The CoT COLLECTION: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning

Seungone Kim\textsuperscript{1,2,3}\textsuperscript{*†} Se June Joo\textsuperscript{1,3}\textsuperscript{*†} Doyoung Kim\textsuperscript{1} Joel Jang\textsuperscript{1,3} Seonghyeon Ye\textsuperscript{1} Jamin Shin\textsuperscript{2,†} Minjoon Seo\textsuperscript{1,†}

KAIST AI\textsuperscript{1} NAVER AI Lab\textsuperscript{2} University of Washington\textsuperscript{3}
{seungone, sejune, minjoon}@kaist.ac.kr jayshin.nlp@gmail.com

Abstract

Language models (LMs) with less than 100B parameters are known to perform poorly on chain-of-thought (CoT) reasoning in contrast to large LMs when solving unseen tasks. In this work, we aim to equip smaller LMs with the step-by-step reasoning capability by instruction tuning with CoT rationales. In order to achieve this goal, we first introduce a new instruction-tuning dataset called the CoT COLLECTION, which augments the existing Flan Collection (including only 9 CoT tasks) with additional 1.84 million rationales across 1,060 tasks. We show that CoT fine-tuning Flan-T5 (3B & 11B) with CoT COLLECTION enables smaller LMs to have better CoT capabilities on unseen tasks. On the BIG-Bench-Hard (BBH) benchmark, we report an average improvement of +4.34\% (Flan-T5 3B) and +2.60\% (Flan-T5 11B), in terms of zero-shot task accuracy. Furthermore, we show that instruction tuning with CoT COLLECTION allows LMs to possess stronger few-shot learning capabilities on 4 domain-specific tasks, resulting in an improvement of +2.24\% (Flan-T5 3B) and +2.37\% (Flan-T5 11B), even outperforming ChatGPT utilizing demonstrations until the max length by a +13.98\% margin. Our code, the CoT COLLECTION data, and model checkpoints are publicly available \textsuperscript{1}.

1 Introduction

Language models (LMs) pre-trained on massive text corpora can adapt to downstream tasks in both zero-shot and few-shot learning settings by incorporating task instructions and demonstrations (Brown et al., 2020; Wei et al., 2021; Sanh et al., 2021; Mishra et al., 2022; Wang et al., 2022b; Iyer et al., 2022; Liu et al., 2022b; Chung et al., 2022; Longpre et al., 2023; Ye et al., 2023). One approach that has been particularly effective in enabling LMs to excel at a multitude of tasks is Chain-of-Thought (CoT) prompting, making LMs generate a rationale to derive its final prediction in a sequential manner (Wei et al., 2022b; Kojima et al., 2022; Zhou et al., 2022; Zhang et al., 2022; Yao et al., 2023).

While CoT prompting works effectively for large LMs with more than 100 billion parameters, it does not necessarily confer the same benefits to smaller LMs (Tay et al., 2022; Suzgun et al., 2022; Wei et al., 2022a; Chung et al., 2022). The requirement of a large number of parameters consequently results in significant computational cost and accessibility issues (Kaplan et al., 2020; Min et al., 2022; Liu et al., 2022b; Mhlanga, 2023; Li et al., 2023).

Recent work has focused on empowering relatively smaller LMs to effectively solve novel tasks as well, primarily through fine-tuning with rationales (denoted as CoT fine-tuning) and applying CoT prompting on a single target task (Shridhar et al., 2022; Ho et al., 2022; Fu et al., 2023). However, solving a single task does not adequately address the issue of generalization to a broad range of unseen tasks. While Chung et al. (2022) leverage 9 publicly available CoT tasks during instruction tuning to solve multiple unseen tasks, the imbalanced ratio compared to 1,827 tasks used for direct fine-tuning results in poor CoT results across smaller LMs (Longpre et al., 2023). In general, the community still lacks a comprehensive strategy to fully leverage CoT prompting to solve multiple unseen novel tasks in the context of smaller LMs.

To bridge this gap, we present the CoT COLLECTION, an instruction tuning dataset that augments 1.84 million rationales from the FLAN Collection (Longpre et al., 2023) across 1,060 tasks. We fine-tune Flan-T5 (3B & 11B) using CoT COLLECTION and denote the resulting model as CoT-T5. We perform extensive comparisons of CoT-T5 and Flan-T5 under two main scenarios: (1) zero-shot learning and (2) few-shot learning.

\textsuperscript{*}denotes equal contribution. Work was done while Seun-gone was interning at NAVER AI Lab.
\textsuperscript{†}Corresponding authors
\textsuperscript{1}https://github.com/kaistAI/CoT-Collection
In the zero-shot learning setting, CoT-T5 (3B & 11B) outperforms Flan-T5 (3B & 11B) by +4.34% and +2.60% on average accuracy across 27 datasets from the Big Bench Hard (BBH) benchmark (Suzgun et al., 2022) when evaluated with CoT prompting. During ablation experiments, we show that CoT fine-tuning T0 (3B) (Sanh et al., 2021) on a subset of the CoT Collection, specifically 163 training tasks used in T0, shows a performance increase of +8.65% on average accuracy across 11 datasets from the P3 Evaluation benchmark. Moreover, we translate 80K instances of CoT COLLECTION into 5 different languages (French, Japanese, Korean, Russian, Chinese) and observe that CoT fine-tuning mT0 (3B) (Muennighoff et al., 2022) on each language results in 2x ~ 10x performance improvement on average accuracy across all 5 languages from the MGSM benchmark (Shi et al., 2022).

In the few-shot learning setting, where LMs must adapt to new tasks with a minimal number of instances, CoT-T5 (3B & 11B) exhibits a +2.24% and +2.37% improvement on average compared to using Flan-T5 (3B & 11B) as the base model on 4 different domain-specific tasks\(^2\). Moreover, it demonstrates +13.98% and +8.11% improvement over ChatGPT (OpenAI, 2022) and Claude (Anthropic, 2023) that leverages ICL with demonstrations up to the maximum input length.

Our contributions are summarized as follows:

- We introduce CoT COLLECTION, a new instruction dataset that includes 1.84 million rationales across 1,060 tasks that could be used for applying CoT fine-tuning to LMs.
- With CoT COLLECTION, we fine-tune Flan-T5, denoted as CoT-T5, which shows a non-trivial boost in zero-shot and few-shot learning capabilities with CoT Prompting.
- For ablations, we show that CoT fine-tuning could improve the CoT capabilities of LMs in low-compute settings by using a subset of CoT COLLECTION and training on (1) smaller number of tasks (T0 setting; 163 tasks) and (2) smaller amount of instances in 5 different languages (French, Japanese, Korean, Russian, Chinese; 80K instances).

\(^2\)We assess the efficacy of our approach on 4 domain-specific datasets, two each from legal and medical fields, namely including LEDGAR (Tuggener et al., 2020), Case Hold (Zheng et al., 2021), MedNLI (Romanov and Shivade, 2018), and PubMedQA (Jin et al., 2019). Each dataset is represented by 64 randomly chosen instances.

2 Related Works

2.1 Chain-of-Thought (CoT) Prompting

Wei et al. (2022b) propose Chain of Thought (CoT) Prompting, a technique that triggers the model to generate a rationale before the answer. By generating a rationale, large LMs show improved reasoning abilities when solving challenging tasks. Kojima et al. (2022) show that by appending the phrase ‘Let’s think step by step’, large LMs could perform CoT prompting in a zero-shot setting. Different work propose variants of CoT prompting such as automatically composing CoT demonstrations (Zhang et al., 2022) and performing a fine-grained search through multiple rationale candidates with a tree search algorithm (Yao et al., 2023). While large LMs could solve novel tasks with CoT Prompting, Chung et al. (2022) and Longpre et al. (2023) show that this effectiveness does not necessarily hold for smaller LMs. In this work, we aim to equip smaller LMs with the same capabilities by instruction tuning on large amount of rationales.

2.2 Improving Zero-shot Generalization

Previous work show that instruction tuning enables generalization to multiple unseen tasks (Wei et al., 2021; Sanh et al., 2021; Aribandi et al., 2021; Ouyang et al., 2022; Wang et al., 2022b; Xu et al., 2022). Different work propose to improve instruction tuning by enabling cross-lingual generalization (Muennighoff et al., 2022), improving label generalization capability (Ye et al., 2022), and training modular, expert LMs (Jang et al., 2023). Meanwhile, a line of work shows that CoT fine-tuning could improve the reasoning abilities of LMs on a single-seen task (Zelikman et al., 2022; Shridhar et al., 2022; Ho et al., 2022; Fu et al., 2023). As a follow-up study, we CoT fine-tune on 1,060 instruction tasks and observe a significant improvement in terms of zero-shot generalization on multiple tasks.

2.3 Improving Few-Shot Learning

For adapting LMs to new tasks with a few instances, recent work propose advanced parameter efficient fine-tuning (PEFT) methods, where a small number of trainable parameters are added (Hu et al., 2021; Lester et al., 2021; Liu et al., 2021, 2022b; Asai et al., 2022; Liu et al., 2022c). In this work, we show that a simple recipe of (1) applying LoRA (Hu et al., 2021) to a LM capable of performing CoT reasoning and (2) CoT fine-tuning on a target task results in strong few-shot performance.
3 The CoT Collection

Despite its effectiveness for CoT fine-tuning, rationale data still remains scarce. To the best of our knowledge, recent work mostly rely on 9 publicly available NLP datasets\(^3\) for fine-tuning with rationales (Zelikman et al., 2022; Shridhar et al., 2022; Chung et al., 2022; Ho et al., 2022; Longpre et al., 2023; Fu et al., 2023). This is due to the difficulty in gathering human-authored rationales (Kim et al., 2023). To this end, we create CoT COLLECTION, an instruction-tuning dataset that includes 1.84 million rationales augmented across 1,060 tasks\(^4\). In this section, we explain the datasets we select to augment into rationales and how we perform the overall augmentation process.

Broad Overview Given an input \(X = [I, z]\) composed of an instruction \(I\), and an instance \(z\) along with the answer \(y\), we obtain a rationale \(r\) by applying in-context learning (ICL) with a large LM. Note that this differs from previous works which focused on generating new instances \(z\) using large LMs (West et al., 2022; Liu et al., 2022a; Kim et al., 2022; Honovich et al., 2022; Wang et al., 2022a; Taori et al., 2023; Chiang et al., 2023) while we extend it to generating new rationales \(r\).

Source Dataset Selection As a source dataset to extract rationales, we choose the Flan Collection (Longpre et al., 2023), consisting of 1,836 diverse NLP tasks from P3 (Sanh et al., 2021), SupernaturalInstructions (Wang et al., 2022b), Flan (Wei et al., 2021), and some additional dialogue & code datasets. We choose 1,060 tasks, narrowing our focus following the criteria as follows:

- Generation tasks with long outputs are excluded since the total token length of appending \(r\) and \(y\) exceeds the maximum output token length (512 tokens) during training.
- Datasets that are not publicly available such as DeepMind Coding Contents and Dr Repair (Yasunaga and Liang, 2020) are excluded.
- Datasets where the input and output do not correspond to each other in the huggingface datasets (Lhoest et al., 2021) are excluded.
- When a dataset appears in common across different sources, we prioritize using the task from P3 first, followed by SNI, and Flan.

3 The 9 available datasets are QASC (Khot et al., 2020), AQvA (Amini et al., 2019), GSM8K (Cobbe et al., 2021), QED (Lamm et al., 2021), StrategyQA (Geva et al., 2021), SenseMaking (Wang et al., 2021), CREAK (Onoe et al., 2021), e-SNLI (Camburu et al., 2018), ECQA (Aggarwal et al., 2021).

4 Following Sanh et al. (2021), we use the notion of ‘task’ referring to each prompt applied to a dataset.
completion, coreference resolution, and word disambiguation, rationales generated by large LMs are very short and uninformative. We exclude these tasks to prevent negative transfer during multitask learning (Aribandi et al., 2021; Jang et al., 2023).

Creating Demonstrations for ICL We first create prompts to apply in-context learning (ICL) with large LMs for augmenting the instances in the selected tasks with rationales. Preparing demonstrations $D^t$ for each task $t$ is the most straightforward, but it becomes infeasible to prepare demonstrations for each task as the number of tasks gets larger. Instead, we assign each task $t$ to $T_k$, a family of tasks that shares a similar task format such as multiple choice QA, closed book QA, and dialogue generation. Each family of tasks share $D^{Tk}$, which consists of $6 \sim 8$ demonstrations. These $6 \sim 8$ demonstrations for each task group $T_k$ is manually created by 3 of the authors in this paper. Specifically, given 136 instances sampled from Flan Collection, two annotators are assigned to write a rationale, and the other third annotator conducts an A/B testing between the two options. We manually create $D^{Tk}$ across $k = 26$ task groups. We include the prompts for all of the different task groups in Appendix D.

**Rationale Augmentation** We use the OpenAI Codex4 to augment rationales. Formally, given $(X_i^t, y_i^t)$, the $i^{th}$ instance of a task $t$, the goal is to generate corresponding rationale $r_i^t$. Note that during preliminary experiments, we found that ordering the label in front of the rationale within the demonstration $D^{Tk}$ was crucial to generate good quality rationales. We conjecture this is because ordering the label in front of the rationale loosens the need for the large LM to solve the underlying task and only focus on generating a rationale. However, we also found that in some tasks such as arithmetic reasoning, large LMs fail to generate good-quality rationales. To mitigate this issue, we apply filtering to the augmented rationales. We provide the criteria used for the filtering phase and the filtered cases at Appendix B. Also, we include analysis of the diversity and quality of CoT COLLECTION compared to the existing 9 CoT tasks and human-authored rationales in Appendix A.

### 4 Experiments

For our main experiments, we use Flan-T5 (Chung et al., 2022) as our base model, and obtain CoT-T5 by CoT fine-tuning on the CoT COLLECTION. Formally, given $X_i^t$, the goal of CoT fine-tuning is to sequentially generate the rationale $r_i^t$ and answer $y_i^t$. To indicate that $r_i^t$ should be generated before $y_i^t$, the trigger phrase ‘Let’s think step by step’ is added during both training and evaluation. We mostly follow the details for training and evaluation from Chung et al. (2022), and provide additional details in Appendix C. In this section, we show how training on CoT COLLECTION enhances zero-shot generalization capabilities (Section 4.2) and few-shot adaptation capabilities (Section 4.3).

#### 4.1 Evaluation

We evaluate under two different evaluation methods: Direct Evaluation and CoT Evaluation. For Direct Evaluation on classification tasks, we follow previous works using verbalizers, choosing the option with the highest probability through comparison of logit values (Schick and Schütze, 2021; Jang et al., 2023).
Sanh et al., 2021; Ye et al., 2022; Jang et al., 2023), and measure the accuracy. For generation tasks, we directly compare the LM’s prediction with the answer and measure the EM score.

When evaluating with CoT Evaluation, smaller LMs including Flan-T5 often do not generate any rationales even with the trigger phrase ‘Let’s think step by step’. Therefore, we adopt a hard constraint of requiring the LM to generate $r^f_t$ with at least a minimum length of 8 tokens. In classification tasks, we divide into two steps where the LM first generates $r^f_t$, and then verbalizers are applied with an indicator phrase ‘[ANSWER]’ inserted between $r^f_t$ and the possible options. For generation tasks, we extract the output coming after the indicator phrase. Accuracy metric is used for classification tasks while EM metric is used for generation tasks.

4.2 Zero-shot Generalization

In this subsection, we show how training with CoT COLLECTION could effectively improve the LM’s ability to solve unseen tasks. We have three difference experimental set-ups, testing different aspects: Setup #1: training on the entire 1060 tasks in CoT COLLECTION and evaluating the reasoning capabilities of LMs with the Bigbench Hard (BBH) benchmark (Suzgun et al., 2022), Setup #2: training only on 163 tasks that T0 (Sanh et al., 2021) used for training (a subset of the CoT COLLECTION), and evaluating the linguistic capabilities of LMs with the P3 evaluation benchmark (Sanh et al., 2021), and Setup #3: training with a translated, subset version of CoT COLLECTION for each five different languages and evaluating how LMs could perform CoT reasoning in multilingual settings using the MGSM benchmark (Shi et al., 2022).

Setup #1: CoT Fine-tuning with 1060 CoT Tasks

We first perform experiments with our main model, CoT-T5, by training Flan-T5 on the entire CoT COLLECTION and evaluate on the BBH benchmark (Suzgun et al., 2022). In addition to evaluating Flan-T5, we compare the performances of different baselines such as (1) T5-LM (Raffel et al., 2020); the original base model of Flan-T5, (2) T0 (Sanh et al., 2021); an instruction-tuned LM trained with P3 instruction dataset, (3) Tk-Instruct (Wang et al., 2022b); an instruction-tuned LM trained with SNI instruction dataset, and (4) GPT-3 (Brown et al., 2020); a pre-trained LLM with 175B parameters. For ablation purposes, we also train T5-LM with CoT COLLECTION (denoted as ‘T5 + CoT FT’). Note that FLAN COLLECTION includes 15 million instances, hence ~8 times larger compared to our CoT COLLECTION.

The results on BBH benchmark are shown across Table 1 and Table 2. In Table 1, CoT-T5 (3B & 11B) achieves a +4.34% and +2.60% improvement over Flan-T5 (3B & 11B) with CoT Evaluation. Surprisingly, while CoT-T5-3B CoT performance improves +4.34% with the cost of 0.96% degradation in Direct Evaluation, CoT-T5-11B’s Direct Evaluation performance even improves, resulting in a +2.57% total average improvement. Since CoT COLLECTION only includes instances augmented with rationales, these results show that CoT fine-tuning could improve the LM’s capabilities regardless of the evaluation method. Also, T5-3B + CoT FT and T5-11B + CoT FT outperforms Flan-T5-3B and Flan-T5-11B by a +1.45% and +3.89% margin, respectively, when evaluated with CoT evaluation. Moreover, T5-3B + CoT FINE-TUNING outperforms ~4 times larger models such as T0-11B and Tk-Instruct-11B in both Direct and CoT Evaluation. The overall results indicate that (1) CoT fine-tuning on a diverse number of tasks enables smaller LMs to outperform larger LMs and (2) training with FLAN Collection and CoT Collection provides complementary improvements to LMs under different evaluation methods; CoT-T5 obtains good results across both evaluation methods by training on both datasets.

In Table 2, CoT-T5-11B obtains same or better results on 15 out of 23 tasks when evaluated with Direct evaluation, and 17 out of 23 tasks when evaluated with CoT Evaluation compared to Flan-

<table>
<thead>
<tr>
<th>Method</th>
<th>CoT</th>
<th>Direct</th>
<th>Total Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5-LM-3B</td>
<td>26.68</td>
<td>26.96</td>
<td>26.82</td>
</tr>
<tr>
<td>T0-3B</td>
<td>26.64</td>
<td>27.45</td>
<td>27.05</td>
</tr>
<tr>
<td>Tk-INSTRUCT-3B</td>
<td>29.86</td>
<td>29.90</td>
<td>29.88</td>
</tr>
<tr>
<td>Tk-INSTRUCT-11B</td>
<td>33.60</td>
<td>30.71</td>
<td>32.16</td>
</tr>
<tr>
<td>T0-11B</td>
<td>31.83</td>
<td>33.57</td>
<td>32.70</td>
</tr>
<tr>
<td>FLAN-T5-3B</td>
<td>34.02</td>
<td>37.14</td>
<td>35.60</td>
</tr>
<tr>
<td>GPT-3 (175B)</td>
<td>38.30</td>
<td>33.60</td>
<td>38.30</td>
</tr>
<tr>
<td>FLAN-T5-11B</td>
<td>38.57</td>
<td>40.99</td>
<td>39.78</td>
</tr>
<tr>
<td>T5-3B + CoT FT</td>
<td>37.95</td>
<td>35.52</td>
<td>36.74</td>
</tr>
<tr>
<td>CoT-T5-3B</td>
<td>38.40</td>
<td>36.18</td>
<td>37.29</td>
</tr>
<tr>
<td>T5-11B + CoT FT</td>
<td>40.02</td>
<td>38.76</td>
<td>39.54</td>
</tr>
<tr>
<td>CoT-T5-11B</td>
<td><strong>42.20</strong></td>
<td><strong>42.56</strong></td>
<td><strong>42.38</strong></td>
</tr>
</tbody>
</table>

Table 1: Evaluation performance on all the 27 unseen datasets from BBH benchmark, including generation tasks. All evaluations are held in a zero-shot setting. The best comparable performances are **bolded** and second best underlined.
Part of the page is not visible, but based on the existing text, it seems to be discussing the performance of different models on various tasks. Here is a structured representation of the visible content:

### Table 2: Evaluation performance on 11 unseen classification datasets from BBH benchmark.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CoT-T5-11B</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>FLAN-T5-11B</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>VICTUNA-13B</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>CHATGPT</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>CODEX</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>GPT-4</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>CoT</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
<tr>
<td>Direct</td>
<td>60.64</td>
<td>61.64</td>
<td>61.64</td>
</tr>
</tbody>
</table>

### Table 3: Evaluation performance on 11 different unseen P3 dataset (SANH ET AL., 2021) categorized into 4 task categories. We report the direct performance of the baselines since they were not CoT fine-tuned on instruction data. The best comparable performances are **bolded** and second best underlined among the open-sourced LMs.

### Setup #2: CoT Fine-tuning with 163 CoT Tasks (T0 Setup)

To examine whether the effect of CoT fine-tuning is dependent on large number of tasks and instances, we use the P3 training subset from the CoT COLLECTION consisted of 644K instances from 163 tasks, and apply CoT fine-tuning to T0 (3B) (SANH ET AL., 2021) and T5-LM (3B) (RAFFEL et al., 2020). Note that T0 is trained with 12M instances, hence ~18.63 times larger. Then, we evaluate on the P3 evaluation benchmark which consists of 11 different NLP datasets. In addition to the baselines from the previous section (T5-LM, T0, and GPT-3), we also include LMs that are trained on the same T0 setup for comparison such as, (1) RoE (JANG ET AL., 2023): a modular expert LM that retrieves different expert models depending on the unseen task, (2) KiC (PAN ET AL., 2022): a retrieval-augmented model that...
is instruction-tuned to retrieve knowledge from a KB memory, and (3) Flipped (Ye et al., 2022): an instruction-tuned model that is trained to generate the instruction in order to resolve the LM overfitting to the output label as baseline models.

The results are shown in Table 3. Surprisingly, T5-3B + CoT FT outperforms T0-3B by a +8.24% margin when evaluated with CoT Evaluation, while using ~18.63 times less instances. This supports that CoT fine-tuning is data efficient, being effective even with less number of instances and tasks. Moreover, T0-3B + CoT FT improves T0-3B by +8.65% on average accuracy. When compared with T0-11B with ~4 times more number of parameters, it achieves better performance at sentence completion, and word sense disambiguation (WSD) tasks, and obtains similar performances at natural language inference and coreference resolution tasks.

Setup #3: Multilingual Adaptation with CoT Fine-tuning In previous work, Shi et al. (2022) proposed MGSM, a multilingual reasoning benchmark composed of 10 different languages. In this subsection, we conduct a toy experiment to examine whether CoT fine-tuning could enable LMs to reason step-by-step in multilingual settings as well, using a subset of 5 languages (Korean, Russian, French, Chinese, Japanese) from MGSM.

In Table 4, current smaller LMs can be divided into three categories: (1) Flan-T5, a LM that is CoT fine-tuned with mostly English instruction data, (2) MT5 (Xue et al., 2021), a LM pretrained on diverse languages, but isn’t instruction tuned or CoT fine-tuned, (3) MT0 (Muennighoff et al., 2022), a LM that is instruction-tuned on diverse languages, but isn’t CoT fine-tuned. In relatively underrepresented languages such as Korean, Japanese, and Chinese, all three LMs get close to zero accuracy.

A natural question arises whether training a multilingual LM that could reason step-by-step on different languages is viable. As a preliminary research, we examine whether CoT Fine-tuning on a single language with a small amount of CoT data could enable LMs to avoid achieving near zero score such as Korean, Chinese and Japanese subsets of MGSM. Since there is no publicly available multilingual instruction dataset, we translate 60K ~ 80K instances from CoT COLLECTION for each 5 languages using ChatGPT (OpenAI, 2022), and CoT fine-tune mT5 and mT0 on each of them.

The results are shown in Table 4. Across all the 5 different languages, CoT fine-tuning brings about non-trivial gains in performance. Even for relatively low-resource languages such as Korean, Japanese, and Chinese, CoT fine-tuning on the specific language allows the underlying LM to perform mathematical reasoning in the target language, which are considered very difficult (Shi et al., 2022). Considering that only a very small number of instances were used for language-specific adaptation (60k-80k), CoT fine-tuning shows potential for efficient language adaptation.

However, it is noteworthy that we limited our setting to training/evaluating on a single target language, without exploring the cross-lingual transfer of CoT capabilities among varied languages. The chief objective of this experimentation was to ascertain if introducing a minimal volume of CoT data could facilitate effective adaptation to the target language, specifically when addressing reasoning challenges. Up to date, no hypothesis has suggested that training with CoT in various languages could enable cross-lingual transfer of CoT abilities among different languages. We identify this as a promising avenue for future exploration.

4.3 Few-shot Generalization

In this subsection, we show how CoT-T5 performs in a few-shot adaptation setting where a limited number of instances from the target task can be used for training, which is sometimes more likely in real-world scenarios.

Dataset Setup We choose 4 domain-specific datasets from legal and medical domains including LEDGAR (Tuggener et al., 2020), Case Hold (Zheng et al., 2021), MedNLI (Romanov and Shivade, 2018), and PubMedQA (Jin et al., 2019). To simulate a few-shot setting, we randomly sample 64 instances from the train split of each dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>ko</th>
<th>ru</th>
<th>fr</th>
<th>zh</th>
<th>ja</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAN-T5-3B</td>
<td>0.0</td>
<td>2.8</td>
<td>7.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>FLAN-T5-11B</td>
<td>0.0</td>
<td>5.2</td>
<td>13.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>MT5-3.7B</td>
<td>0.0</td>
<td>1.2</td>
<td>2.0</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>mT0-3.7B</td>
<td>0.0</td>
<td>4.8</td>
<td>7.2</td>
<td>1.6</td>
<td>2.4</td>
</tr>
<tr>
<td>GPT-3 (175B)</td>
<td>0.0</td>
<td>4.4</td>
<td>10.8</td>
<td>6.8</td>
<td>0.8</td>
</tr>
<tr>
<td>MT5-3.7B + CoT FT</td>
<td>3.2</td>
<td>6.8</td>
<td>9.6</td>
<td>6.0</td>
<td>7.6</td>
</tr>
<tr>
<td>mT0-3.7B + CoT FT</td>
<td>7.6</td>
<td>10.4</td>
<td>15.6</td>
<td>11.2</td>
<td>11.0</td>
</tr>
</tbody>
</table>

Table 4: Evaluation performance on MGSM benchmark (Shi et al., 2022) across 5 languages (Korean, Russian, French, Chinese, Japanese, respectively). All evaluations are held in a zero-shot setting with CoT Evaluation except GPT-3 using a 6-Shot prompt for ICL. The best comparable performances are bolded and second best underlined. Note that 'MT5-3.7B + CoT FT' and 'mT0-3.7B + CoT FT' are trained on a single language instead of multiple languages as mT5 and mT0.
Table 5: Evaluation performance on 4 domain-specific datasets. FT. denotes Fine-tuning, CoT FT. denotes CoT fine-tuning, and CoT PT. denotes CoT Prompting. The best comparable performances are **bolded** and second best *underlined*. For a few-shot adaptation, we use 64 randomly sampled instances from each dataset.

We report the average accuracy across 3 runs with different random seeds. We augment rationales for the 64 training instances using the procedure described in Section 3 for the rationale augmentation phase, utilizing the MCQA prompt from P3 dataset. In an applied setting, practitioners could obtain rationales written by human experts.

**Training Setup** We compare Flan-T5 & CoT-T5, across 3B and 11B scale and explore 4 different approaches for few-shot adaptation: (1) regular fine-tuning, (2) CoT fine-tuning, (3) LoRA fine-tuning, and (4) LoRA CoT fine-tuning. When applying Lora, we use a rank of 4 and train for 1K steps following Liu et al. (2022b). This results in training 2.35M parameters for 3B scale models and 4.72M parameters for 11B scale models. Also, we include Claude (Anthropic, 2023) and ChatGPT (OpenAI, 2022) as ICL baselines by appending demonstrations up to maximum context length\(^6\). Specifically, for CoT prompting, the demonstrations are sampled among 64 augmented rationales are used.

**Effect of LoRA** The experimental results are shown in Table 5. Overall, CoT fine-tuning CoT-T5 integrated with LoRA obtains the best results overall. Surprisingly for Flan-T5, applying full fine-tuning obtains better performance compared to its counterpart using LoRA fine-tuning. However, when using CoT-T5, LoRA achieves higher performance compared to full fine-tuning. We conjecture this to be the case because introducing only a few parameters enables CoT-T5 to maintain the CoT ability acquired during CoT fine-tuning.

**Fine-tuning vs. CoT Fine-tuning** While CoT fine-tuning obtains similar or lower performance compared to regular fine-tuning in Flan-T5, CoT-T5 achieves higher performance with CoT fine-tuning compared to Flan-T5 regular fine-tuning. This results in CoT-T5 in combination with CoT fine-tuning showing the best performance in few-shot adaptation setting.

**Fine-tuning vs. ICL** Lastly, fine-tuning methods obtain overall better results compared to ICL methods utilizing much larger, proprietary LLMs. We conjecture this to be the case due to the long input length of legal and medical datasets, making appending all available demonstrations (64) impossible. While increasing the context length could serve as a temporary solution, it would still mean that the inference time will increase quadratically in proportion to the input length, which makes ICL computationally expensive.

### Analysis of CoT Fine-tuning

In this section, we conduct experiments to address the following two research questions:

- For practitioners, is it more effective to augment CoT rationales across diverse tasks or more instances with a fixed number of tasks?
- During CoT fine-tuning, does the LM maintain its performance on in-domain tasks without any catastrophic forgetting?
5.1 Scaling the number of tasks & instances

In our main experiments, we used a large number of instances (1.84M) across a large number of tasks (1,060) to apply CoT fine-tuning. A natural question arises: “Is it more effective to increase the number of tasks or the number of instances?” To address this question, we conduct an experiment of randomly sampling a small number of instances within the CoT COLLECTION and comparing the BBH performance with (1) a baseline that is only CoT fine-tuned with the existing 9 CoT tasks and (2) CoT-T5 that fully utilizes all the 1.84M instances. Specifically, we sample 10K, 100K instances within the CoT COLLECTION and for the 9 CoT tasks, we fully use all the 180K instances. As CoT-T5, we use Flan-T5 as our base model and use the same training configuration and evaluation setting (CoT Eval) during our experiments.

The results are shown in Figure 3, where surprisingly, only using 10K instances across 1,060 tasks obtains better performance compared to using 180K instances across 9 tasks. This shows that maintaining a wide range of tasks is more crucial compared to increasing the number of instances.

5.2 In-domain Task Accuracy of CoT-T5

It is well known that LMs that are fine-tuned on a wide range of tasks suffer from catastrophic forgetting (Chen et al., 2020; Jang et al., 2021, 2023), a phenomenon where an LM improves its performance on newly learned tasks while the performance on previously learned tasks diminishes.

While CoT-T5 uses the same tasks as its base model (Flan-T5), we also check whether CoT fine-tuning on a wide range of tasks could possibly harm performance. For this purpose, we use the test set of 5 tasks within the CoT COLLECTION, namely ANLI-R1, ANLI-R2, ANLI-R3, RTE, and Winograd. Note that this differs with the Setup #2 in the main experiments in that we use different base models (T0 vs Flan-T5), and the tasks are already used for CoT fine-tuning.

Results are shown in Figure 4, where CoT-T5 consistently improves in-domain accuracy on the learned tasks as well. However, we conjecture that this is because we used the exact same task that Flan-T5 used to CoT fine-tuned CoT-T5. Adding additional tasks that were not used to train Flan-T5 and CoT-T5 could show different results, and we leave additional exploration of catastrophic forgetting during CoT fine-tuning to future work.

6 Conclusion

In this work, we show that augmenting rationales from an instruction tuning data using LLMs (Open AI Codex), and CoT fine-tuning could improve the reasoning capabilities of smaller LMs. Specifically, we construct CoT COLLECTION, a large-scale instruction-tuning dataset with 1.84M CoT rationales extracted across 1,060 NLP tasks. With our dataset, we CoT fine-tune Flan-T5 and obtain CoT-T5, which shows better zero-shot generalization performance and serves as a better base model when training with few number of instances. We hope CoT COLLECTION could be beneficial in the development of future strategies for advancing the capabilities of LMs with CoT fine-tuning.
Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00264, Comprehensive Video Understanding and Generation with Knowledge-based Deep Logic Neural Network, 50%; No.2021-0-02068, Artificial Intelligence Innovation Hub, 20%) and KAIST-NAVER Hypercreative AI Center.

Limitations

Recently, there has been a lot of focus on distilling the ability to engage in dialogues with long-form outputs in the context of instruction following (Taori et al., 2023; Chiang et al., 2023). Since our model CoT-T5 is not trained to engage in dialogues with long-form responses from LLMs, it does not necessarily possess the ability to be applied in chat applications. In contrast, our work focuses on improving the zero-shot and few-shot capabilities by training on academic benchmarks (CoT COLLECTION, Flan Collection), where LMs trained with chat data lack on. Utilizing both long-form chat data from LLMs along with instruction data from academic tasks has been addressed in future work (Wang et al., 2023). Moreover, various applications have been introduced by using the FEEDBACK COLLECTION to train advanced chat models 7.

Also, since CoT-T5 uses Flan-T5 as a base model, it doesn’t have the ability to perform step-by-step reasoning in diverse languages. Exploring how to efficiently and effectively train on CoT data from multiple languages is also a promising and important line of future work. While Shi et al. (2022) has shown that large LMs with more than 100B parameters have the ability to write CoT in different languages, our results show that smaller LMs show nearly zero accuracy when solving math problems in different languages. While CoT fine-tuning somehow shows slight improvement, a more comprehensive strategy of integrating the ability to write CoT in diverse language would hold crucial.

In terms of reproducibility, it is extremely concerning that proprietary LLMs shut down such as the example of the Codex, the LLM we used for rationale augmentation. We provide additional analysis on how different LLMs could be used for this process in Appendix A. Also, there is room of improvement regarding the quality of our dataset by using more powerful LLMs such as GPT-4 and better prompting techniques such as Tree of Thoughts (ToT) (Yao et al., 2023). This was examined by later work in Mukherjee et al. (2023) which used GPT-4 to augment 5 million rationales and Yue et al. (2023) which mixed Chain-of-Thoughts and Program of Thoughts (PoT) during fine-tuning. Using rationales extracted using Tree of Thoughts (Yao et al., 2023) could also be explored in future work.

References


A Analysis of CoT COLLECTION

Non-cherry picked rationales within CoT COLLECTION are shown in Table 8. We perform an analysis regarding the quality, diversity, and reproducibility of rationale within the CoT COLLECTION.

Diversity of Rationales To take a look into the diversity of CoT COLLECTION, we use Berkeley Neural Parser (Kitaev and Klein, 2018; Kitaev et al., 2019) and parse rationales. More specifically, the verb which is closest to the root of the parse tree along the noun object is extracted. We compare this with the rationales from the 9 CoT datasets used in Chung et al. (2022). As shown in Figure 5, CoT COLLECTION have diverse textual formats included compared to the 9 existing CoT datasets that have a high proportion assigned to ‘answer question’ and ‘consider following’.

Quality of Rationales To ensure the quality of CoT COLLECTION, we use ROSCOE (Golovneva et al., 2022), a suite of metrics designed to evaluate rationales under different criteria within semantic alignment, semantic similarity, logical inference, language coherence. We compare with human-authored rationales obtained during Prompt Creation in Section 3. The 13 ROSCOE scores are shown in Table 6. The results show that CoT COLLECTION include CoT rationales that are faithful, less repetitive, informative, and logical even when compared to human-authored rationales. Yet, we find that machine-generated rationales tend to have higher perplexity, leading to lower language coherence scores. We conjecture this is because including diverse textual formats leads may result in relatively higher perplexity (Holtzman et al., 2019).

Is CoT COLLECTION Reproducible? One could doubt whether CoT COLLECTION is reproducible due to the usage of OpenAI model in the process of CoT rationale augmentation. In this section, we test different LLMs to generate 150 rationales randomly sampled from CoT COLLECTION, and compare the ROSCOE score (Golovneva et al., 2022) in order to assess the quality. We use Bard (Google, 2023), Claude (Anthropic, 2023), for comparing with OpenAI Codex. The comparison of quality is shown in Figure 6. The results show that different LLMs are able to produce high quality rationales in terms of semantic alignment and language coherence.

B Filtering CoT COLLECTION

Filtering After generating multiple rationales, we filter to ensure high quality. We apply the following criteria to filter instances:

- We exclude rationales that do not include the ground truth answer when splitted by white spaces. While a rationale that doesn’t include the answer isn’t necessarily a bad rationale, we found it is effective to exclude inconsistent ones.
- We exclude CoT rationales that exceed the maximum output length, where we constrain the sum of \( r \) and \( y \) to be shorter than 512 tokens.
- We exclude rationales that are identical to previously augmented ones during our process.
- We exclude rationales that include repetitive sentences within the context.

Table 6: Comparison of the quality between human-authored rationales and machine-generated rationales. 13 label-free metrics from ROSCOE (Golovneva et al., 2022) is used.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Human</th>
<th>CoT Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>faithfulness</td>
<td>0.8836</td>
<td>0.8914</td>
</tr>
<tr>
<td>Semantic faithfulness</td>
<td>0.8756</td>
<td>0.8793</td>
</tr>
<tr>
<td>Alignment</td>
<td>0.9376</td>
<td>0.9419</td>
</tr>
<tr>
<td>informativeness_step</td>
<td>0.9519</td>
<td>0.9521</td>
</tr>
<tr>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>discourse_representation</td>
<td>0.4835</td>
<td>0.2797</td>
</tr>
<tr>
<td>Similarity</td>
<td>0.2453</td>
<td>0.2910</td>
</tr>
<tr>
<td>perplexity_step</td>
<td>0.0198</td>
<td>0.0012</td>
</tr>
<tr>
<td>perplexity_chain</td>
<td>0.0475</td>
<td>0.0255</td>
</tr>
<tr>
<td>repetition_word</td>
<td>0.0144</td>
<td>0.0088</td>
</tr>
<tr>
<td>repetition_sent</td>
<td>0.8883</td>
<td>0.8721</td>
</tr>
<tr>
<td>grammar_step</td>
<td>0.8013</td>
<td>0.7724</td>
</tr>
<tr>
<td>grammar_step_max</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9. Metrics Human CoT Collection

Moreover, OpenAI announced to stop its support on Codex model starting from June, 2023.
C Training and Evaluation Details of CoT-T5

<table>
<thead>
<tr>
<th>Params</th>
<th>Model</th>
<th>Batch size</th>
<th>LR</th>
<th>Optimizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>3B</td>
<td>CoT-T5-3B</td>
<td>64</td>
<td>5e-5</td>
<td>AdamW</td>
</tr>
<tr>
<td>11B</td>
<td>CoT-T5-11B</td>
<td>8</td>
<td>1e-4</td>
<td>Adafactor</td>
</tr>
</tbody>
</table>

| Table 7: Hyperparameters used for fine-tuning CoT-T5. |

We mostly follow the fine-tuning details of Chung et al. (2022) to train CoT-T5. The hyperparameters used for training CoT-T5 are shown in Table 7. We find 3B and 11B sized LMs converge well using different optimizers. While CoT-T5-3B tends to converge well using AdamW, CoT-T5-11B is well optimized using Adafactor. For both sizes, we train with 1 epoch using CoT COLLECTION which takes 1 day (3B) and 7 days (11B) when 8 A 100 (80GB) GPUs are used. For both settings, we use a gradient accumulation step of 8.

For sampling training instances, we sample instances from Flan Collection (Longpre et al., 2023) by using the proportion of 23.94%(FLAN), 30.85%(P3), 7.89%(Existing 9 CoT datasets), 25.47%(SNI) and 11.85%(other dialogue & code datasets). This is done by sampling 400 instances (FLAN), 300 instances (P3), 150 instances (SNI), 4000 instances (Existing 9 CoT datasets), and 300 instances (other dialogue & code datasets), respectively. We generate 5 rationales per instance and then apply filtering, leading to the final set of CoT COLLECTION, which is consisted of 1.84 million instances and rationales across 1,060 tasks.

During evaluation, we found that using nucleus sampling (Holtzman et al., 2019) with p = 0.8 and no_repeat_n_gram = 3 was very effective in generating good-quality rationales.
Figure 5: The top 20 common root verbs (inner circle) and their top 4 noun objects (outer circle) within the rationales of the 9 CoT tasks used in Chung et al. (2022) (left side) and our CoT COLLECTION with 1,060 tasks (right side).

Figure 6: Comparison of the quality of augmenting CoT rationales using different LLMs by reference-free ROSCOE scores. Semantic Alignment(SA), Semantic Similarity(SS), Logical Inference(LI), Coherence(LC).
Read the following context and choose the best option to answer the question.

Context: From the triple loop she executed after five minutes, he suspected that she had managed to sneak off for flying more often than during the summers and wondered just how diligent George was about practicing on his own time.

Question: What is a triple loop?

Options:
A. None of the above choices
B. 3 round trips on a plane route
C. 3 neckties worn on flights
D. A feature pulled off by making 3 loops in the air.

Answer: D

Rationale: A triple loop is a maneuver in which an aircraft makes three loops. Based on the context, it can be inferred that a triple loop involves making 3 loops in the air. Since this is what A means, D should be chosen as the correct answer.

Read the passage below and choose the right answer to the following question (choices are more or less):

On a hot day, an endotherm may give off more heat by increasing blood flow to the surface of the body. After Sarah finishes her walk, she notices that her blood is not really pumping hard. Sarah can know that since it’s a cool day, she’s giving off what kind of heat compared to usual more or less?

Answer: less

Rationale: The given passage states that an endotherm gives off more heat by increasing blood flow to the surface of its body. Since Sarah notices that her blood is not really pumping that hard due to cooler weather, it can be inferred that she’s giving off less heat compared to usual.

Pubertal development also affects circulatory and respiratory systems as an adolescent’s heart and lungs increase in both size and capacity. These changes lead to increased strength and tolerance for exercise. Sex differences are apparent as males tend to develop larger hearts and lungs; higher systolic blood pressure, a lower resting heart rate, a greater capacity for carrying oxygen to the blood, a greater power for neutralizing the chemical products of muscular exercise, higher blood hemoglobin and more red blood cells.

Which sex tends to develop larger hearts and lungs?

Answer: males

Rationale: The last sentence of the paragraph states that males tend to develop larger hearts and lungs.

In this task, you are given a string of characters. You need to remove duplicate characters from the string if any, and return the resulting string.

Input: nXnGFzHFYudOiHyWLZXQfpw

Output: nXGFzHYudOiyWLZQfpw

Rationale: There are no duplicate characters in the given string, so the resulting string will be the same as the input.

In this task you will be given a list of numbers and you should remove all duplicates in the list. If every number is repeated in the list an empty list should be returned. Your list should be numbers inside brackets, just like the given list.

Input: {3, 5, 1, 0, 3}

Output: {5, 1, 0}

Rationale: The given list is {3, 5, 1, 0, 3}. We can see that ‘3’ appears in the list twice. Therefore, we need to keep only one of the ‘3’s and remove the other one. The resulting list is {5, 1, 0}.

How to keep your phone cool in the car

Keep your phone on a vent mount with air conditioning on if you need to use it.

If you need to use your phone while you’re in your car, purchase a mount that attaches to one of the dashboard air vents. Turn on the air conditioner so it keeps your phone cold.

Can we draw the following conclusion?

keep your phone and tablet cool while mounted on an air conditioner vent if you need to use them while in the car.

Options:
- Yes
- It’s impossible to say
- No

Answer: It’s impossible to say

Rationale: While the passage does mention that your phone should be mounted on an air vent, it doesn’t specify whether this is referring to a car or home AC. Additionally, we don’t know from the context if other devices such as tablets can also be kept cool using this method. So, "It’s impossible to say."

Give me a possible correct answer to the question "who nominated stephen g. breyer?"

Bill Clinton

Rationale: Stephen G. Breyer was a nominee for the Supreme Court by Bill Clinton and was confirmed on July 29, 1994. Although he wasn’t nominated in 1998 or 1999, this is still a valid answer because it happened within 5 years of those dates before.

Write down the solution for this math problem:

Solve 291c = 264c + 189 for c.

Answer: 7

Rationale: 291c – 264c = 189
STEP 2: 27c = 189
STEP 3: c = 7

Table 8: Example of rationales within COT COLLECTION including tasks such as Multiple choice QA (MCQA), Extractive QA (ExQA), Closed-book QA (CBQA), Formal Logic, Natural Language Inference (NLI), and Arithmetic.
You are given a math word problem and you are supposed to only use subtraction on the numbers embedded in the text to answer the following question and then only report the final numerical answer.

Context: Jake has 8 fewer peaches and 10 more apples than Steven. Steven has 11 apples and 18 peaches.

Question: How many more peaches than apples does Steven have?

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>CoT Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>You are given a math word problem and you are supposed to only use subtraction on the numbers embedded in the text to answer the following question and then only report the final numerical answer.</td>
<td>7</td>
<td>To solve this problem, we need to find the number of apples and peaches that Steven has. From the text, we know that Jake has 8 fewer peaches than Steven and 10 more apples than Steven. We also know that Steven has 11 apples and 18 peaches. Using these clues, we can write two equations: Steven’s peaches = Jake’s Peaches + 8 Steven’s Apples = Jake’s Apples - 10 We can then add these two equations together: (Steven’s Peaches + Steven’s Apples) = (Jake’s Peaches + 8) + (Jake’s Apples - 10) (18 + 11) = (Jake’s Peaches + 8) + (Jake’s Apples - 10) 29 = Jake’s peaches + Jake’s apples - 2 a = 29 - 2 = 27 b = 18 + 11 - a c = b / 2 d = c ^ 2</td>
</tr>
<tr>
<td>What is the solution? Solve 690c + 14495 = -4825 for c.</td>
<td>-28</td>
<td>690c + 14495 = -4825 Subtracting 4825 from both sides: 690c + 9669 = 0 Dividing both sides by 690: c = 14.02391304348</td>
</tr>
</tbody>
</table>

Table 9: Two representative rationale examples that are filtered with our filtering criteria. It demonstrates that not including the ground truth token within the rationale could indicate that the LLM failed to generate a consistent rationale, often leading to degeneration. This phenomenon is mostly found on math datasets.
D Prompts used for Rationale Augmentation

Among the 26 demonstrations used for each family of tasks when augmenting CoT rationales, we include 4 examples, other demonstrations for task families can be seen at this link.

---

### Example 1
Instruction and Question
Skylar had stopped breathing but Lee held Skylar’s arms away, in vain. Given the context, Lee wants to do what next?

Possible answers: beg the doctors to try again, get her up and walking home, bring Skylar back

Answer: beg the doctors to try again

Rationale
The context of the situation is that Skylar has stopped breathing and Lee is holding Skylar’s arms. This suggests that Skylar has received medical attention and is no longer alive.

So the answer is to beg the doctors to try again.

---

### Example 2
Instruction and Question
Do you think the right answer to the question “what can run on alcoholic fermentation of the glucose in corn or other plants?” is “bicycles”?

Options
- No
- Yes

Answer: No

Rationale
Alcoholic fermentation is a process that occurs in living organisms, most commonly yeast, to convert glucose into energy. This process does not produce enough energy to power something like a bicycle.

So the answer is No.

---

### Example 3
Instruction and Question
Eventually the cyst can get so large that part of the wall weakens and fluid begins to leak out, that is the cyst ruptures. It is similar to popping a water balloon. When this happens, the fluid may irritate the pelvic lining and cause pain, often severe. Most of the time the pain begins to gradually subside over the course of the next few days.

What type of doctor might the person have had to see?

- A paediatrician
- A gynaecologist
- None of the above choices

Answer: None of the above choices

Rationale
The context mentions a cyst in the pelvic region and the symptoms of the cyst rupturing. This suggests that the issue is related to the reproductive or urinary system. Given that the issue is related to the reproductive or urinary system, it is likely that the person would have had to see a specialist in those areas.

So the answer is that the person would have had to see a gynaecologist, as gynaecologists specialize in female reproductive health, including issues related to the uterus, ovaries, and other reproductive organs.

---

### Example 4
Instruction and Question
Process:
- Get some seeds
- Pick a spot to plant them
- Dig a hole in the dirt
- Put the seed in the hole
- Pour some water on the seed and hole
- Cover up the hole
- Press down on it
- Spray some plant food on it

Question: suppose if no water is available, how will it affect vegetables?

Options
- A: more
- B: less

Answer: B

Rationale
The process of planting seeds requires water, which is essential for the growth of plants. Water helps transport nutrients, regulate temperature, and absorb minerals from the soil. If no water is available, the seed will not germinate and the plant will not grow properly, leading to fewer vegetables. Hence, the absence of water will result in “less” vegetables, making the answer “B.”

---

### Example 5
Instruction and Question
Read the following conversation and answer the question.

M: I’m worried about my sister, Lucy. W: Why is that? M: She thinks she’s too fat. W: Is she? M: No, but she keeps skipping meals. Then, she only eats chips and drinks cola. W: I used to do that. It’s called binge. It was no fun! M: Why did you stop doing it? W: Well, my doctor told me to eat when I’m hungry. She said, “Eat till you’re full or you’ll eat too much later.” She said a lot of girls ruin their health this way. M: Did she say what to eat? W: She said, “Eat fruit, vegetables, meats and grains. Have regular meals and snacks. Get exercise, too.” Question: What’s the relationship between the speakers?

Options
- Doctor and patient.
- Friends.
- Brother and sister.

Answer: Friends

Rationale
The conversation between two people, a man(M) and a woman(W), is discussing the man’s sister who has a negative view of her own body and is skipping meals. The woman shares that she used to engage in similar behavior, and explains that her doctor advised her to eat a balanced diet and exercise. It is likely that the relationship between the speakers is that of friends.

Therefore, the answer is Friends.

---

### Example 6
Instruction and Question
The context of the situation is that Skylar has stopped breathing and Lee is holding Skylar’s arms. This suggests that Skylar has received medical attention and is no longer alive.

Out of the given options, “a gynaecologist” is the most likely specialist that the person would have had to see, as gynaecologists specialize in female reproductive health, including issues related to the uterus, ovaries, and other reproductive organs.

In conclusion, based on the given context, the most likely answer is a gynaecologist.

---
Prompts 1: Demonstration used for tasks in family of MCQA (T0)
W: I used to do that. It’s called binge. It was no fun!
M: Why did you stop doing it?
W: Well, my doctor told me to eat … financial woes in the
1970s sent the chain on a decline and most of its restaurants out of business by the 1980s.

[Instruction and Question]
Q: What’s the relationship between the speakers?

[Options]
- Doctor and patient
- Friends
- Brother and sister

[Answer]
Friends

[Rationale]
The conversation between two people, a man M and a woman W, is discussing the man’s sister who has a negative view of her own body and is skipping meals. The woman shares that she used to engage in similar behavior, and explains that her doctor advised her to eat a balanced diet and exercise.

Among the three options, 
- First, since the woman is sharing her experience of what the doctor told her, it is unlikely she would be a medical professional. So the answer couldn’t be doctor and patient.
- Second, given that the man is talking about his sister, the woman is unlikely to be the man’s sister.
- Lastly, because the conversation is informal and the speakers are discussing a personal matter it is likely that the relationship between the speakers is that of friends.

Therefore, the answer is Friends.

[Example 6]

[Instruction and Question]
Question: The best place for the troublemaker to sneak a smoke while he was in a yard, was in the bathroom? 

[Options]
- playground  
- school  
- at hotel
- theater
- convenience store

[Answer]
convenience store 

[Example 7]

[Instruction and Question]
So I was looking at quotes for our tshirts for ice hockey, and I came across one that sums up how I’m feeling right now. When you talk about an injury and the kind of depression you go through, it’s not just because you’re out of shape and you can’t go out and play. You’re missing a part of you.

Why is the narrator so down in the dumps? Pick the best answer from the following options: 
A. They are unable to play ice crouching. 
B. They are hurt. 
C. They are unable to play ice skating. 
D. They are unable to play in the yard.

[Options]
A
B
C
D

[Answer] B

[Rationale]
The quote in the question mentions the narrator’s feelings about an injury and the depression that follows. The narrator says that it’s not just about being out of shape and unable to play, but rather they feel like they are missing a part of themselves. This implies that the reason for their depression is linked to their physical inability to participate in a sport or activity they enjoy. Based on the context, it can be inferred that the narrator is upset because they are hurt and unable to play ice hockey.

So, the answer is B.

[Example 8]

[Instruction and Question]
Q: What does the ubiquitous generator use to generate currents?

[Options]
- inertia – water – gravity – magnetism

[Answer] magnetism

[Rationale]
A generator is a device that converts mechanical energy into electrical energy. This is done by using a magnetic field to induce a current in the generator’s windings. This current is then sent to an external circuit, where it can be used to power electrical devices. The generator uses magnetism to generate currents, so the answer is magnetism.

Prompts 2: Demonstration used for tasks in family of Commonsense (SNI)
Today, two Biff Burger restaurants remain. Biff Burger came about in Clearwater Florida. OPTIONS: Yes– No

[Answer] Yes

[Rationale] To validate the claim "Biff Burger came about in Clearwater Florida," we should find a supporting evidence within the context. We could see that "Biff Burger (Best In Fast Food)" was a burger fast food franchise that originated in Clearwater, Florida, supports the claim we are trying to validate. So the answer is "Yes".

[Example 5] [Instruction and Question] 4. There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room?

[Answer] 29

[Rationale] There were 9 computers in the beginning, so 5 * 4 = 20 computers were added. That means in total 4 * 5 = 20 computers were added. There were 9 computers in the beginning, so

Prompts 3: Demonstration used for tasks in family of NLI (FLAN)

[Example 1] [Instruction and Question] Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

[Answer] 8

[Rationale] Jason had 20 lollipops. Since he only has 12 now, he must have given the rest to Denny. If they ate 35, how many pieces do they have left in total?

[Example 2] [Instruction and Question] Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

[Answer] 39

[Rationale] Leah had 32 chocolates and Leah’s sister had 42. If they ate 35, how many pieces do they have left in total? There were originally 32 + 42 = 74 chocolates. 35 have been eaten. So in total they still have 74 − 35 = 39 chocolates.

[Example 3] [Instruction and Question] Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

[Answer] 8

[Rationale] Jason had 20 lollipops. Since he only has 12 now, he must have given the rest to Denny. The number of lollipops he has given to Denny must have been 20 − 12 = 8 lollipops.

[Example 4] [Instruction and Question] Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

[Answer] 8

[Rationale] Jason had 20 lollipops. Since he only has 12 now, he must have given the rest to Denny. The number of lollipops he has given to Denny must have been 20 − 12 = 8 lollipops.

[Example 5] [Instruction and Question] Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

[Answer] 8

[Rationale] Jason had 20 lollipops. Since he only has 12 now, he must have given the rest to Denny. The number of lollipops he has given to Denny must have been 20 − 12 = 8 lollipops.

[Example 6] [Instruction and Question] Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

[Answer] 8

[Rationale] Jason had 20 lollipops. Since he only has 12 now, he must have given the rest to Denny. The number of lollipops he has given to Denny must have been 20 − 12 = 8 lollipops.
now there are 9 + 20 = 29 computers.

[Example 7]  
[Instruction and Question]  
Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?

[Answer]  
33

[Rationale]  
Michael initially had 58 balls. He lost 23 on Tuesday, so after that he has $58 - 23 = 35$ balls. On Wednesday he lost 2 more so now he has $35 - 2 = 33$ balls.

[Example 8]  
[Instruction and Question]  
Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

[Answer]  
8

[Rationale]  
She bought 5 bagels for $3 each. This means she spent $5 \times $3 = $15$ on the bagels. She had $23 in beginning, so now she has $23 - $15 = $8$.

[Example 9]  
[Instruction and Question]  
Prompts 4: Demonstration used for tasks in family of Arithmetic (FLAN)