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

Recent work has shown that large language models are capable of generating natural language reasoning steps or Chains-of-Thoughts (CoT) to answer a multi-step question when prompted to do so. This is insufficient, however, when the necessary knowledge is not available or up-to-date within a model's parameters. A straightforward approach to address this is to retrieve text from an external knowledge source using the question as a query and prepend it as context to the model's input. This, however, is also insufficient for multi-step QA where *what to retrieve* depends on *what has already been derived*. To address this issue we propose IRCoT, a new approach that interleaves retrieval with CoT for multi-step QA, guiding the retrieval with CoT and in turn using retrieved results to improve CoT. Our experiments with GPT3 show substantial improvements in retrieval (up to 22 points) and downstream QA (up to 16 points) over the baselines on four datasets: HotpotQA, 2WikiMultihopQA, MuSiQue, and IIRC. Notably, our method also works well for much smaller models such as T5-Flan-large (0.7B) without any additional training.[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 successfully applied in settings where all the information needed to answer the question is either provided as context (e.g., algebra questions) or assumed to be present in the parameters of the models (e.g., commonsense reasoning). 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).

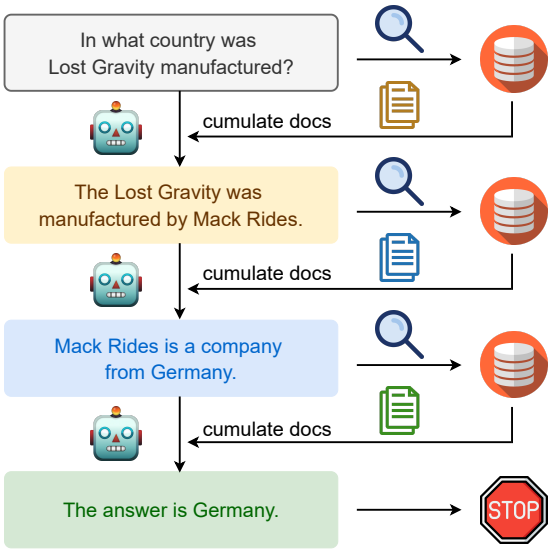

Figure 1: IRCoT interleaves chain-of-thought (CoT) generation and knowledge retrieval steps in order to guide the retrieval by CoT and vice-versa. This interleaving allows retrieving more relevant information for later reasoning steps, compared to standard retrieval using solely the question as the query.

We ask the following question: *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 reasoning done so far, and iterate. As an example, consider the question illustrated in Figure 1, *"In*

[1]The code and prompts are available at https://github.com/StonyBrookNLP/ircot.

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 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. Figure 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 repeat the following two steps alternatingly: (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 knowledge paragraphs and add them 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 12-22 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 16 F1 points) in downstream few-shot QA performance. Our approach also works on much smaller T5-Flan models (11B, 3B, and 0.7B) showing similar trends. In particular, we find QA using T5-Flan-XL (3B) with IRCoT even outperforms the 58X larger GPT3 with a one-step question-based retrieval. Lastly, we compare our work to contemporaneous works on few-shot prompting for open-domain QA (Khot et al., 2022; Press et al., 2022; Yao et al., 2022), and show that our QA scores are higher than these previous reported systems[4].

In summary, we make the following contributions:

1. A novel retrieval method, IRCoT, that leverages language models' chain-of-thought generation capabilities to guide the retrieval and uses retrieval in turn to improve models' CoT reasoning.

2. Demonstrate that IRCoT improves both retrieval and few-shot QA peformance on four multi-step open-domain QA datasets.

3. Demonstrate that IRCoT leads to improved performance with both large-scale (175B parameter models) as well as smaller scale models (Flan-T5-*, $\leq$11B) without any training.

2 Related Work

Prompting LMs for Multi-Step Reasoning

Large Language Models (LLMs) have been shown to learn various tasks by just 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 demonstrations (Wei et al., 2022) or in many cases even without any demonstrations (Kojima et al., 2022). These capabilities have also been extended to smaller models

[2] IRCoT stands for Interleaved Retrieval guided by Chain-of-Thought Reasoning.

[3] We explain later (in the Metric section and Footnote 10) the appropriateness of this metric in our setting as opposed to more mainstream information recall metrics.

[4] Note that due to different design choices made in prior works, it is not possible to make an apples-to-apples comparison and we mainly report them here for completeness.

by additional fine-tuning (Zelikman et al., 2022; Chung et al., 2022).

Despite the recent progress in CoT-based prompting for reasoning, their value in improving the retrieval and QA for multi-step open-domain questions has been relatively underexplored. There have been few recent works that have also explored this particular problem. Nakano et al. (2021) used GPT-3 to answer long-form questions by first interacting with a browser over multiple steps to retrieve evidence and then generating the answer. This approach also relies on human annotations of the interactions with the browser, i.e., not few-shot. Lazaridou et al. (2022) proposed augmenting few-shot QA prompting with the context retrieved from Google Search results. They do not employ multi-step retrieval though and so is closer to our one-step retrieval baseline.

Contemporaneous to our work, three new approaches have been proposed for mutli-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., 2022) is a general framework that decomposes a task and delegates sub-tasks to appropriate sub-models. Similar to SelfAsk, they decompose questions and delegate retrieval to a BM25-based retriever. Both these approaches are not developed for CoT reasoning or focus on the retrieval problem for multi-step QA. Additionally, they require a single-hop QA model to answer their decomposed questions. Recently, the ReAct (Yao et al., 2022) system was proposed that frames the problem as generating a sequence of reasoning and action steps. Their reasoning and action steps are more complex than simple CoT and BM25 retrieval steps in our system. They also rely on prompting much larger models (PaLM-540B) and require fine-tuning for smaller models. In general, unlike our work, none of these contemporaneous works have been shown to be effective for smaller models without any training.

Although a direct comparison with these approaches is not straightforward as the knowledge sources (Wikipedia vs Google), APIs (Wikipedia vs BM25), sub-sampled prompt examples, sub-sampled test sets are different, we do show our ODQA results are significantly better than all their reported numbers where available (see § 5).

Supervised Multi-Step Open-Domain QA

Several previous works have explored iterative retrieval for open-domain QA in the fully supervised setting. Das et al. (2019) proposes an iterative retrieval model that retrieves with neural query representation, applies the reading comprehension model on them and reformulates query representation, and uses it again to perform the retrieval. Feldman and El-Yaniv (2019) apply similar neural query reformulation idea for multihop open-domain QA. Xiong et al. (2021) extend 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) leverage the graph structure induced by the entity-links present in Wikipedia paragraphs. Unlike our method, all of these methods rely on the supervised training on a large-scale dataset and can not be easily applied to a few-shot setting.

In this line of work, GoldEn (Gold Entity) retriever (Qi et al., 2019) is the closest to ours. Like us, they also iteratively generate queries based on the retrieved paragraphs and use an off-the-shelf retriever to retrieve the next round of paragraphs. However, GoldEn requires training data for the next query generator instead of our approach to use few-shot CoT reasoning. Additionally it is unclear how to extend their heuristics to datasets other than HotpotQA.

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).

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

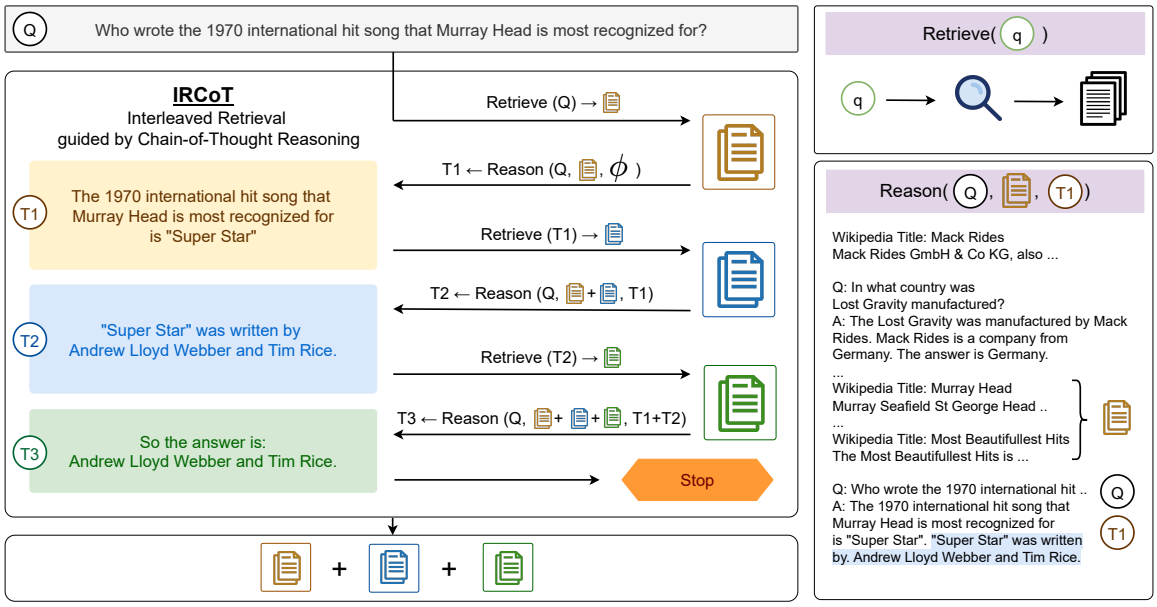

Figure 2: IRCoT interleaves chain-of-thought (CoT) generation and retrieval steps to guide the retrieval by CoT and vice-versa. We start by retrieving K documents using the question as they query and repeat two steps alternatingly until termination. (i) `reason`-step generates next CoT sentence based 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.

zero/few-shot Chain-of-Thought (CoT) generation capabilities (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 K that collectively support the reasoning chain and the answer.

The overview of IRCoT is given in Figure 2. We start by gathering 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 like the following:

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

For the paragraphs in the in-context demonstrations, we show the model ground-truth supporting paragraphs and M randomly sampled paragraphs shuffled and concatenated together in the above format. For the 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[5] 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 cumulating set of the collected paragraphs. We don't allow more paragraphs to be added once its size reaches a fixed upper cap.[6] We keep this limit so that we can fit in at least a few demonstrations

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

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 Appendix B for details.

4 Experimental Setup

4.1 Datasets

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), **MuSiQue** (Trivedi et al., 2022) and **IIRC** (Ferguson et al., 2020).

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[7]. 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[8].

MuSiQue and IIRC come with unanswerable questions, but we limit our experiments to the answerable questions. For each dataset, we randomly sample 100 questions from the original development set and use it as our development set for tuning hyperparameters. We sample 500 questions from the remaining development set and use it as our test set.

4.2 Models

Retriever. We use BM25 (Robertson et al., 2009) implemented in Elasticsearch[9] 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 the optimal $K \in \{5, 7, 9, 11, 13, 15\}$ based 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 T5-Flan (Chung et al., 2022) of different sizes as its underlying CoT generator.

For demonstrating in-context examples to these language models, we hand-wrote CoTs for 15 questions for all the datasets and used them in all the experiments (see App. §B). At test time, we dynamically pack as many demonstrations as possible within the model's context length limit. The context limit for GPT3 (`code-davinci-002`) is 8000 wordpieces. Flan-T5-* doesn't have any hard limit as it uses relative position embeddings. But we limit Flan-T5's context to 6000 wordpieces, which is the maximum we could fit in the memory our GPUs (80G A100).

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 (see § 3.1).

Retrieval Metric: We allow a maximum of 15

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

[8] 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.

[9] `https://www.elastic.co/`

paragraphs for all the 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 development 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.[10]

QA Reader. To implement the QA reader, we use the same language models as used in the reason-step of IRCoT Retriever. We found that QA readers implemented with T5-Flan-* perform better with the Direct Prompting strategy and GPT3 performs better with CoT Prompting strategy (see Appendix A). So we use Direct prompting strategy for QA with T5-Flan and CoT with GPT3 for the experiments.

The QA reader has only one hyperparameter M: the number of distractor paragraphs in the in-context demonstrations. We search for it in $M \in \{1, 2, 3\}$. When used in conjunction with IRCoT retriever this choice 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: **OneR-[LM-name]** and **IRCoT-[LM-name]**. For IRCoT, the choice of LM for the CoT generator and the reader is kept the same. We also experiment with retriever-less QA readers **NoR-[LM-name]** 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 maximizes the answer F1 on the development set.

Special handling 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 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.

[10]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 by question, 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.

5 Results

We begin with our three main results, demonstrating the effectiveness of (a) IRCoT as a retriever, (b) QA models build using IRCoT's output as end-task models, and (c) IRCoT built using smaller LMs. We will then discuss ablation results and our other findings.

IRCoT outperforms one-step retrieval. Figure 3 compares one-step retriever (OneR) with IRCoT retrievers made from `Flan-T5-XXL` and `GPT3` models. For both models, IRCoT significantly outperforms one-step retrieval in all the datasets. For `Flan-T5-XXL`, IRCoT improves our recall metric, relative to one-step retrieval, on HotpotQA by 8.0, on 2WikiMultihopQA by 14.4, on MuSiQue by 3.8, and on IIRC by 8.4 points. For `GPT3`, IRCoT improves our recall metric on the four datasets by 10.8, 21.6, 12.0, and 22.0 points, respectively.

IRCoT QA outperforms NoR and OneR QA. Figure 4 compares open-domain QA performance using no retrieval (NoR), one-step retriever (OneR), and IRCoT retriever made from `Flan-T5-XXL` and `GPT3` models. For `Flan-T5-XXL`, IRCoT QA outperforms OneR QA on HotpotQA by 10.9, on 2WikiMultihopQA by 16.0, on MuSiQue by 5.1 and IIRC by 2.2 F1 points. For `GPT3`, IRCoT QA outperforms OneR QA on HotpotQA by 8.5, on 2WikiMultihopQA by 15.8, and on MuSiQue by 9.2 F1 points. For `GPT3`, IRCoT doesn't improve the QA score on IIRC, despite significantly improved retrieval (22 points as shown in Figure 3). This is likely because the knowledge required to do well on IIRC is already there in GPT3, which is also evidenced by its NoR QA score being roughly the same as OneR QA and IRCoT QA scores. For all other datasets and model combinations, NoR QA is significantly worse than IRCoT QA indicating that models' parametric knowledge is not enough for these tasks.

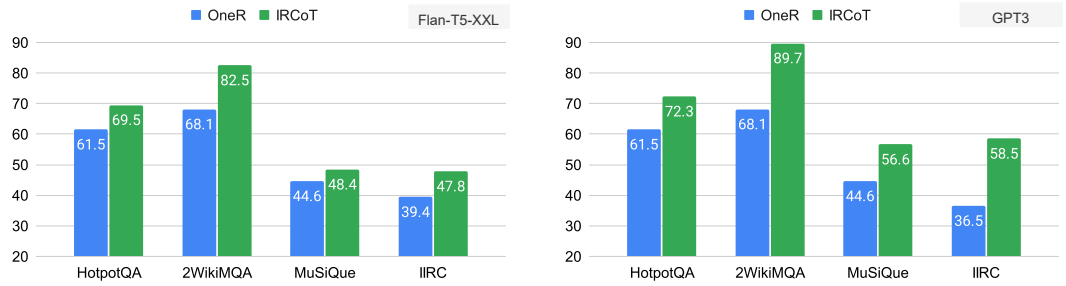

Figure 3: Retrieval recall for one-step retriever (OneR) and IRCoT instantiated from `Flan-T5-XXL` (left) and `GPT3` (right) models. IRCoT outperforms OneR for both models and all datasets.

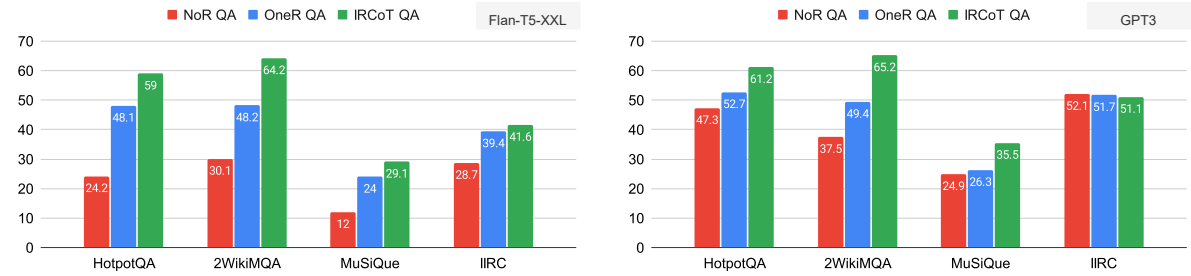

Figure 4: Answer F1 for ODQA model made using (i) no retriever (NoR QA) (ii) one-step retriever (OneR QA) and (iii) IRCoT QA instantiated from `Flan-T5-XXL` (left) and `GPT3` (right) models. IRCoT QA outperforms OneR QA and NoR QA for both models and all datasets, except for GPT3 on IIRC.

IRCoT is effective for smaller models too. To understand how effective IRCoT is at different language model sizes, we show the scaling plots in Figure 5. We compare the recall for OneR and IRCoT using `Flan-T5-base` (0.2B), `Flan-T5-large` (0.7B), `Flan-T5-XL` (3B), `Flan-T5-XXL` (11B), and GPT3 `code-davinci-002` (175B). We see that IRCoT with even the smallest model (0.2B) is better than OneR, and the performance roughly improves with the model size. This shows that the CoT generation capabilities of even small models can be leveraged for improving retrieval. Furthermore, we also show the effect of model size on the QA score in Figure 6. For all model sizes except the smallest (0.2B), we see IRCoT QA is better than OneR QA. Moreover, IRCoT with a 3B-sized model even outperforms OneR and NoR with the 58 times larger 175B-sized GPT3 model in all three datasets.

5.1 Ablations and Other Findings

Separate reader in IRCoT QA helps. 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 1 shows, for both Flan-T5-XXL and GPT3, this doesn't work as well. This illustrates the importance of having a separate QA reader after the retrieval step. This separate reader has a chance to consider all collected evidence together and rethink the full reasoning in this new light.

	Model	HotpotQA	2WikiMQA	MuSiQue	IIRC
Flan	IRCoT QA	**59.0**	**64.2**	**29.1**	**41.6**
	w/o reader	52.3	57.9	23.4	40.4
GPT3	IRCoT QA	**61.2**	**65.2**	**35.5**	**51.1**
	w/o reader	59.1	65.1	30.7	49.5

Table 1: 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 hurts the performance showing its importance.

Comparison to Few-shot Prompting for ODQA We compare IRCoT QA with four recent approaches to using large language models for open-domain QA. Internet-Augmented QA (Lazaridou et al., 2022) does (one-step) Google Search retrieval, performs additional LLM-based filtering

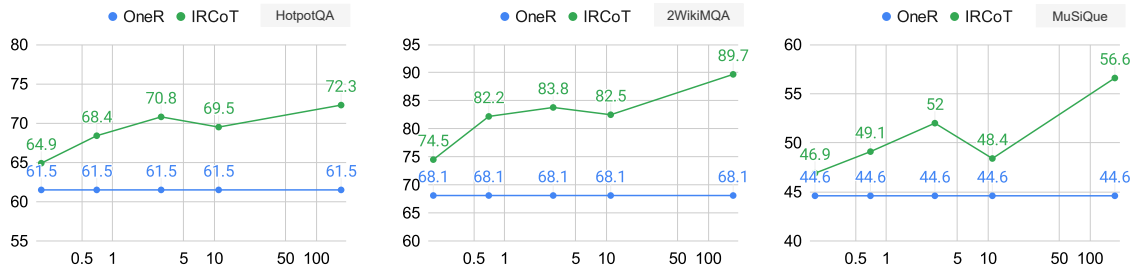

Figure 5: Retrieval recall for OneR (bottom line) and IRCoT (top line) for language models of increasing sizes: Flan-T5-base (0.3B), Flan-T5-large (0.7B), Flan-T5-XL (3B), Flan-T5-XXL (11B), GPT3 (175B) on HotpotQA (left), 2WikiMultihopQA (middle), MuSiQue (right). IRCoT outperforms OneR for all model sizes, including the 0.3B model, and the difference roughly grows with model size. Note that OneR does not interleave any LM in its retrieval and hence has a fixed score.

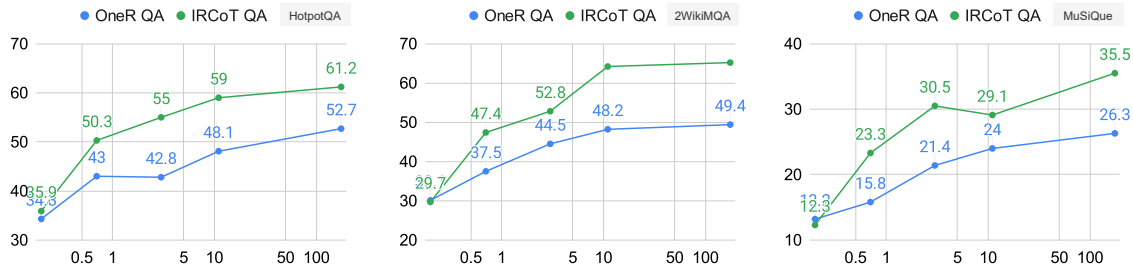

Figure 6: Answer F1 for ODQA models made using OneR (bottom line) and IRCoT (top line) for language models of increasing sizes: Flan-T5-base (0.3B), Flan-T5-large (0.7B), Flan-T5-XL (3B), Flan-T5-XXL (11B), GPT3 (175B) on HotpotQA (left), 2WikiMultihopQA (middle), MuSiQue (right). IRCoT QA outperforms 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.

.

on it, and then prompts an LLM to answer the question using the resulting context. It uses the Gopher 280B language model. 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., 2022) 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. Since most of these methods use different knowledge sources or APIs and are built using different LLMs, 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, as summarized in Table 2. We find that 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 on these datasets via retrieval augmented LLMs.

6 Conclusions

Chain-of-thought prompting has significantly improved prompting-based large language models' ability to perform multi-step reasoning. In this work, 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. One-step question based retrieval is insufficient for such tasks as what information is needed for later steps is not evident from the question alone. To address this, we introduced IRCoT, which uses interleaved chain-of-thought reasoning and retrieval steps that guide each other step-

Model	$HpQA^{Br}$	HpQA	2WikiMQA	MQ^{2H}
InterAug	– | –	30.3 | –	– | –	– | –
ReAct	– | –	35.1 | –	– | –	– | –
SelfAsk	– | –	– | –	40.1 | –	15.2 | –
DecomP	– | 50.0	– | –	– | 59.3	– | –
IRCoT QA	**46.8 | 58.8**	**50.4 | 61.2**	**53.4 | 65.2**	**31.9 | 42.0**

Table 2: Our Answer (EM | F1) scores compared with previously published scores on four recent large language model based ODQA systems. '–' indicates the score for that dataset isn't available for the corresponding system. $HpQA^{Br}$ refers to 'bridge' typed question of HotpotQA. MQ^{2H} refers to 2-hop subset of MuSiQue. IRCoT QA (ours) outperforms all four previous systems by a large margin. Note: These are not head-to-head comparisons as different methods use different APIs, knowledge sources, and even LLMs. We report these numbers mainly for completeness.

by-step. Empirical evaluations on four datasets showed that IRCoT significantly improves retrieval performance as well as few-shot open-domain QA performance when compared to one-step retrieval, for both large and relatively smaller scale language models.

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.

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. 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*.

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. *arXiv preprint arXiv:2210.02406*.

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.

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, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

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, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

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 Direct vs CoT Prompting Readers

Table 3 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.

B 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 3.2[11]. 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 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's training format and found it to help its CoT generation. Note that each of these Listings having 20 annotations, but we randomly sampled 15 of them for our experiments. Please check the code for details.

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

		Flan-T5-XXL				GPT3			
	Model	HotpotQA	2WikiMQA	MuSiQue	IIRC	HotpotQA	2WikiMQA	MuSiQue	IIRC
ZeroR QA	Direct	**24.2**	**30.1**	**12.0**	**28.7**	41.4	34.9	18.1	41.6
	CoT	22.3	30.0	9.8	24.5	**47.3**	**37.5**	**24.9**	**52.1**
OneR QA	Direct	**48.1**	**48.2**	**24.0**	**39.4**	49.7	41.1	18.3	41.6
	CoT	42.8	45.5	16.6	34.0	**52.7**	**49.4**	**26.3**	**51.7**
IRCoT QA	Direct	**59.0**	**64.2**	**29.1**	**41.6**	59.9	60.2	34.1	49.2
	CoT	51.0	51.3	22.9	36.1	**61.2**	**65.2**	**35.5**	**51.1**

Table 3: 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.

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.