Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions

Harsh Trivedi[†] **Niranjan Balasubramanian**[†] **Tushar Khot**[‡] **Ashish Sabharwal**[‡]

[†]Stony Brook University
Stony Brook, U.S.A.
{hjtrivedi,niranjan}@cs.stonybrook.edu

[‡]Allen Institute for AI
Seattle, U.S.A.
{tushark,ashishs}@allenai.org

Abstract

Prompting-based large language models (LLMs) are surprisingly powerful at generating natural language reasoning steps or Chains-of-Thoughts (CoT) for multi-step question answering (QA). They struggle, however, when the necessary knowledge is either unavailable to the LLM or not up-to-date within its parameters. While using the question to retrieve relevant text from an external knowledge source helps LLMs, we observe that this one-step retrieve-and-read approach is insufficient for multi-step QA. Here, *what to retrieve* depends on *what has already been derived*, which in turn may depend on *what was previously retrieved*. To address this, we propose IRCoT, a new approach for multi-step QA that interleaves retrieval with steps (sentences) in a CoT, guiding the retrieval with CoT and in turn using retrieved results to improve CoT. Using IRCoT with GPT3 substantially improves retrieval (up to 21 points) as well as downstream QA (up to 15 points) on four datasets: HotpotQA, 2WikiMultihopQA, MuSiQue, and IIRC. We observe similar substantial gains in out-of-distribution (OOD) settings as well as with much smaller models such as Flan-T5-large without additional training. IRCoT reduces model hallucination, resulting in factually more accurate CoT reasoning.[1].

1 Introduction

Large language models are capable of answering complex questions by generating step-by-step natural language reasoning steps—so called chains of thoughts (CoT)—when prompted appropriately (Wei et al., 2022). This approach has been successful when all information needed to answer the question is either provided as context (e.g., algebra questions) or assumed to be present in the model's parameters (e.g., commonsense reasoning).

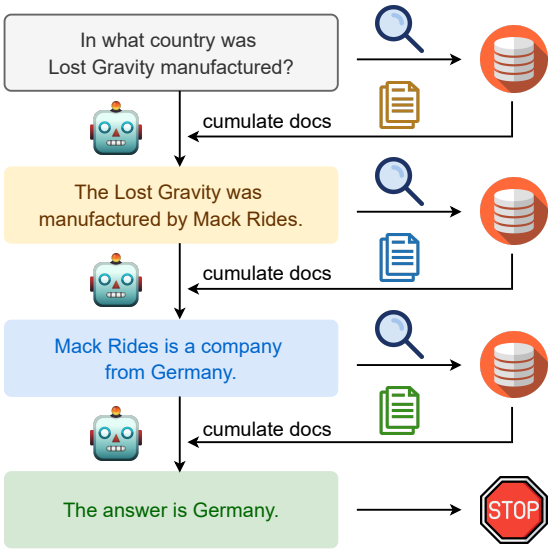

Figure 1: IRCoT interleaves chain-of-thought (CoT) generation and knowledge retrieval steps in order to guide the retrieval by CoT and vice-versa. This interleaving allows retrieving more relevant information for later reasoning steps, compared to standard retrieval using solely the question as the query.

However, for many open-domain questions, all required knowledge is not always available or up-to-date in models' parameters and it's beneficial to retrieve knowledge from external sources (Lazaridou et al., 2022; Kasai et al., 2022).

How can we augment chain-of-thought prompting for open-domain, knowledge-intensive tasks that require complex, multi-step reasoning?

While a *one-shot* retrieval from a knowledge source based solely on the question can successfully augment LMs with relevant knowledge for many factoid-based tasks (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022), this strategy has clear limitations for more complex multi-step reasoning questions. For such questions, one often must retrieve partial knowledge, perform partial reasoning, retrieve additional information based on the outcome of the partial

[1]Code, data, and prompts are available at https://github.com/stonybrooknlp/ircot

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10014–10037

reasoning done so far, and iterate. As an example, consider the question illustrated in Fig. 1, *"In what country was Lost Gravity manufactured?"*. The Wikipedia document retrieved using the question (in particular, the roller coaster Lost Gravity) as the query does not mention where Lost Gravity was manufactured. Instead, one must first infer that it was manufactured by a company called Mack Rides, and then perform further retrieval, guided by the inferred company name, to obtain evidence pointing to the manufacturing country.

Thus, the retrieval and reasoning steps must inform each other. Without retrieval, a model is likely to generate an incorrect reasoning step due to hallucination. Additionally, without generating the first reasoning step, the text supporting the second step can't be identified easily given the lack of lexical or even semantic overlap with the question. In other words, we need retrieved facts in order to generate factually correct reasoning steps and the reasoning steps to retrieve relevant facts.

Based on this intuition, we propose an *interleaving approach* to this problem, where the idea is to use retrieval to guide the chain-of-thought (CoT) reasoning steps and use CoT reasoning to guide the retrieval. Fig. 1 shows an overview of our retrieval method, which we call IRCoT.[2] We begin by retrieving a base set of paragraphs using the question as a query. Subsequently, we alternate between the following two steps: (i) *extend CoT*: use the question, the paragraphs collected thus far, and the CoT sentences generated thus far to generate the next CoT sentence; (ii) *expand retrieved information*: use the last CoT sentence as a query to retrieve additional paragraphs to add to the collected set. We repeat these steps till the CoT reports an answer or we reach the maximum allowed number of reasoning steps. Upon termination, all collected paragraphs are returned as the retrieval outcome. Finally, we use these as the context for answering the question via direct QA prompting (Brown et al., 2020) or CoT prompting (Wei et al., 2022).

We evaluate the efficacy of our system on 4 multi-step reasoning datasets under an open-domain setting: HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and IIRC (Ferguson et al., 2020). Our experiments using OpenAI GPT3 (`code-davinci-002`) (Brown et al., 2020; Ouyang et al., 2022; Chen et al., 2021) demonstrate that retrieval using IRCoT is substantially more effective than the baseline, one-step, question-based retrieval by 11-21 recall points under a fixed-budget optimal recall setup.[3] When IRCoT is used in conjunction with a prompting-based reader, it also leads to substantial improvement (up to 15 F1 points) in downstream few-shot QA performance and reduces factual errors in generated CoT by up to 50%. Our approach also works on much smaller Flan-T5 models (11B, 3B, and 0.7B) showing similar trends. In particular, we find QA using Flan-T5-XL (3B) with IRCoT even outperforms the 58X larger GPT3 with a one-step question-based retrieval. Furthermore, these improvements also hold up in an out-of-distribution (OOD) setting where the demonstrations from one dataset are used when testing on another dataset. Lastly, we note that our QA scores exceed those reported by recent works on few-shot prompting for open-domain QA (ODQA) (Khot et al., 2023; Press et al., 2022; Yao et al., 2022), although a fair apples-to-apples comparison with them isn't possible (cf. Appendix C).

In summary, our main **contribution** is a novel retrieval method, IRCoT, that leverages LMs' chain-of-thought generation capabilities to guide retrieval and uses retrieval in turn to improve CoT reasoning. We demonstrate that IRCoT:

1. improves both retrieval and few-shot QA performance on several multi-step open-domain QA datasets, in both IID and OOD settings;
2. reduces factual errors in generated CoTs; and
3. improves performance with both large-scale (175B models) as well as smaller-scale models (Flan-T5-*, $\leq$11B) without any training.

2 Related Work

Prompting for Open-Domain QA. LLMs can learn various tasks by simply using a few examples as prompts (Brown et al., 2020). They've also been shown to answer complex questions by producing step-by-step reasoning (chain-of-thoughts, or CoT) when prompted with a few or zero demonstrations (Wei et al., 2022; Kojima et al., 2022). Prompting has been applied to open-domain QA (Lazaridou et al., 2022; Sun et al., 2022; Yu et al., 2023) but its value in improving retrieval and QA for multi-step open-domain questions remains relatively underexplored.

[2] Interleaved Retrieval guided by Chain-of-Thought.

[3] We explain later (in the Metric section and Footnote 7) the appropriateness of this metric in our setting as opposed to more mainstream information recall metrics.

Recently three approaches have been proposed for multi-step open-domain QA. SelfAsk (Press et al., 2022) prompts LLMs to decompose a question into subquestions and answers subquestions by a call to Google Search API. DecomP (Khot et al., 2023) is a general framework that decomposes a task and delegates sub-tasks to appropriate submodels. They also decompose questions but delegate retrieval to a BM25-based retriever. Both of these approaches are not developed for CoT reasoning, do not focus on the retrieval problem, and require a single-hop QA model to answer the decomposed questions. Recently proposed ReAct (Yao et al., 2022) system frames the problem as generating a sequence of reasoning and action steps. These steps are much more complex, rely on much larger models (PaLM-540B), and require fine-tuning to outperform CoT for multi-step ODQA. Furthermore, none of these works have been shown to be effective for smaller models without any training. While a direct comparison with these approaches is not straightforward (difference in knowledge corpus, LLMs, examples), we find that our ODQA performance is much higher than all their reported numbers where available (§5).

Supervised Multi-Step Open-Domain QA. Prior work has explored iterative retrieval for open-domain QA in a fully supervised setting. Das et al. (2019) proposes an iterative retrieval model that retrieves using a neural query representation and then updates it based on a reading comprehension model's output. Feldman and El-Yaniv (2019) apply similar neural query reformulation idea for multihop open-domain QA. Xiong et al. (2021) extends the widely-used Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) to multihop setting, which has since been improved by Khattab et al. (2021). Asai et al. (2020) leverages the graph structure induced by the entity links present in Wikipedia paragraphs to perform iterative multi-step retrieval. GoldEn (Gold Entity) retriever (Qi et al., 2019) iteratively generates text queries based on paragraphs retrieved from an off-the-shelf retriever but requires training data for this next query generator. Nakano et al. (2021) used GPT3 to answer long-form questions by interacting with the browser but relied on human annotations of these interactions. All of these methods rely on supervised training on a large-scale dataset and can not be easily extended to a few-shot setting.

3 Chain-of-Thought-Guided Retrieval and Open-Domain QA

Our goal is to answer a knowledge-intensive multi-step reasoning question Q in a few-shot setting by using a knowledge source containing a large number of documents. To do this we follow a `retrieve-and-read` paradigm (Zhu et al., 2021), where the retriever first retrieves documents from the knowledge source and the QA model reads the retrieved documents and the question to generate the final answer. Our contribution is mainly in the `retrieve` step (§3.1), and we use standard prompting strategies for the `read` step (§3.2).

As noted earlier, for multi-step reasoning, retrieval can help guide the next reasoning step, which in turn can inform what to retrieve next. This motivates our interleaving strategy, discussed next.

3.1 Interleaving Retrieval with Chain-of-Thought Reasoning

Our proposed retriever method, IRCoT, can be instantiated from the following three ingredients: (i) a base retriever that can take a query and return a given number of paragraphs from a corpus or knowledge source; (ii) a language model with zero/few-shot Chain-of-Thought (CoT) generation capabilities; and (iii) a small number of annotated questions with reasoning steps explaining how to arrive at the answer in natural language (chain of thoughts) and a set of paragraphs from the knowledge source that collectively support the reasoning chain and the answer.

The overview of IRCoT is given in Fig. 2. We first gather a base set of paragraphs by retrieving K paragraphs using the question Q as the query. Then, we interleave two steps (`reason` and `retrieve`) iteratively until the termination criterion is met.

The **retrieval-guided reasoning step ("Reason")** generates the next CoT sentence using the question, the paragraphs collected thus far, and the CoT sentences generated thus far. The prompt template for the task looks as follows:

```
Wikipedia Title: <Page Title>
<Paragraph Text>
...
Wikipedia Title: <Page Title>
<Paragraph Text>

Q: <Question>
A: <CoT-Sent-1> ... <CoT-Sent-n>
```

For in-context demonstrations, we use the complete CoT in the above format. For a test instance,

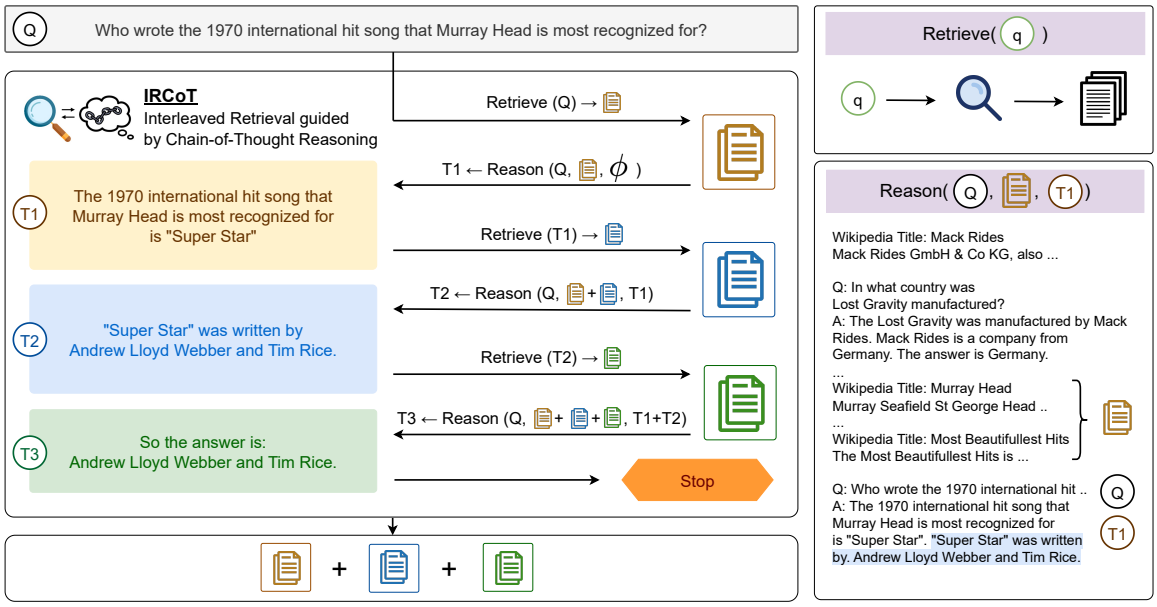

Figure 2: IRCoT interleaves chain-of-thought (CoT) generation and retrieval steps to guide the retrieval by CoT and vice-versa. We start by retrieving K documents using the question as they query and repeat two steps alternatingly until termination. (i) `reason`-step generates next CoT sentence based on the question, so far retrieved paragraphs, and CoT sentences. (ii) `retrieve`-step retrieves K more paragraphs based on the last CoT sentence. The process terminates when the generated CoT has "answer is" or the number of steps exceeds a threshold. The collection of all paragraphs is returned as the retrieval result on the termination.

we show the model only the CoT sentences generated thus far and let it complete the rest. Even though the model may output multiple sentences, for each `reason-step`, we only take the first generated sentence and discard the rest.

For the paragraphs in the in-context demonstrations, we use ground-truth supporting paragraphs and M randomly sampled paragraphs shuffled and concatenated together in the above format. For a test instance, we show all the paragraphs collected thus far across all the previous `retrieve-steps`.

If the generated CoT sentence has the "answer is:" string or the maximum number of steps[4] has been reached, we terminate the process and return all collected paragraphs as the retrieval result.

The **CoT-guided retrieval step ("Retrieve")** uses the last generated CoT sentence as a query to retrieve more paragraphs and adds them to the collected paragraphs. We cap the total number of collected paragraphs[5] so as to fit in at least a few demonstrations in the model's context limit.

3.2 Question Answering Reader

The QA reader answers the question using retrieved paragraphs taken from the retriever. We consider two versions of the QA reader implemented via two prompting strategies: `CoT Prompting` as proposed by Wei et al. (2022), Direct Prompting as proposed by Brown et al. (2020). For CoT prompting, we use the same template as shown in §3.2, but at test time we ask the model to generate the full CoT from scratch. The final sentence of CoT is expected to be of the form "answer is: ...", so that the answer can be extracted programmatically. If it's not in that form, the full generation is returned as the answer. For Direct Prompting, we use the same template as CoT Prompting but the answer field ("A: ") contains only the final answer instead of CoT. See App. G for details.

4 Experimental Setup

We evaluate our method on 4 multi-step QA datasets in the open-domain setting: **HotpotQA** (Yang et al., 2018), **2WikiMultihopQA** (Ho et al., 2020), answerable subset of **MuSiQue** (Trivedi et al., 2022), and answerable subset of **IIRC** (Ferguson et al., 2020). For HotpotQA, we use the Wikipedia corpus that comes with it for the open-domain setting. For each of the other three datasets, which originally come in a reading comprehension or mixed setting, we used the associated contexts to construct a

[4] set to 8 in our experiments.
[5] set to 15 in our experiments.

corpus for our open-domain setting (see App. A for details). For each dataset, we use 100 randomly sampled questions from the original development set for tuning hyperparameters, and 500 other randomly sampled questions as our test set.

4.1 Models

Retriever. We use BM25 (Robertson et al., 2009) implemented in Elasticsearch[6] as our base retriever. We compare two retriever systems:

(i) **One-step Retriever (OneR)** uses the question as a query to retrieve K paragraphs. We select $K \in \{5, 7, 9, 11, 13, 15\}$ that's best on the dev set.

(ii) **IRCoT Retriever** is our method described in §3. We use BM25 as its underlying retriever and experiment with OpenAI GPT3 (`code-davinci-002`) (Brown et al., 2020; Ouyang et al., 2022; Chen et al., 2021) and Flan-T5 (Chung et al., 2022) of different sizes as its CoT generator.

For demonstrating in-context examples to these LMs, we wrote CoTs for 20 questions for all the datasets (see App. §G). We then create 3 demonstration ("training") sets by sampling 15 questions each for each dataset. For each experiment, we search for the best hyperparameters for the dev set using the first demonstration set and evaluate each demonstration set on the test set using the selected hyperparameters. We report the mean and standard deviation of these 3 results for each experiment.

At test time, we pack as many demonstrations as possible within the model's context length limit. The context limit for GPT3 (`code-davinci-002`) is 8K word pieces. Flan-T5-* doesn't have any hard limit as it uses relative position embeddings. But we limit Flan-T5's context to 6K word pieces, which is the maximum we could fit in the memory of our 80G A100 GPUs.

IRCoT Retriever has one key hyperparameter: $K \in \{2, 4, 6, 8\}$, the number of paragraphs to retrieve at each step. Additionally, when creating "training" demonstrations for IRCoT's Reasoner module, we use gold paragraphs and a smaller number $M \in \{1, 2, 3\}$ of distractor paragraphs (§3.1).

Retrieval Metric: We allow a maximum of 15 paragraphs for all retriever systems and measure the recall of the gold paragraphs among the retrieved set of paragraphs. We search for the hyperparameter K (and M for IRCoT) that maximizes the recall on the dev set and use it on the test set. The reported metric can thus be viewed as the *fixed-budget optimal recall* for each system considered.[7]

QA Reader. To implement the reader, we use the same LMs as used in the `reason-step` of IRCoT Retriever. We found that QA readers implemented with Flan-T5-* perform better with the Direct Prompting strategy and GPT3 performs better with CoT Prompting strategy (see App. E). Hence we use Direct prompting strategy for QA with Flan-T5-* and CoT with GPT3 for the experiments.[8]

The QA reader has one hyperparameter M: the number of distractor paragraphs in the in-context demonstrations. We search for M in $\{1, 2, 3\}$. When used in conjunction with IRCoT retriever M is tied for the CoT generator and the reader.

Open-Domain QA (ODQA) Models. Putting retrievers and readers together, we experiment with ODQA models constructed from the various language models denoted as **OneR QA** and **IRCoT QA**. For IRCoT QA, the choice of LM for the CoT generator and the reader is kept the same. We also experiment with retriever-less QA readers **NoR QA** to assess how well LMs can answer the question from their parametric knowledge alone. To select the best hyperparameters for the ODQA model, we search for the hyperparameters K and M that maximize the answer F1 on the development set.

IIRC is structured slightly differently from the other datasets, in that its questions are grounded in a main passage and other supporting paragraphs come from the Wikipedia pages of entities mentioned in this passage. We slightly modify the retrievers and readers to account for this (see App. B).

5 Results

IRCoT retrieval is better than one-step. Fig. 3 compares OneR with IRCoT retrievers made from

[6] https://www.elastic.co/

[7] Note that our retrieved documents are not ranked, making standard information retrieval metrics such as MAP and DCG inapplicable. Further, we can only limit the number of retrieved paragraphs *per step* to K. Since the total number of reasoning steps varies for questions, and in some cases, we don't even obtain all K paragraphs in a given step, the total number of retrieved paragraphs also varies (even though capped at 15). This makes Recall@k, Precision@k, etc., also not applicable as metrics for any given k.

[8] IRCoT, by construction, produces a CoT as a part of its retrieval process. Thus, instead of having a separate post-hoc reader, one can also just extract the answer from the CoT generated during retrieval. However, we found this to be a suboptimal choice, so we always use a separate reader (see App. F).

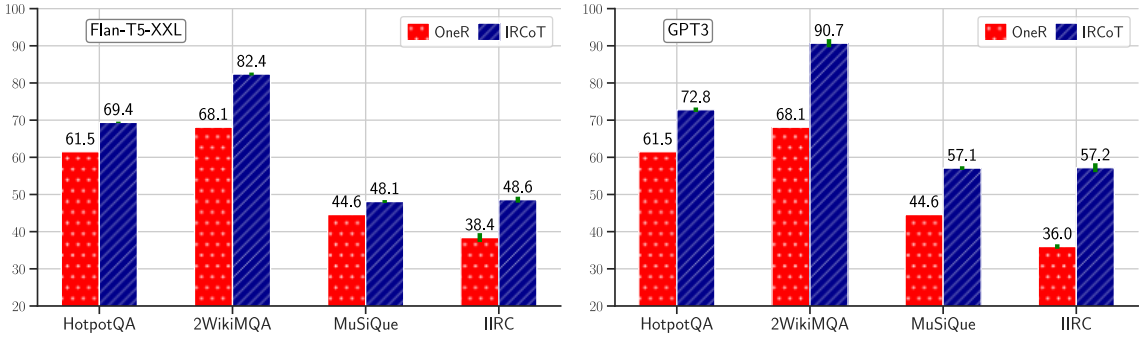

Figure 3: Retrieval recall for one-step retriever (OneR) and IRCoT instantiated from `Flan-T5-XXL` (left) and `GPT3` (right) models. IRCoT outperforms OneR for both models and all datasets.

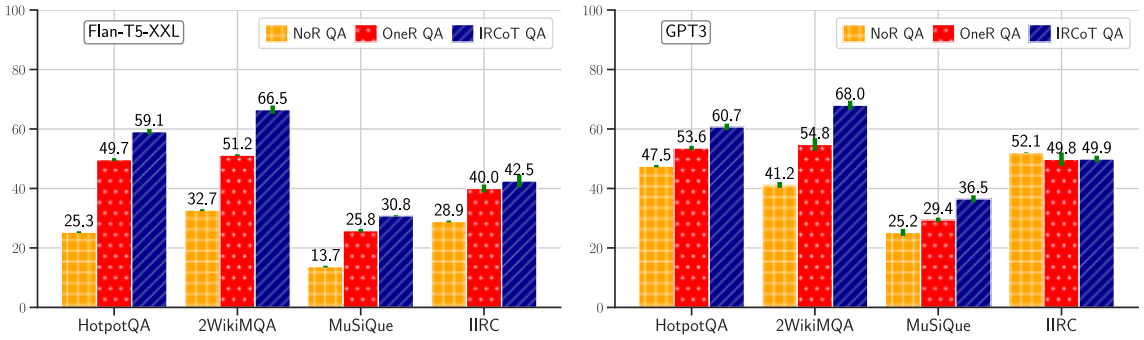

Figure 4: Answer F1 for ODQA model made using (i) no retriever (NoR QA) (ii) one-step retriever (OneR QA) and (iii) IRCoT QA instantiated from `Flan-T5-XXL` (left) and `GPT3` (right) models. IRCoT QA outperforms OneR QA and NoR QA for both models on all datasets, except for GPT3 on IIRC.

`Flan-T5-XXL` and `GPT3` LMs. For both models, IRCoT significantly outperforms one-step retrieval across all datasets. For `Flan-T5-XXL`, IRCoT improves our recall metric relative to one-step retrieval, on HotpotQA by 7.9, on 2WikiMultihopQA by 14.3, on MuSiQue by 3.5, and on IIRC by 10.2 points. For `GPT3`, this improvement is by 11.3, 22.6, 12.5, and 21.2 points, respectively.

IRCoT QA outperforms NoR and OneR QA. Fig. 4 compares ODQA performance using NoR, OneR and IRCoT retriever made from `Flan-T5-XXL` and `GPT3` LMs. For `Flan-T5-XXL`, IRCoT QA outperforms OneR QA on HotpotQA by 9.4, on 2WikiMultihopQA by 15.3, on MuSiQue by 5.0 and IIRC by 2.5 F1 points. For `GPT3`, the corresponding numbers (except for IIRC) are 7.1, 13.2, and 7.1 F1 points. For `GPT3`, IRCoT doesn't improve the QA score on IIRC, despite significantly improved retrieval (21 points as shown in Fig. 3). This is likely because IIRC relevant knowledge may already be present in GPT3, as also evidenced by its NoR QA score being similar. For other datasets and model combinations, NoR QA is much worse than IRCoT QA, indicating the limits of the models' parametric knowledge.

IRCoT is effective in OOD setting. Since CoT may not always be easy to write for new datasets, we evaluate NoR, OneR, and IRCoT on generalization to new datasets, i.e. OOD setting. To do so, we use prompt demonstrations from one dataset to evaluate on another dataset.[9] For all pairs of the datasets[10] and for both `Flan-T5-XXL` and `GPT3`, we find the same trend as in the IID setting: IRCoT retrieval outperforms OneR (Fig. 5), and IRCoT QA outperforms both OneR QA and NoR QA (Fig. 6).

IRCoT generates CoT with fewer factual errors. To assess whether our approach also improves the factuality of generated CoTs, we manually annotated CoTs generated by NoR QA, OneR QA, and IRCoT QA using GPT3 for 40 randomly sampled questions from each of the four datasets. We considered CoT to have a factual error if at least one

[9]We use the evaluation dataset's corpus for retrieval.

[10]We skip IIRC in this exploration as the task is structured a bit differently and requires special handling (see App. B).

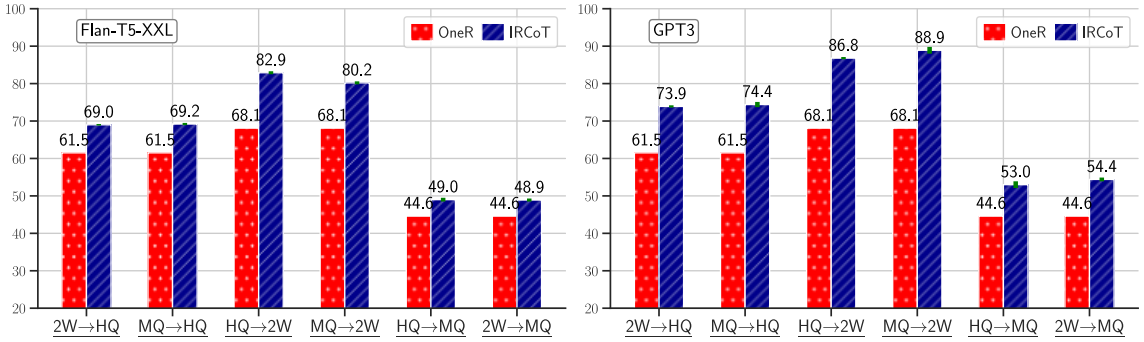

Figure 5: Retrieval recall for OneR and IRCoT using Flan-T5-XXL (Left) and GPT3 (Right) in out-of-distribution (OOD) setting. HQ (HotpotQA), 2W (2WikiMultihopQA), MQ (MuSiQue). The result X→Y indicates prompt demonstrations are from dataset X and evaluation is on dataset Y. IRCoT outperforms OneR in such an OOD setting.

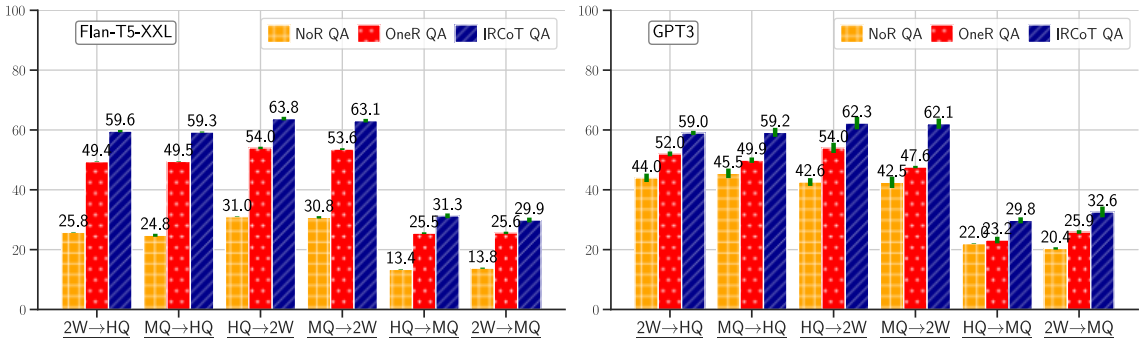

Figure 6: Answer F1 for NoR QA, OneR QA and IRCoT QA using Flan-T5-XXL (Left) and GPT3 (Right) in out-of-distribution (OOD) setting. HQ (HotpotQA), 2W (2WikiMultihopQA), MQ (MuSiQue). The result X→Y indicates prompt demonstrations are from dataset X and evaluation is on dataset Y. IRCoT QA outperforms OneR QA and NoR QA in such OOD setting.

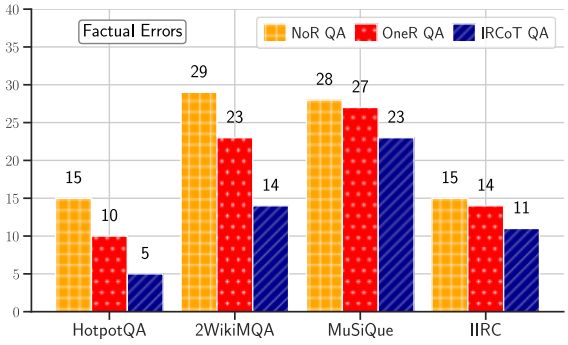

Figure 7: Number of questions, out of 40, where CoT generated by GPT3 using different methods has at least 1 factual error. Factual errors: IRCoT < OneR < NoR.

of the facts[11] is not true.[12] As Fig. 7 shows, NoR makes the most factual errors, OneR makes fewer, and IRCoT the least. In particular, IRCoT reduces the factual errors over OneR by 50% on HotpotQA and 40% on 2WikiMultihopQA.

Table 2 illustrates how the CoT predictions for different methods vary qualitatively. Since NoR relies completely on parametric knowledge, it often makes a factual error in the first sentence, which derails the full CoT. OneR can retrieve relevant information closest to the question and is less likely to make such errors early on, but it still makes errors later in the CoT. IRCoT, on the other hand, is often able to prevent such errors in each step.

IRCoT is also effective for smaller models. To see how effective IRCoT is at different LM sizes, we show the scaling plots in Fig. 8.[13] We compare the recall for OneR and IRCoT using `Flan-T5` {base (0.2B), large (0.7B), XL (3B), XXL (11B)}, and GPT3 `code-davinci-002` (175B). IRCoT with even the smallest model (0.2B) is better than

[11] all sentences before the final "answer is:" sentence.

[12] Note that factual error doesn't necessarily mean the predicted answer is incorrect and vice-versa. This is because the model can generate a wrong answer despite all correct facts, and vice-versa. We also account for the possibility of answer annotation errors in the original datasets.

[13] We skip IIRC here as the smaller models are not good at identifying Wikipedia titles from a paragraph and a question which is necessary for IIRC (see App. B).

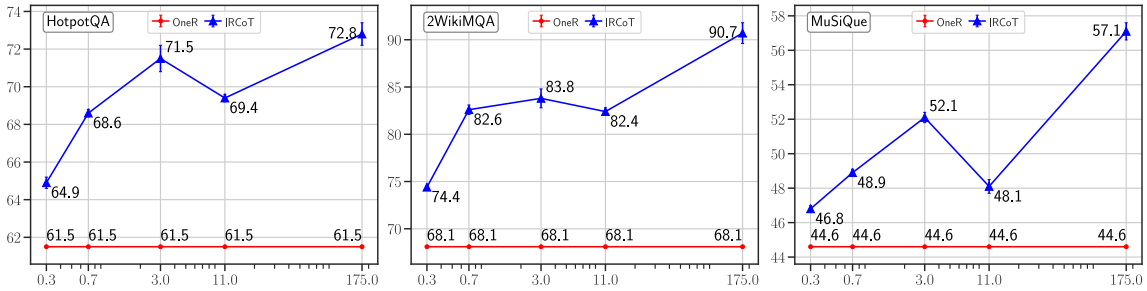

Figure 8: Retrieval recall for OneR (bottom) and IRCoT (top) for LMs of increasing sizes: Flan-T5 {base (0.2B), large (0.7B), XL (3B), XXL (11B)} and GPT3 (175B) on HotpotQA, 2WikiMultihopQA, MuSiQue. IRCoT outperforms OneR for all model sizes, including the 0.3B model, and the difference roughly grows with model size. Note: OneR doesn't use LM in its retrieval and so has a fixed score.

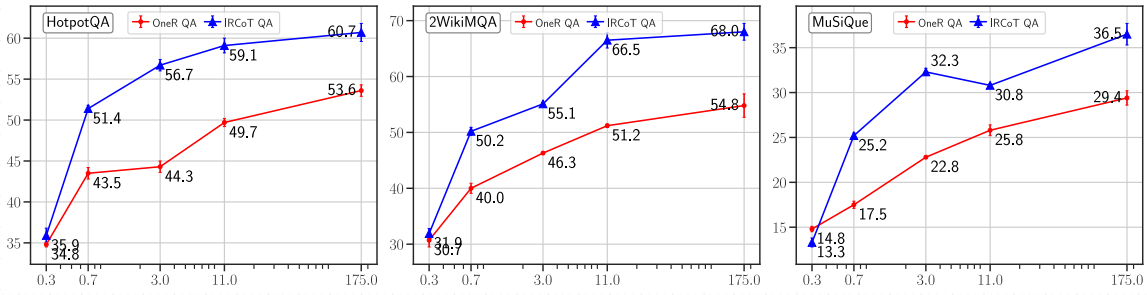

Figure 9: Answer F1 for ODQA models made using OneR (bottom) and IRCoT (top) for LMs of increasing sizes: Flan-T5 {base (0.2B), large (0.7B), XL (3B), XXL (11B)} and GPT3 (175B) on HotpotQA, 2WikiMultihopQA and MuSiQue. IRCoT QA outperforms OneR QA for all model sizes except for the smallest, 0.3B. IRCoT with 3B model even outperforms OneR with 58X larger GPT3 model showing the value of improved retrieval.

OneR, and the performance roughly improves with the model size. This shows the CoT generation capabilities of even small models can be leveraged for improving retrieval. Furthermore, we show the effect of model size on the QA score in Fig. 9. For all sizes except the smallest (0.2B), we see IRCoT QA is better than OneR QA. Moreover, IRCoT with a 3B model even outperforms OneR and NoR with a 58X larger 175B GPT3 model in all datasets.

IRCoT is SOTA for few-shot multistep ODQA.[14] We compare IRCoT QA with five recent approaches to using LLMs for ODQA: Internet-Augmented QA (Lazaridou et al., 2022), RECITE (Sun et al., 2022) ReAct (Yao et al., 2022), SelfAsk (Press et al., 2022), and DecomP (Khot et al., 2022). Although these are not head-to-head comparisons as different methods use different APIs, knowledge sources, and even LLMs (see App. C for details), it is still informative to explore, in a leaderboard-style fashion, how IRCoT performs relative to the best numbers published for these recent systems.

[14] App. §C reports updated SOTA numbers, including contemporaneous and newer works.

Model	$HpQA^{Br}$	HpQA	2WikiMQA	MQ^{2H}
InterAug	– | –	30.3 | –	– | –	– | –
RECITE	– | –	37.1 | 48.4	– | –	– | –
ReAct	– | –	35.1 | –	– | –	– | –
SelfAsk	– | –	– | –	40.1 | –	15.2 | –
DecomP	– | 50.0	– | –	– | 59.3	– | –
IRCoT QA	**45.8 | 58.5**	**49.3 | 60.7**	**57.7 | 68.0**	**34.2 | 43.8**

Table 1: Comparison with other LLM-based ODQA systems on EM and F1 scores. '–': score is unavailable. $HpQA^{Br}$: Bridge questions subset of HotpotQA. MQ^{2H}: MuSiQue 2-hop questions. IRCoT QA with GPT3 (ours) outperforms other systems by a large margin. Note: Comparisons aren't head-to-head as discussed in the text. App. §C reports updated SOTA numbers, including contemporaneous and newer works.

As shown in Table 1, IRCoT QA significantly outperforms all of these recent systems by a large margin, setting a new state of the art in terms of what's achievable via retrieval-augmented LLMs (without supervised training).

6 Conclusions

Chain-of-thought prompting has significantly improved LLMs' ability to perform multi-step reason-

	Question	Gold Facts	NoR QA	OneR QA	IRCoT QA
HotpotQA	The tap dancing soloist in My Feet Keep Dancing also starred in what 1935 MGM musical?	⇒ The tap dancing soloist in My Feet Keep Dancing is Fayard Nicholas. ⇒ Fayard Nicholas starred in the 1935 MGM musical An All-Colored Vaudeville Show.	The tap dancing soloist in My Feet Keep Dancing is Eleanor Powell. Eleanor Powell also starred in the 1935 MGM musical Broadway Melody of 1936. So the answer is: Broadway Melody of 1936.	The tap dancing soloist in My Feet Keep Dancing is Fayard Nicholas. Fayard Nicholas also starred in the 1935 MGM musical Top Hat. So the answer is: Top Hat.	The tap dancing soloist in My Feet Keep Dancing is Fayard Nicholas. Fayard Nicholas starred in the 1935 MGM musical An All-Colored Vaudeville Show. So the answer is: An All-Colored Vaudeville Show.
2WikiMQA	Who is the father-in-law of Queen Hyojeong?	⇒ Queen Hyojeong is the wife of King Heonjong of Joseon. ⇒ King Heonjong of Joseon is the son of Crown Prince Hyomyeong.	Queen Hyojeong has a husband named King Gojong. King Gojong has a father named Heungseon Daewongun. Thus, Queen Hyojeong has a father-in-law named Heungseon Daewongun. So the answer is: Heungseon Daewongun.	Queen Hyojeong is the wife of King Heonjong of Joseon. King Heonjong of Joseon is the son of King Sejo of Joseon. Thus, King Sejo of Joseon is the father-in-law of Queen Hyojeong. So the answer is: King Sejo of Joseon.	Queen Hyojeong is the wife of King Heonjong of Joseon. King Heonjong of Joseon is the son of Crown Prince Hyomyeong. Thus, Crown Prince Hyomyeong is the father-in-law of Queen Hyojeong. So the answer is: Crown Prince Hyomyeong.
MuSiQue	What is the name of the castle in the city where the performer of A Collection 1984–1989 was born?	⇒ A Collection 1984–1989 was performed by Jane Siberry. ⇒ Jane Siberry was born in Toronto. ⇒ The castle in Toronto is the Casa Loma.	The performer of A Collection 1984–1989 is The The. The The was born in London. The name of the castle in London is the Tower of London. So the answer is: the Tower of London.	A Collection 1984–1989 was performed by Jane Siberry. Jane Siberry was born in Toronto. The castle in Toronto is Peqin Castle. So the answer is: Peqin Castle.	A Collection 1984–1989 was performed by Jane Siberry. Jane Siberry was born in Toronto. The castle in Toronto is the Casa Loma. So the answer is: Casa Loma.

Table 2: Example CoTs generated by GPT3 with different methods. Since NoR relies on parametric knowledge, it often makes a factual error in the first sentence derailing the full CoT. OneR can retrieve relevant information closest to the question and is less likely to make such errors early on, but it still makes errors later in the CoT. As IRCoT performs retrieval after each step, it is often able to prevent such errors in each step. More examples are in App. D.

ing. We leveraged this ability to improve retrieval, and in turn, improve QA performance for complex knowledge-intensive open-domain tasks in a few-shot setting. We argued that one-step question-based retrieval is insufficient for such tasks, and introduced IRCoT, which uses interleaved CoT reasoning and retrieval steps that guide each other step-by-step. On four datasets, IRCoT significantly improves both retrieval and QA performance when compared to one-step retrieval, for both large and relatively smaller-scale LMs. Additionally, CoTs generated by IRCoT contain fewer factual errors.

Limitations

IRCoT relies on the base LM to have a zero or few-shot CoT-generation ability. While this is commonly available in large LMs (over 100B), it's not as common for small LMs (under 20B), which to some extent limits IRCoT adoptability. Given the recent surge of interest (Tay et al., 2023; Magister et al., 2022; Ho et al., 2022), however, smaller LMs will likely increasingly acquire such ability, making IRCoT compatible with many more LMs.

IRCoT also relies on the base LM to support long inputs as multiple retrieved paragraphs need to fit in the LM's input, in addition to at least a few demonstrations of QA or CoT with paragraphs. This was supported by the models we used as `code-davinci-002` (GPT3) allows 8K tokens and Flan-T5-* uses relative position embeddings making it as extensible as the GPU memory constraints allow. Future work can explore strategies to rerank and select the retrieved paragraphs instead of passing all of them to the LM to alleviate the need for the LM to support long input.

The performance gain of IRCoT retriever and QA (over OneR and ZeroR baselines) come with an additional computational cost. This is because IRCoT makes a separate call to an (L)LM for each sentence of CoT. Future work can focus on, for instance, dynamically deciding when to retrieve more information and when to perform additional reasoning with the current information.

Lastly, a portion of our experiments was carried out using a commercial LLM API from OpenAI (`code-davinci-002`). This model was deprecated by OpenAI after our submission making the reproduction of these experiments challenging despite our best efforts, just like any other work using such APIs. The trends discussed in the paper (IRCoT > OneR > NoR), we believe, would still hold. Additionally, all our experiments using Flan-T5-*, which exhibit similar trends as that of GPT3, will remain reproducible, thanks to its publicly available model weights.

Ethical Considerations

Language models are known to hallucinate incorrect and potentially biased information. This is especially problematic when the questions asked to it are of a sensitive nature. While retrieval-augmented approaches such as ours are expected to alleviate this issue to some extent by grounding generation in external text, this by no means solves the problem of generating biased or offensive statements. Appropriate care should thus be taken if deploying such systems in user-facing applications.

All the datasets and models used in this work are publicly available with permissible licenses. HotpotQA has CC BY-SA 4.0 license[15], 2WikiMultihopQA has Apache-2.0 license[16], MuSiQUe and IIRC have CC BY 4.0 license[17], and Flan-T5-* models have Apache-2.0 license.

Acknowledgments

We thank the reviewers for their valuable feedback and suggestions. We also thank OpenAI for providing access to the `code-davinci-002` API. This material is based on research supported in part by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003, in part by the National Science Foundation under the award IIS #2007290, and in part by an award from the Stony Brook Trustees Faculty Awards Program.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309, Florence, Italy. Association for Computational Linguistics.

James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. 2020. IIRC: A dataset of incomplete information reading comprehension questions. In *EMNLP*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

[15] https://creativecommons.org/licenses/by-sa/4.0/

[16] https://www.apache.org/licenses/LICENSE-2.0

[17] https://creativecommons.org/licenses/by/4.0

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Xanh Ho, A. Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *COLING*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

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.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. RealTime QA: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Baleen: Robust multi-hop reasoning at scale via condensed retrieval. In *Advances in Neural Information Processing Systems*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. 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 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *TACL*, 10:539–554.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. 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*.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Constructing Retrieval Corpora

HotpotQA already comes with the associated Wikipedia corpus for the open-domain setting, so we use it directly. 2WikiMultihopQA and MuSiQue, however, are originally reading comprehension datasets. Questions in 2WikiMultihopQA and MuSiQue are associated with 10 and 20 paragraphs respectively, 2-4 of which are supporting and others are non-supporting. To turn these datasets into an open-domain setting, we make two corpora, one for each dataset, by combining all supporting and non-supporting paragraphs for all its questions in the train, development, and test sets. IIRC is originally a mix between reading comprehension and an open-domain setting. Each question is grounded in one main paragraph, which contains links to multiple Wikipedia pages with several paragraphs each. We create a corpus out of all the paragraphs from all the Wikipedia pages present in the dataset.[18] We do assume the availability of the main passage which doesn't need to be retrieved and is always present. We don't assume the availability of Wikipedia links in the main passage, however, to keep the retrieval problem challenging.[19]

B Special Handling of Models for IIRC

IIRC is slightly different from the other datasets, in that the question is grounded in the main passage and other supporting paragraphs come from the Wikipedia pages of entities mentioned in this passage. We modify the retrievers and readers to account for this difference: (i) We always keep the main passage as part of the input to the model regardless of the retrieval strategy used. (ii) For all the retrieval methods, we first prompt the model to generate a list of Wikipedia page titles using the main passage and the question. We map these generated titles to the nearest Wikipedia page titles in the corpus (found using BM25), and then the rest of the paragraph retrieval queries are scoped within only those Wikipedia pages.

To prompt the model to generate Wikipedia page titles using the main passage and the question for IIRC, we use the following template.

```
Wikipedia Title: <Main Page Title>
<Main Paragraph Text>

Q: The question is: '<Question>'. Generate titles
of <N> Wikipedia pages that have relevant
information to answer this question.
A: ["<Title-1>", "<Title-2>", ...]
```

For "training", i.e., for demonstrations, N (≤ 3) is the number of supporting Wikipedia page titles for the question. At test time, since the number of supporting page titles is unknown, we use a fixed value of 3. We found this trick of prompting the model to generate more titles at the test time improves its recall over letting the model decide by itself how many titles to generate.

C Comparison with Previous Systems for ODQA with LLMs

We showed a leaderboard-style comparison with previous approaches to using large language models for open-domain QA in § 5. We noted though that the comparison is not head-to-head given various differences. We briefly describe each method and the differences in API, LLM, retrieval corpus, and other choices here.

Internet-Augmented QA (Lazaridou et al., 2022) does (one-step) Google Search retrieval, performs additional LLM-based filtering on it, and then prompts an LLM to answer the question using the resulting context. It uses the Gopher 280B language model. RECITE (Sun et al., 2022) bypasses the retrieval and instead prompts an LLM to first generate (recite) one or several relevant passages from its own memory, and generate the answer conditioned on this generation. They experiment with many LLMs, the highest performing of which is `code-davinci-002` which we report here. ReAct (Yao et al., 2022) prompts LLMs to produce reasoning and action traces where actions are calls to a Wikipedia API to return the summary for a given Wikipedia page title. It uses the PALM 540B model. SelfAsk (Press et al., 2022) prompts LLMs to decompose a question into subquestions and answers these subquestions by issuing separate calls to the Google Search API. It uses the GPT3 (`text-davinci-002`) model. Finally, DecomP (Khot et al., 2023) is a general framework that decomposes a task and delegates sub-tasks to appropriate sub-models. Similar to our system, it uses BM25 Search and the GPT3 (`code-davinci-002`) model. And lastly,

[18] Following are the corpus sizes for the datasets: HotpotQA (5,233,329), 2WikiMultihopQA (430,225), MuSiQue has (139,416), and IIRC (1,882,415)

[19] IIRC corpus has a positional bias, i.e., the majority of supporting paragraphs are always within the first few positions of the Wikipedia page. To keep the retrieval problem challenging enough we shuffle the paragraphs before indexing the corpus, i.e., we don't use positional information in any way.

Model	$HpQA^{Br}$	HpQA	2WikiMQA	MQ^{2H}	MQ
InterAug (Lazaridou et al., 2022)	– | –	30.3 | –	– | –	– | –	– | –
RECITE (Sun et al., 2022)	– | –	37.1 | 48.4	– | –	– | –	– | –
ReAct (Yao et al., 2022)	– | –	35.1 | –	– | –	– | –	– | –
SelfAsk (Press et al., 2022)	– | –	– | –	40.1 | –	15.2 | –	– | –
DecomP (Khot et al., 2022)	– | 50.0	– | –	– | 59.3	– | –	– | –
DecomP (Khot et al., 2023) *	– | –	– | 53.5	– | **70.8**	– | –	– | 30.9
DSP (Khattab et al., 2023) *	– | –	**51.4 | 62.9**	– | –	– | –	– | –
IRCoT QA (ours)	**45.8 | 58.5**	49.3 | 60.7	57.7 | 68.0	**34.2 | 43.8**	**26.5 | 36.5**

Table 3: Extended comparison with published LLM-based ODQA systems (as of May 25, 2023) on EM and F1 scores (with new numbers marked with *). '–': score is unavailable. $HpQA^{Br}$: Bridge questions subset of HotpotQA. MQ^{2H}: MuSiQue 2-hop questions. IRCoT remains SOTA for MuSiQue and is close to SOTA for HotpotQA and 2WikiMultihopQA. Note the comparisons here are not head-to-head as discussed in the text.

		Flan-T5-XXL				GPT3			
	Model	HotpotQA	2WikiMQA	MuSiQue	IIRC	HotpotQA	2WikiMQA	MuSiQue	IIRC
ZeroR QA	Direct	**25.3**± **0.3**	**32.7**± **0.3**	**13.7**± **0.3**	**28.9**± **0.3**	41.0± 1.1	38.5± 1.1	19.0± 1.2	40.9± 0.7
	CoT	22.9± 0.1	31.7± 1.5	10.3± 0.5	24.4± 0.1	**47.5**± **0.4**	**41.2**± **1.0**	**25.2**± **1.2**	**52.1**± **0.1**
OneR QA	Direct	**49.7**± **0.5**	**51.2**± **0.3**	**25.8**± **0.6**	**40.0**± **1.3**	50.7± 0.1	46.4± 2.9	20.4± 0.3	40.1± 0.9
	CoT	43.1± 0.7	47.8± 0.9	17.6± 0.2	34.5± 1.5	**53.6**± **0.7**	**54.8**± **2.1**	**29.4**± **0.8**	**49.8**± **2.3**
IRCoT QA	Direct	**59.1**± **0.9**	**66.5**± **1.4**	**30.8**± **0.2**	**42.5**± **2.1**	60.6± 1.0	63.5± 2.7	36.0± 0.5	47.9± 2.3
	CoT	52.0± 0.6	55.1± 1.0	24.9± 1.0	36.5± 1.3	**60.7**± **1.1**	**68.0**± **1.5**	**36.5**± **1.2**	**49.9**± **1.1**

Table 4: Answer F1 for different ODQA models made from NoR, One and IRCoT retrievals, and Direct and CoT prompting readers. For `Flan-T5-XXL`, Direct prompting is a better choice for the reader, and for `GPT3`, CoT prompting is a better choice for the reader. Hence, we make different reader choices for Flan-T5 and GPT3 for the experiments in the main paper. Note that IRCoT QA > OneR QA > ZeroR QA holds up regardless of this choice.

DSP (Khattab et al., 2023) provides a way to programmatically define interactions between LLM and retrieval for ODQA (e.g., via question decomposition), bootstrap demonstrations for such a program, and use them to make the answer prediction. It uses GPT3.5 LLM with ColBERT-based retrieval. Since most of these methods use different knowledge sources or APIs and are built using different LLMs and retrieval models, it's difficult to make a fair scientific comparison across these systems. Additionally, the evaluations in the respective papers are on different random subsets (from the same distribution) of test instances.

Despite these differences, it is still informative to explore, in a leaderboard-style fashion, how IRCoT performs relative to the best numbers published for these recent systems. Table 3 shows results from different systems, including contemporaneous and newer numbers. The two new systems in this table (relative to Table 1) are DecomP (newer version) and DSP. While IRCoT remains SOTA on MuSiQue, DSP outperforms it on HotpotQA by 2.0 points and the newer version of Decomp outperforms IRCoT on 2WikiMultihopQA by 2.8 points. We speculate DecomP performs well on 2WikiMultihopQA because it has only a few easy-to-predict decomposition patterns, which DecomP's question decomposition can leverage. The lack of such patterns in HotpotQA and MuSiQue causes it to underperform compared to IRCoT. Lastly, it will be useful to assess whether DSP, which is hardcoded for 2-hop questions like that of HotpotQA, will work well for a dataset with a varied number of hops like that of MuSiQue. We leave this further investigation to future work.

D Additional CoT Generation Examples

Table 5 provides illustrations, in addition to the ones provided in Table 2, for how the CoT generations for NoR QA, OneR QA, and IRCoT QA methods vary. This gives an insight into how IRCoT improves QA performance. Since NoR relies completely on parametric knowledge, it often makes a factual error in the first sentence, which derails the full reasoning chain. Some of this factual information can be fixed by OneR, especially information closest to the question (i.e., can be retrieved using the question). This is insufficient for fixing

	Question	Gold Facts	NoR QA	OneR QA	IRCoT QA
HotpotQA	What was the island, on which Marinelli Glacier is located, formerly known as?	⇒ Marinelli Glacier is located on the island of Tierra del Fuego. ⇒ The island of Tierra del Fuego was formerly known as Isla de Xativa.	Marinelli Glacier is located on the island of Graham Land. Graham Land was formerly known as Graham's Land. So the answer is: Graham's Land.	Marinelli Glacier is located on Isla Grande de Tierra del Fuego. Isla Grande de Tierra del Fuego was formerly known as Tierra del Fuego. So the answer is: Tierra del Fuego.	Marinelli Glacier is located on the island of Tierra del Fuego. The island of Tierra del Fuego was formerly known as Isla de Xativa. So the answer is: Isla de Xativa.
2WikiMQA	Who is the child of the director of film Mukhyamantri (1996 Film)?	⇒ The director of Mukhyamantri (1996 film) is Anjan Choudhury. ⇒ Anjan Choudhury has a child named Sandip Choudhury.	The film Mukhyamantri was directed by S. V. Rajendra Singh Babu. S. V. Rajendra Singh Babu has a child named S. V. Rajendra Singh Babu. So the answer is: S. V. Rajendra Singh Babu.	Mukhyamantri (1996 film) was directed by Anjan Choudhury. Anjan Choudhury has a child named Aniruddha Choudhury. So the answer is: Aniruddha Choudhury.	The director of Mukhyamantri (1996 film) is Anjan Choudhury. Anjan Choudhury has a child named Sandip Choudhury. So the answer is: Sandip Choudhury.
MuSiQue	When did the Governorship end of the city where the author of Sacerdotii Nostri Primordia died?	⇒ Sacerdotii Nostri Primordia was written by Pope John XXIII. ⇒ Pope John XXIII died in Vatican City. ⇒ The Governorship of Vatican City ended in 1952.	The author of Sacerdotii Nostri Primordia is Pope Pius IX. Pope Pius IX died in the city of Rome. The Governorship of Rome ended in 1870. So the answer is: 1870.	Sacerdotii Nostri Primordia was written by Pope John XXIII. Pope John XXIII died in the city of Rome. The Governorship of Rome ended in 1870. So the answer is: 1870.	Sacerdotii Nostri Primordia was written by Pope John XXIII. Pope John XXIII died in Vatican City. The Governorship of Vatican City ended in 1952. So the answer is: 1952.

Table 5: Additional CoTs generated by GPT3 with different methods. ZeroR is most prone to factual errors. OneR often fixes some of the factual information which is closest to the question but doesn't always fix it all the way. Since IRCoT retrieves after each step, it can also fix the errors at each step. More examples are in Table 2.

all the mistakes. Since IRCoT involves retrieval after each step, it can fix errors at each step.

E Direct vs CoT Prompting Readers

Table 4 compares reader choice (Direct vs CoT Prompting) for Flan-T5-XXL and GPT3. We find that Flan-T5-XXL works better with Direct Prompting as a reader and GPT3 works better with CoT Prompting as a reader. Therefore, for the experiments in the main paper, we go with this choice. Note though that the trends discussed in § 5 (IRCoT QA > OneR QA > ZeroR QA) hold regardless of the choice of the reader.

F Separate Reader in IRCoT QA

IRCoT, by construction, produces a CoT as a part of its retrieval process. So, instead of having a separate post-hoc reader, one can also just extract the answer from the CoT generated during retrieval. As Table 6 shows the effect of such an ablation. For `Flan-T5-XXL` having a separate reader is significantly better. For GPT3, this is not always true, but at least a model with a separate reader is always better or close to the one without. So overall we go with the choice of using the reader for the experiments in this paper.

	Model	HotpotQA	2WikiMQA	MuSiQue	IIRC
Flan	IRCoT QA	**59.1**± 0.9	**66.5**± 1.4	**30.8**± 0.2	**42.5**± 2.1
	w/o reader	52.6± 0.3	60.9± 0.6	24.9± 0.2	40.3± 0.2
GPT3	IRCoT QA	60.7± 1.1	68.0± 1.5	**36.5**± 1.2	**49.9**± 1.1
	w/o reader	**61.0**± 0.7	**70.4**± 1.5	31.5± 0.6	48.4± 1.0

Table 6: Answer F1 of IRCoT QA with and without a separate reader for `Flan-T5-XXL` (top two rows) and `GPT3` (bottom two rows). When the reader is not used, the answer is extracted from the CoT generated by IRCoT while doing the retrieval. Ablating the reader usually hurts the performance.

G Prompts

Our manually written chain-of-thought annotations for HotpotQA, 2WikiMultihopQA, MuSiQue, and IIRC are given in Listing 1, 2, 3 and 4 respectively. Our prompts for GPT3 CoT Prompting are the same as these, except they have Wikipipedia paragraphs on the top of the questions as shown in § 3.1[20]. Our prompts for GPT3 Direct Prompting are the same as that of CoT prompting, except have the answer after "A:" directly. Our prompts for Flan-T5-* are slightly different from that of GPT3. For CoT Prompting, we prefix the question

[20]We are not showing the paragraphs in the paper for brevity but they can be obtained from the released code.

line: "Q: Answer the following question by reasoning step-by-step. <actual-question>". For Direct Prompting, we prefix the question line: "Q: Answer the following question. <actual-question>". We did this to follow Flan-T5-*'s training format and found it to help its CoT generation.

Listing 1: Chain-of-Thought annotations for HotpotQA.

Q: Jeremy Theobald and Christopher Nolan share what profession?
A: Jeremy Theobald is an actor and producer. Christopher Nolan is a director, producer, and screenwriter. Therefore, they both share the profession of being a producer. So the answer is: producer.

Q: What film directed by Brian Patrick Butler was inspired by a film directed by F.W. Murnau?
A: Brian Patrick Butler directed the film The Phantom Hour. The Phantom Hour was inspired by the films such as Nosferatu and The Cabinet of Dr. Caligari. Of these Nosferatu was directed by F.W. Murnau. So the answer is: The Phantom Hour.

Q: How many episodes were in the South Korean television series in which Ryu Hye−young played Bo−ra?
A: The South Korean television series in which Ryu Hye−young played Bo−ra is Reply 1988. The number of episodes Reply 1988 has is 20. So the answer is: 20.

Q: Vertical Limit stars which actor who also played astronaut Alan Shepard in "The Right Stuff"?
A: The actor who played astronaut Alan Shepard in "The Right Stuff" is Scott Glenn. The movie Vertical Limit also starred Scott Glenn. So the answer is: Scott Glenn.

Q: What was the 2014 population of the city where Lake Wales Medical Center is located?
A: Lake Wales Medical Center is located in the city of Polk County, Florida. The population of Polk County in 2014 was 15,140. So the answer is: 15,140.

Q: Who was born first? Jan de Bont or Raoul Walsh?
A: Jan de Bont was born on 22 October 1943. Raoul Walsh was born on March 11, 1887. Thus, Raoul Walsh was born the first. So the answer is: Raoul Walsh.

Q: In what country was Lost Gravity manufactured?
A: The Lost Gravity (roller coaster) was manufactured by Mack Rides. Mack Rides is a German company. So the answer is: Germany.

Q: Which of the following had a debut album entitled "We Have an Emergency": Hot Hot Heat or The Operation M.D.?
A: The debut album of the band "Hot Hot Heat" was "Make Up the Breakdown". The debut album of the band "The Operation M.D." was "We Have an Emergency". So the answer is: The Operation M.D..

Q: In which country did this Australian who was detained in Guantanamo Bay detention camp and published "Guantanamo: My Journey" receive para−military training?
A: The Australian who was detained in Guantanamo Bay detention camp and published "Guantanamo: My Journey" is David Hicks. David Hicks received his para−military training in Afghanistan. So the answer is: Afghanistan.

Q: Does The Border Surrender or Unsane have more members?
A: The Border Surrender band has following members: Keith Austin, Simon Shields, Johnny Manning and Mark Austin. That is, it has 4 members. Unsane is a trio of 3 members. Thus, The Border Surrender has more members. So the answer is: The Border Surrender.

Q: Which band formed first, Sponge Cola or Hurricane No. 1?
A: Sponge Cola band was formed in 1998. Hurricane No. 1 was formed in 1996. Thus, Hurricane No. 1 band formed the first. So the answer is: Hurricane No. 1.

Q: James Paris Lee is best known for investing the Lee−Metford rifle and another rifle often referred to by what acronymn?
A: James Paris Lee is best known for investing the Lee−Metford rifle and LeeâĂŞEnfield series of rifles. LeeâĂŞEnfield is often referred to by the acronym of SMLE. So the answer is: SMLE.

Q: Who was born first, James D Grant, who uses the pen name of Lee Child, or Bernhard Schlink?
A: James D Grant, who uses the pen name of Lee Child, was born in 1954. Bernhard Schlink was born in 1944. Thus, Bernhard Schlink was born first. So the answer is: Bernhard Schlink.

Q: Which American neo−noir science fiction has Pierce Gagnon starred?
A: Pierce Gagnon has starred in One Tree Hill, Looper, Wish I Was Here and Extant. Of these, Looper is an American neo−noir science fiction. So the answer is: Looper.

Q: What year did Edburga of Minster−in−Thanet's father die?
A: The father of Edburga of Minster−in−Thanet is King Centwine. Centwine died after 685. So the answer is: after 685.

Q: Were Lonny and Allure both founded in the 1990s?
A: Lonny (magazine) was founded in 2009. Allure (magazine) was founded in 1991. Thus, of the two, only Allure was founded in 1990s. So the answer is: no.

Q: The actor that stars as Joe Proctor on the series "Power" also played a character on "Entourage" that has what last name?

A: The actor that stars as Joe Proctor on the series "Power" is Jerry Ferrara. Jerry Ferrara also played a character on Entourage named Turtle Assante. Thus, Turtle Assante's last name is Assante. So the answer is: Assante.

Q: Nobody Loves You was written by John Lennon and released on what album that was issued by Apple Records, and was written, recorded, and released during his 18 month separation from Yoko Ono?
A: The album issued by Apple Records, and written, recorded, and released during John Lennon's 18 month separation from Yoko Ono is Walls and Bridges. Nobody Loves You was written by John Lennon on Walls and Bridges album. So the answer is: Walls and Bridges.

Q: How many awards did the "A Girl Like Me" singer win at the American Music Awards of 2012?
A: The singer of "A Girl Like Me" singer is Rihanna. In the American Music Awards of 2012, Rihana won one award. So the answer is: one.

Q: who is older Jeremy Horn or Renato Sobral ?
A: Jeremy Horn was born on August 25, 1975. Renato Sobral was born on September 7, 1975. Thus, Jeremy Horn is older. So the answer is: Jeremy Horn.

Listing 2: Chain-of-Thought annotations for 2WikiMultihopQA.

Q: When did the director of film Hypocrite (Film) die?
A: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is: 19 June 2013.

Q: Do director of film Coolie No. 1 (1995 Film) and director of film The Sensational Trial have the same nationality?
A: Coolie No. 1 (1995 film) was directed by David Dhawan. The Sensational Trial was directed by Karl Freund. David Dhawan's nationality is India. Karl Freund's nationality is Germany. Thus, they do not have the same nationality. So the answer is: no.

Q: Are both Kurram Garhi and Trojkrsti located in the same country?
A: Kurram Garhi is located in the country of Pakistan. Trojkrsti is located in the country of Republic of Macedonia. Thus, they are not in the same country. So the answer is: no.

Q: Who was born first out of Martin Hodge and Ivania Martinich?
A: Martin Hodge was born on 4 February 1959. Ivania Martinich was born on 25 July 1995. Thus, Martin Hodge was born first. So the answer is: Martin Hodge.

Q: Which film came out first, The Night Of Tricks or The Genealogy?
A: The Night of Tricks was published in the year 1939. The Genealogy was published in the year 1979. Thus, The Night of Tricks came out first. So the answer is: The Night Of Tricks.

Q: When did the director of film Laughter In Hell die?
A: The film Laughter In Hell was directed by Edward L. Cahn. Edward L. Cahn died on August 25, 1963. So the answer is: August 25, 1963.

Q: Which film has the director died later, The Gal Who Took the West or Twenty Plus Two?
A: The film Twenty Plus Two was directed by Joseph M. Newman. The Gal Who Took the West was directed by Frederick de Cordova. Joseph M. Newman died on January 23, 2006. Fred de Cordova died on September 15, 2001. Thus, the person to die later from the two is Twenty Plus Two. So the answer is: Twenty Plus Two.

Q: Who is Boraqchin (Wife Of ÃŰgedei)'s father−in−law?
A: Boraqchin is married to ÃŰgedei Khan. ÃŰgedei Khan's father is Genghis Khan. Thus, Boraqchin's father−in−law is Genghis Khan. So the answer is: Genghis Khan.

Q: What is the cause of death of Grand Duke Alexei Alexandrovich Of Russia's mother?
A: The mother of Grand Duke Alexei Alexandrovich of Russia is Maria Alexandrovna. Maria Alexandrovna died from tuberculosis. So the answer is: tuberculosis.

Q: Which film has the director died earlier, When The Mad Aunts Arrive or The Miracle Worker (1962 Film)?
A: When The Mad Aunts Arrive was directed by Franz Josef Gottlieb. The Miracle Worker (1962 film) was directed by Arthur Penn. Franz Josef Gottlieb died on 23 July 2006. Arthur Penn died on September 28, 2010. Thus, of the two, the director to die earlier is Franz Josef Gottlieb, who directed When The Mad Aunts Arrive. So the answer is: When The Mad Aunts Arrive.

Q: Which album was released earlier, What'S Inside or Cassandra'S Dream (Album)?
A: What's Inside was released in the year 1995. Cassandra's Dream (album) was released in the year 2008. Thus, of the two, the album to release earlier is What's Inside. So the answer is: What's Inside.

Q: Are both mountains, Serre Mourene and Monte Galbiga, located in the same country?
A: Serre Mourene is located in Spain. Monte Galbiga is located in Italy. Thus, the two countries are not located in the same

country. So the answer is: no.

Q: What is the date of birth of the director of film Best Friends (1982 Film)?
A: The film Best Friends was directed by Norman Jewison. Norman Jewison was born on July 21, 1926. So the answer is: July 21, 1926.

Q: Which film has the director born first, Two Weeks With Pay or Chhailla Babu?
A: Two Weeks with Pay was directed by Maurice Campbell. Chhailla Babu was directed by Joy Mukherjee. Maurice Campbell was born on November 28, 1919. Joy Mukherjee was born on 24 February 1939. Thus, from the two directors, Chhailla Babu was born first, who directed Two Weeks With Pay. So the answer is: Two Weeks With Pay.

Q: Who is the grandchild of Krishna Shah (Nepalese Royal)?
A: Krishna Shah has a child named Rudra Shah. Rudra Shah has a child named Prithvipati Shah. Thus, Krishna Shah has a grandchild named Prithvipati Shah. So the answer is: Prithvipati Shah.

Q: When was the director of film P.S. Jerusalem born?
A: P.S. Jerusalem was directed by Danae Elon. Danae Elon was born on December 23, 1970. So the answer is: December 23, 1970.

Q: Which album was released more recently, If I Have to Stand Alone or Answering Machine Music?
A: If I Have to Stand Alone was published in the year 1991. Answering Machine Music was released in the year 1999. Thus, of the two, the album to release more recently is Answering Machine Music. So the answer is: Answering Machine Music.

Q: Where did the director of film Maddalena (1954 Film) die?
A: The film Maddalena is directed by Augusto Genina. Augusto Genina died in Rome. So the answer is: Rome.

Q: When did the director of film The Boy And The Fog die?
A: The director of The Boy and the Fog is Roberto GavaldÃşn. Roberto GavaldÃşn died on September 4, 1986. So the answer is: September 4, 1986.

Q: Are the directors of films The Sun of the Sleepless and Nevada (1927 film) both from the same country?
A: The director of Sun of the Sleepless is Temur Babluani. The director of Nevada (1927 film) is John Waters. John Waters is from the country of America. Temur Babluani is from the country of Georgia. Thus, John Walters and Temur Babluani are not from the same country. So the answer is: no.

Listing 3: Chain-of-Thought annotations for MuSiQue.

Q: When did the first large winter carnival take place in the city where CIMI−FM is licensed to broadcast?
A: CIMI−FM is licensed to broadcast in Quebec City. The first large winter carnival in Quebec City took place in 1894. So the answer is: 1894.

Q: When was Neville A. Stanton's employer founded?
A: The employer of Neville A. Stanton is University of Southampton. The University of Southampton was founded in 1862. So the answer is: 1862.

Q: What county is Hebron located in, in the same province the Heritage Places Protection Act applies to?
A: Heritage Places Protection Act applies to the jurisdiction of Prince Edward Island. Hebron, Prince Edward Island is located in the Prince County. So the answer is: Prince County.

Q: What weekly publication in the Connecticut city with the most Zagat rated restaurants is issued by university of America−Lite: How Imperial Academia Dismantled Our Culture's author?
A: The author of America−Lite: How Imperial Academia Dismantled Our Culture is David Gelernter. David Gelernter was educated at the Yale University. The city in Connecticut that has the highest number of Zagat−rated restaurants is New Haven. The weekly publication in New Haven that is issued by Yale University is Yale Herald. So the answer is: Yale Herald.

Q: What is the headquarters for the organization who sets the standards for ISO 21500?
A: The standards for ISO 21500 were set by International Organization for Standardization. The International Organization for Standardization has headquarters in Geneva. So the answer is: Geneva.

Q: What did the publisher of Banjo−Tooie rely primarily on for its support?
A: The publisher of Banjo−Tooie is Nintendo. Nintendo relied primarily for its support on first−party games. So the answer is: first−party games.

Q: In which county was the birthplace of the Smoke in tha City performer?
A: The performer of Smoke in tha City is MC Eiht. MC Eiht's birthplace is Compton. Compton is located in the county of Los Angeles County. So the answer is: Los Angeles County.

Q: What region of the state where Guy Shepherdson was born, contains SMA Negeri 68?

A: Guy Shepherdson was born in Jakarta. SMA Negeri 68 Jakarta is located in Central Jakarta. So the answer is: Central Jakarta.

Q: When did Britain withdraw from the country containing Hoora?
A: Hoora is in the country of Bahrain. Britain withdrew from Bahrain in 1971. So the answer is: 1971.

Q: Where does the Snake River start, in the state where Lima Mountain is located?
A: Lima Mountain is located in the state of Minnesota. The snake river in Minnesota starts in southern Aitkin County. So the answer is: southern Aitkin County.

Q: What shares a border with RiviÃĺre−Verte in the province WRSU−FM broadcasts in?
A: WRSU−FM was licensed to broadcast to New Brunswick. RiviÃĺre−Verte, New Brunswick shares border with Edmundston. So the answer is: Edmundston.

Q: When was the state of emergency declared in the country where the Senate is located?
A: The Senate is in the country of Kenya. The state of emergency was declared in Kenya on 20 October 1952. So the answer is: 20 October 1952.

Q: How long is the US border with the country that borders the state where Finding Dory takes place?
A: Finding Dory is supposed to take place in California. The country that shares a border with California is Mexico. The length of the us border with Mexico is 1,989 mi. So the answer is: 1,989 mi.

Q: What genre is the record label of the performer of So Long, See You Tomorrow associated with?
A: The performer of So Long, See You Tomorrow is Bombay Bicycle Club. The record label of Bombay Bicycle Club is Island Records. The genre of Island Records is jazz. So the answer is: jazz.

Q: When did the first large winter carnival happen in Olivier Robitaille's place of birth?
A: Olivier Robitaille was born in Quebec City. The first large winter carnival in Quebec City happened in the 1894. So the answer is: 1894.

Q: What is the genre of the record label of the band that performed on the Crush Tour?
A: The Crush Tour is performed by the band Bon Jovi. The record label of Bon Jovi is Island Records. The genre of Island Records is jazz. So the answer is: jazz.

Q: When was the first railway line constructed between Kotri and the city where Marie Adelaide Leprosy Centre is located?
A: Marie Adelaide Leprosy Centre is located in Karachi. The first railway line between Kotri and Karachi was constructed in April 1858. So the answer is: April 1858.

Q: Where is the crying stone found in the country in which Raphael Tuju holds citizenship?
A: Raphael Tuju is a citizen of Kenya. The crying stone in Kenya is found along the highway towards Kisumu. So the answer is: along the highway towards Kisumu.

Q: When did Britain withdraw from the country where the village of Wadyan is found?
A: Wadyan is in the country of Bahrain. Britain withdraw from Bahrain in 1971. So the answer is: 1971.

Q: How many countries in Pacific National University's continent are recognized by the organization that mediated the truce ending the Iran−Iraq war?
A: Pacific National University is located in Khabarovsk, Russia Khabarovsk, Russian is in the continent of Asia. The entity that mediated the truce which ended the Iran−Iraq War is the UN. The number of member states that UN recognises in Asia is 53. So the answer is: 53.

Listing 4: Chain-of-Thought annotations for IIRC.

Q: What is the age difference between the kicker and the quarterback for the Chargers?
A: The kicker for the Chargers is Nate Kaeding. The quarterback (QB) for the Chargers is Philip Rivers. Nate Kaeding was born in the year 1982. Philip Rivers was born in the year 1981. Thus, the age difference between them is of 1 year. So the answer is: 1.

Q: How many years was the ship that took the battalion from New South Wales to Ceylon in service?
A: The ship that took the battalion from New South Wales to Ceylon is General Hewitt. General Hewitt was launched in Calcutta in 1811. General Hewitt was sold for a hulk or to be broken up in 1864. So she served for a total of 1864 − 1811 = 53 years. So the answer is: 53.

Q: What year was the theatre that held the 2016 NFL Draft built?
A: The theatre that held the 2016 NFL Draft is Auditorium Theatre. The Auditorium Theatre was built in 1889. So the answer is: 1889.

Q: How long had Milan been established by the year that Nava returned there as a reserve in the first team's defense?
A: Nava returned to Milan as a reserve in the first team's defense in the year 1990. Milan had been established in the year

1899. Thus, Milan had been established for 1990 − 1899 = 91 years when Milan returned to Milan as a reserve in the first team's defense. So the answer is: 91.

Q: When was the town Scott was born in founded?
A: Scott was born in the town of Cooksville, Illinois. Cooksville was founded in the year 1882. So the answer is: 1882.

Q: In what country did Wright leave the French privateers?
A: Wright left the French privateers in Bluefield's river. Bluefields is the capital of the South Caribbean Autonomous Region (RAAS) in the country of Nicaragua. So the answer is: Nicaragua.

Q: Who plays the A−Team character that Dr. Hibbert fashioned his hair after?
A: Dr. Hibbert fashioned his hair after Mr. T from The A−Team. Mr T.'s birthname is Lawrence Tureaud. So the answer is: Lawrence Tureaud.

Q: How many people attended the conference held near Berlin in January 1942?
A: The conference held near Berlin in January 1942 is Wannsee Conference. Wannsee Conference was attended by 15 people. So the answer is: 15.

Q: When did the country Ottwalt went into exile in founded?
A: Ottwalt went into exile in the country of Denmark. Denmark has been inhabited since around 12,500 BC. So the answer is: 12,500 BC.

Q: When was the J2 club Uki played for in 2001 founded?
A: The J2 club that Uki played for is Montedio Yamagata. Montedio Yamagata was founded in 1984. So the answer is: 1984.

Q: When was the person who produced A Little Ain't Enough born?
A: A Little Ain't Enough was produced by Bob Rock. Bob Rock was born on April 19, 1954. So the answer is: April 19, 1954.

Q: Which of the schools Fiser is affiliated with was founded first?
A: The schools that Fiser is affiliated with (1) Academy of Music, University of Zagreb (2) Mozarteum University of Salzburg (3) Croatian Music Institute orchestra. Academy of Music, University of Zagreb was founded in the year 1829. Mozarteum University of Salzburg was founded in the year 1841. Croatian Music Institute was founded in the year 1827. Thus, the school founded earliest of these is Croatian Music Institute. So the answer is: Croatian Music Institute.

Q: How many casualties were there at the battle that Dearing fought at under Jubal Early?
A: Under Jubal Early, Dearing fought the First Battle of Bull Run. First Battle of Bull Run has 460 union casualties and 387 confederate casualties. Thus, in total the First Battle of Bull Run had 460 + 387 = 847 casualties. So the answer is: 847.

Q: Which of the two congregations which provided leadership to the Pilgrims was founded first?
A: The congregations which provided leadership to the Pilgrims are Brownists and Separatist Puritans. Brownist was founded in 1581. The Separatist Puritans was founded in 1640. Thus, Brownist was founded first. So the answer is: Brownist.

Q: How long had the Rock and Roll Hall of Fame been open when the band was inducted into it?
A: The band was inducted into Rock and Roll Hall of Fame in the year 2017. Rock and Roll Hall of Fame was established in the year of 1983. Thus, Rock and Roll Hall of Fame been open for 2018 − 1983 = 34 years when the band was inducted into it. So the answer is: 34.

Q: Did the Lord Sewer who was appointed at the 1509 coronation live longer than his king?
A: Lord Sewer who was appointed at the 1509 coronation was Robert Radcliffe, 1st Earl of Sussex. Lord Sever's king in 1509 was Henry VIII of England. Robert Radcliffe, 1st Earl of Sussex was born in the year 1483, and died in the year 1542. So Robert lived for 1542 − 1483 = 59 years. Henry VIII of England was born in the year 1491 and died in the year 1547. So Henry VIII lived for 1547 − 1491 = 56 years. Thus, Robert Radcliffe lived longer than Henry VIII. So the answer is: yes.

Q: When was the place near where Manuchar was defeated by Qvarqvare established?
A: Manuchar was defeated by Qvarqvare near Erzurum. Erzurum was founded during the Urartian period. So the answer is: Urartian period.

Q: What year was the man who implemented the 46 calendar reform born?
A: The man who implemented the 46 calendar reform is Julius Caesar. Julius Caesar was born in the year 100 BC. So the answer is: 100 BC.

Q: How many years after the first recorded Tommy John surgery did Scott Baker undergo his?
A: The first recorded Tommy John surgery happened when it was invented in the year 1974. Scott Baker underwent Tommy John surgery in the year 2012. Thus, Scott Baker underwent Tommy John surgery 2012 − 1974 = 38 years after it was first recorded. So the answer is: 38.

Q: Which was the older of the two players who found the net in the Double−Headed Eagle of the North in the sixth final for PAOK?

A: The two players who found the net in the Double−Headed Eagle of the North in the sixth final for PAOK are Koudas and Matzourakis. Koudas was born on 23 November 1946. Matzourakis was born on 6 June 1949. Thus, the older person among the two is Koudas. So the answer is: Koudas.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
8

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
8

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4,5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.