Answer 1 compare to Answer 2? (required)

Better	Slightly Better	Not sure	Slightly Worse	Worse
○	○	○	○	○

Feedback

> If there are any data-related concerns or issues, please report them here

Figure 7: Pairwise preference: annotation interface.

Pairwise Evaluation: Instruction and Rubric

We will show you a query and a pair of answers to the query. You need to provide your preference over this pair of answers.

First, try your best to determine whether the information in an answer can help truthfully answer the query. Then rate your preference based on Helpfulness and Truthfulness.

- Helpfulness: information that is helpful/relevant to answer the query. An ideal answer consists of only information that is helpful/relevant to answer the query.
- Truthfulness: information that you believe is correct to answer the query. By our definition, truthful information should be helpful information. If you find it difficult to determine the truthfulness of some information, consider it untruthful. Often time, this is due to not enough context provided in the answer. Another source of untruthfulness is when conflicting information presented, and the answer does not reconcile them in a coherent way.

<rubric>
Here is how you judge (in the order of importance),
- If one answer has all truthful information while the other has some untruthful information, prefer the all truthful one.
- If both have some untruthful information, prefer the one with less untruthful information.
- If both have all truthful information, prefer the one with more truthful or helpful information.
- If two answers look equally good, or it is too hard to judge using the 3 cases above, then you are our "not sure" which one is better.
</rubric>

Table 13: Instruction and rubrics for pairwise evaluation. We use this template across all LLM evaluators. This can be considered as the "system" instruction for GPT-4 and LLama-3 models.

Pairwise Evaluation: Example Template

Query is in the <query></query> tags. Answer 1 is in <answer 1></answer 1>, and Answer 2 is in <answer 2></answer 2>.

<query>
x.question
</query>

<answer 1>
x.response1
</answer 1>

<answer 2>
x.response2
</answer 2>

Review the rubric in <rubric> tags,
- if you prefer <answer 1>, output 1.
- if you prefer <answer 2>, output 2.
- if you are not sure, output 0.

First, think step by step, put your thinking in <thinking></thinking> tags. Your thinking must be shorter than 50 words. Then, provide your rating inside <rating></rating> tags. Remember your rating should be 0 if you are not sure, and your rating must be either 0, 1, or 2.

Table 14: Example template for pairwise evaluation. This template is used for both ICL examples and the final test example.

Label	1
Query	difference between publicly and publically.
Answer 1	Both 'publicly' and 'publically' bear no difference in meaning, as they are essentially alternative spellings of the same concept. Publicly is more widely used, but the existence of 'publically' in reputable sources like the OED means it cannot be dismissed as simply incorrect. Some opinions hold that 'publicly' is the older irregular form, still preached by a lot of grammars, and 'publically,' on the other hand, is the newer and regular form.
Answer 2	There is no difference in meaning between 'publicly' and 'publically'; they are alternative spellings of the same word.
Reason	Both <answer 1> and <answer 2> are truthful. However, <answer 1> provides more truthful information as the context to compare the two terms. Therefore, <answer 1> is better.
Label	2
Query	what did European/American historical cooks do with the egg whites?
Answer 1	Historical European and American cooks used egg whites for making egg white omelettes and egg white pasta, as well as for stiffening clothing, similar to how starch is used today.
Answer 2	Egg whites have found their place in various non-culinary applications throughout history, such as in the clarification of beer and wine, in the conservation of books through bookbinding and gilding, and in makeup as an ancient form of nail polish. They were also utilized historically as a stiffening agent for clothing, similar to how starch is used today. The culinary landscape was not left untouched, with egg whites making their way into recipes for omelettes and pastas as early as the 15th century.
Reason	Both <answer 1> and <answer 2> provide several usages of egg whites for European/American cooks. <answer 2> clearly provides more options with fully explained details. Therefore, <answer 2> is better.
Label	0
Query	should utf-16 be considered harmful?
Answer 1	The question of whether UTF-16 should be considered harmful is subject to differing opinions. One perspective suggests that UTF-16 is harmful due to a historical misconception about character encoding and recommends UTF-8 as the superior choice for various text interfaces. Another viewpoint argues that UTF-16 is not inherently harmful, emphasizing its utility in specific scenarios where it serves as a compromise between simplicity and compactness. The choice between UTF-16 and other encodings like UTF-8 depends on the specific requirements of the application, such as compatibility with ASCII or the need to efficiently encode certain character sets.
Answer 2	UTF-16 should not be considered harmful. However, contrasting views argue that UTF-16 should indeed be considered harmful. Some argue that the very reason UTF-16 exists is because some time ago there used to be a misguided belief that WideChar is going to be what UCS-4 now is. Additionally, the harmfulness of UTF-16 is tied to issues with exercising code.
Reason	Both <answer 1> and <answer 2> reconcile the two conflicting views with detailed explanation. I am not sure which one is better.

Table 15: In-context examples for pairwise evaluation. Labels 1, 2, and 3 mean "answer 1 is better", "answer 2 is better" and "tie", respectively. "Reason" is a model's chain-of-thought output.

Passage 1	INTRODUCTION: Cidofovir (Vistide) is an antiviral marketed for the treatment of cytomegalovirus retinitis. Clinical efficacy has been reported with its broad antiviral spectrum that includes poxvirus, human papilloma virus and Herpes simplex. In immunodepressed patients, these infectious dermatoses are often recurrent and resistant. In an open study, we assessed the efficacy and clinical tolerance of cidofovir gel at 1 p. 100.PATIENTS AND METHODS: Twelve HIV-infected adults were included. Cidofovir gel at 1 p. 100 was applied directly on the lesions, once a day, for two weeks on the molluscum and condylomas, four weeks on the warts and one week on
Passage 2	(S)-1-(3-Hydroxy-2-phosphonylmethoxypropyl)cytosine (HPMPC, Cidofovir, Vistide) is an acyclic nucleoside phosphonate with broad-spectrum activity against a wide variety of DNA viruses including herpesviruses [Herpes simplex virus type 1 (HSV-1) and type 2 (HSV-2), varicella-zoster virus (VZV), cytomegalovirus (CMV), Epstein-Barr virus (EBV), human herpesvirus type 6 (HHV-6) and equine and bovine herpesviruses], papovaviruses [human polyoma virus and human papilloma virus (HPV)], adeno-, irido-, hepadna-, and poxviruses. HPMPC has proved effective against these viruses in different cell culture systems and/or animal models. The mechanism of action of HPMPC is based upon the interaction of its active intracellular metabolite, the diphosphorylated HPMPC derivative HPMPCpp, with
Passage 3	Cidofovir (CDF) or Vistid is a monophosphate nucleoside analogue that inhibits the DNA polymerase of herpes viruses including the cytomegalovirus (CMV). CDF is active on GCV-resistant strains with a mutation on the phosphotransferase gene (UL97). However, DNA polymerase gene mutations that induce resistance to GCV are responsible for cross-resistance to CDF. Resistance phenotypes to GCV and CDF were determined for 57 CMV strains isolated from blood and urine samples. Sixteen strains were recovered after CDF therapy. Of the remaining 41 CDF-naive strains, 34 were susceptible and seven resistant to GCV. Fifty percent inhibitory concentrations (IC50) for CDF were in the
Passage 4	We employed a murine model to test the concept of using an aerosolized, long-acting antiviral drug to protect humans against smallpox. We previously showed that a low dose of aerosolized cidofovir (HPMPC [Vistide]) was highly protective against subsequent aerosolized cowpox virus challenge and was more effective than a much larger dose of drug given by injection, suggesting that aerosolized cidofovir is retained in the lung. Because the nephrotoxicity of cidofovir is a major concern in therapy, delivering the drug directly to the respiratory tract might be an effective prophylactic strategy that maximizes the tissue concentration at the site of initial
Passage 5	(S)-1-[3-hydroxy-2-(phosphonomethoxy)propyl]cytosine [corrected] (HPMPC, cidofovir, CDV, Vistide) is an acyclic nucleoside analogue with a potent and selective activity against a broad spectrum of DNA viruses including the poxviruses. In this study we present the results of different treatment regimens in lambs experimentally infected with orf virus with different cidofovir formulations prepared in Beeler basis and Unguentum M. Our results show that choice of excipient, concentration of codofovir [corrected] and treatment regimen were all important to the clinical outcome of the therapy. Whilst one particular regimen appeared to exacerbate the lesion, treatment with 1% (w/v) cidofovir cream, prepared in Beeler basis, for
Query	**which virus is cidofovir (vistide) indicated for?**
GPT-4o	**<thinking>** Cidofovir (Vistide) is indicated for cytomegalovirus (CMV) retinitis, as mentioned in passage 1. **</thinking>**
	<answer> I couldn't find an answer. **</answer>**
GPT-4o*	Cidofovir (Vistide) is indicated for the treatment of cytomegalovirus (CMV) retinitis. It also has broad-spectrum activity against various DNA viruses, including herpesviruses, human papilloma virus (HPV), and poxviruses.
GPT-4-turbo	**<thinking>** Review passages for specific indications of Cidofovir (Vistide). Passage1 mentions it's marketed for cytomegalovirus retinitis. This directly answers the query. **</thinking>**
	<answer> Cidofovir (Vistide) is indicated for the treatment of cytomegalovirus retinitis. **</answer>**
Mixtral-8x22B-Instruct-v0.1	**<thinking>** Cidofovir (Vistide) is indicated for a wide variety of DNA viruses including herpesviruses, papovaviruses, adeno-, irido-, hepadna-, and poxviruses. **</thinking>**
	<answer> Cidofovir (Vistide) is indicated for a wide variety of DNA viruses including herpesviruses, papovaviruses, adeno-, irido-, hepadna-, and poxviruses. It is specifically mentioned to be effective against herpes simplex virus type 1 (HSV-1) and type 2 (HSV-2), varicella-zoster virus (VZV), cytomegalovirus (CMV), Epstein-Barr virus (EBV), human herpesvirus type 6 (HHV-6), human polyoma virus, human papilloma virus (HPV), and poxviruses. **</answer>**
LFRQA	Cidofovir is commonly used in the treatment of cytomegalovirus (CMV) infection and disease.

Passage 1	net capital losses inside a mutual fund are not distributed to shareholders but are retained within the fund to be written off against future capital gains. See also this previous answer or this one.
Passage 2	Generally speaking, each year, mutual funds distribute to their shareholders the dividends that are earned by the stocks that they hold and also the net capital gains that they make when they sell stocks that they hold. If they did not do so, the money would be income to the fund and the fund would have to pay taxes on the amount not distributed. (On the other hand, net capital losses are held by the fund and carried forward to later years to offset future capital gains). You pay taxes on the amounts of the distributions declared
Passage 3	the net gains are negative, i.e. losses, they are not passed on to the shareholders. See the last paragraph below). A shareholder doesn't have to reinvest the distribution amount into the mutual fund: the option of receiving the money as cash always exists, as does the option of investing the distribution into a different mutual fund in the same family, e.g. invest the distributions from Vanguard's S&P 500 Index Fund into Vanguard's Total Bond Index Fund (and/or vice versa). This last can be done without needing a brokerage account, but doing it across fund families will require the money
Passage 4	A mutual fund makes distributions of its dividends and capital gains, usually once a year, or seminanually or quarterly or monthly etc; it does not distribute any capital losses to its shareholders but holds them for offsetting capital gains in future years, (cf, this answer of mine to a different question). A stock pays dividends; a stock neither has nor does it distribute capital gains: you get capital gains (or losses) when you sell the shares of the stock, but these are not called distributions of any kind. Similarly, you incur capital gains or losses when you
Passage 5	points out, a shareholder in a mutual fund must pay taxes (in the appropriate categories) on the distributions from the fund even though no actual cash has been received because the entire distribution has been reinvested. It is worth keeping in mind that when the mutual fund declares a distribution (say $1.22 a share), the Net Asset Value per share drops by the same amount (assuming no change in the prices of the securities that the fund holds) and the new shares issued are at this lower price. That is, there is no change in the value of the
Query	**net loss not distributed by mutual funds to their shareholders?**
GPT-4o	**<thinking>** Net capital losses are not distributed to shareholders but are retained within the mutual fund to offset future capital gains. **</thinking>**
	<answer> I couldn't find an answer. **</answer>**
GPT-4o*	Net capital losses inside a mutual fund are not distributed to shareholders but are retained within the fund to be written off against future capital gains.
GPT-4-turbo	**<thinking>** The query asks about the distribution of net losses by mutual funds. Passages 1, 2, and 4 mention that net capital losses are not distributed but retained to offset future gains. **</thinking>**
	<answer> Net capital losses inside a mutual fund are not distributed to shareholders. Instead, they are retained within the fund to be written off against future capital gains. **</answer>**
Mixtral-8x22B-Instruct-v0.1	**<thinking>** The passages state that mutual funds do not distribute net capital losses to their shareholders but retain them within the fund to offset future capital gains. **</thinking>**
	<answer> Mutual funds do not distribute net capital losses to their shareholders. Instead, these losses are retained within the fund to be written off against future capital gains. **</answer>**
LFRQA	When you experience a loss within a mutual fund due to a decrease in share value, such a loss is not reported to you on Form 1099-DIV, and you are not able to claim this loss on your tax return. Gains and losses that occur within a mutual fund exist only internally and are not realized by you as a shareholder until assets are sold.

Figure 8: GPT-4O's original "I couldn't find an answer" response compared with GPT-4-TURBO and MIXTRAL-8X22B, and its answer without CoT (*). Blue highlights in the passages indicate helpful information to answer the query. **<answer>** tags are added to help differentiate from **<thinking>**.

Passage 1	Leptons and quarks are fermions. (Fermions are particles with half integer spins.) You can, like the author has, divide them into three generations on basis of their masses. The Higgs boson is a boson. (Bosons are particles with integer spins.) The Higgs boson (which happens to be electrically neutral) is part of a completely different category of particles, and it cannot be a part of the three generations you've mentioned. Also, I suppose the author was talking about fermions and not charged particles, because neutrinos are electrically neutral. Also remember that electric neutrality is not a criteria which determines
Passage 2	A fermion is just a particle of half-integer spin. Being a lepton for a particle is a matter of definition of global symmetries of the theory. This means that a lepton can in principle be both a fermion or a boson, although all known leptons are fermions (electron, muon, tau and their neutrinos). One example of bosonic lepton is the weak triplet Higgs boson of the type-2 see-saw models. In supersymmetric extensions of the Standard Model there are scalar partners of the ordinary fermions. All these scalar have spin-0 and some of them are leptons, yet another example of non-fermionic
Passage 3	Fundamental fermions like quarks and leptons are described by the spinor field, while gauge bosons like photons are described by the vector field. They together with the Higgs bosons are currently what we have in the Standard Model for elementary particles.
Passage 4	Usually the particles that are exchanged between matter particles are fundamental bosons. This is because they lack lepton number and baryon number. There are it turns out some bosons that have fermions in them(like the pions). The difference between a boson and a fermion comes mostly from their spins. A fermion has a fractional spin of anything than can be divided by 1/2 times Planck's constant. Or just 1/2 of spin. Bosons have integer spins. Of course fermions follow Pauli exclusion principle as you say. Even a Helium-4 nucleus is a boson however it is not a force carrier. As
Passage 5	Elementary particles are classified into two groups: Bosons & Fermions. Fermions comes with two families: quarks and leptons. Leptons come with three generations (till date no fourth generation leptons observed). Same is true for quarks as well.The first generation consists of electron e^{-} and electron-neutrino τ_{e}. Standard way of representing them goes as follows \begin{equation} L_{e} = \begin{bmatrix} \nu_{e}\\e^{-} \end{bmatrix}\qquad L_{\mu} = \begin{bmatrix} \nu_{\mu}\\ \mu^{-} \end{bmatrix}\qquad L_{\tau} = \begin{bmatrix} \nu_{\tau}\\ \tau^{-} \end{bmatrix} \end{equation} Where L_{e} stands for electron family (or first generation), L_{μ} stands for muon family (or second generation) and L_{τ} stands for tauon family (or third generation) of
Query	**are leptons fermions or bosons?**
GPT-4o	**<thinking>** Passages 1, 2, and 5 state that leptons are fermions. Passage 2 also mentions that all known leptons are fermions. **</thinking>**
	<answer> I couldn't find an answer. **</answer>**
GPT-4o*	Leptons are fermions. All known leptons, such as the electron, muon, tau, and their neutrinos, are fermions, which are particles with half-integer spins. (Passages 1, 2, and 5)
GPT-4-turbo	**<thinking>** Passage 1 and 2 discuss leptons primarily as fermions but also mention exceptions where they can be bosons in specific theoretical models. **</thinking>**
	<answer> Leptons are generally fermions, as all known leptons like electrons, muons, and tau particles are fermions. However, in certain theoretical models like the type-2 see-saw models and supersymmetric extensions of the Standard Model, leptons can also be bosons. **</answer>**
Mixtral-8x22B-Instruct-v0.1	**<thinking>** Leptons are mentioned as fermions in multiple passages. **</thinking>**
	<answer> Leptons are fermions, as they are particles with half-integer spins. This is mentioned in passages 1, 2, and 5. **</answer>**
LFRQA	A lepton can in principle be a fermion or a boson, although all known leptons are fermions (electron, muon, tau and their neutrinos). A fermion is any particle characterized by FermiDirac statistics and obeying the Pauli exclusion principle, for example, quarks.

Passage 1	Yes, right now means at "this exact moment". But "now" gives a longer lapse of time--perhaps in the next hour or so. "Right now" is also a panicked expression whereas "now" gives the feeling of being more relaxed.
Passage 2	and expectation of change. Right now it is too hot to go outside … but soon, in the evening it will be fine. I can't talk right now … I'll get back to you as soon as I can. (Note: later answers have correctly pointed out that "right now" can bring an element of imperative that "currently" does not). Grammar: There is a grammatical difference between the two. Currently is an adverb. It describes something taking place (verb). Right now is a noun - it is a time. So they can be used interchangeably in sentences, but not without slightly
Passage 3	They are almost the same, usually interchangable, but in certain contexts they might be interpreted slightly differently depending on the listener/reader. "Right now" could be interpreted as meaning "in this instant", or "at this very moment". Whereas "currently" is still referring to the present, but a present that is still ongoing, possibly for a longer period than "right now". For example: "No thank you, I don't want coffee right now" - meaning, I don't want a coffee now at 11.03am, but I might want tea at 11.57am. "No thank you, I don't currently want coffee" - this could be interpreted
Passage 4	"Currently" is continuous (a small window of time). "Right now" is discrete (at a precise moment in time).
Passage 5	Right often adds emphasis, as in 'I want you to do it right now.' In your example right now means 'at exactly the present moment', but it leaves the reader with the hope that a connection might be possible in the not too distant future.
Query	**what is the difference between now and right now?**
GPT-4o	**<thinking>** Passage 1 explains the difference between "now" and "right now," indicating that "right now" is more immediate and urgent, while "now" is more relaxed and can refer to a slightly longer time frame. **</thinking>**
	<answer> I couldn't find an answer. **</answer>**
GPT-4o*	"Right now" means at "this exact moment" and often conveys urgency or a panicked expression. "Now" refers to a longer lapse of time, such as the next hour or so, and gives a more relaxed feeling.
GPT-4-turbo	**<thinking>** The passages discuss the nuances between "now" and "right now." I need to synthesize these differences based on the given information without adding external knowledge. **</thinking>**
	<answer> The difference between "now" and "right now" as described in the passages is primarily in the immediacy and duration they imply. "Now" suggests a slightly longer lapse of time, perhaps within the next hour, and carries a more relaxed feeling. In contrast, "right now" refers to "this exact moment," indicating a more urgent and immediate timeframe. Additionally, "right now" can express a sense of panic or imperative, whereas "now" does not carry these connotations. **</answer>**
Mixtral-8x22B-Instruct-v0.1	**<thinking>** The passages suggest that "now" refers to a longer lapse of time, while "right now" refers to the exact moment. "Right now" can also convey a sense of urgency or panic. **</thinking>**
	<answer> "Now" refers to a longer lapse of time, perhaps in the next hour or so, and gives a more relaxed feeling. On the other hand, "right now" refers to the exact moment and can convey a sense of urgency or panic. **</answer>**
LFRQA	Including the word "right" often serves to stress the immediacy or importance of an action, suggesting it should be done in the immediate present. When "right now" is used, it usually implies that there's an existing condition hindering an action or connection, yet it anticipates that the situation will improve soon. On the other hand, if "just now" is leveraged, it could indicate a recent change that's not necessarily time-related, or it could carry the same sense of immediacy as "right now" but with a bit softer emphasis on the present moment. Comparatively, "right now" conveys urgency or a sense of crisis, whereas using "now" on its own offers a more relaxed timeframe, potentially extending to the next hour or so.

Figure 9: GPT-4o's original "I couldn't find an answer" response compared with GPT-4-Turbo and Mixtral-8x22b, and its answer without CoT (*). Blue highlights in the passages indicate helpful information to answer the query. **<answer>** tags are added to help differentiate from **<thinking>**.