arXiv:2210.06726v1 [cs.CL] 13 Oct 2022

Explanations from Large Language Models Make Small Reasoners Better

Shiyang Li[1], **Jianshu Chen**[2], **Yelong Shen**[3], **Zhiyu Chen**[1], **Xinlu Zhang**[1] , **Zekun Li**[1]
Hong Wang[1], **Jing Qian**[1], **Baolin Peng**[3], **Yi Mao**[3], **Wenhu Chen**[4] and **Xifeng Yan**[1]
[1]University of California, Santa Barbara
[2]Tencent AI Lab, [3]Microsoft
[4]University of Waterloo, Vector Institute
`{shiyangli,zhiyuchen,xinluzhang,zekunli,hongwang600,jing_qian,xyan}@cs.ucsb.edu`
`jianshuchen@tencent.com`, `wenhuchen@uwaterloo.ca`
`{yelong.shen,bapeng,maoyi}@microsoft.com`

Abstract

Integrating free-text explanations to in-context learning of large language models (LLM) is shown to elicit strong reasoning capabilities along with reasonable explanations. In this paper, we consider the problem of leveraging the explanations generated by LLM to improve the training of small reasoners, which are more favorable in real-production deployment due to their low cost. We systematically explore three explanation generation approaches from LLM and utilize a multi-task learning framework to facilitate small models to acquire strong reasoning power together with explanation generation capabilities. Experiments on multiple reasoning tasks show that our method can consistently and significantly outperform finetuning baselines across different settings, and even perform better than finetuning/prompting a 60x larger GPT-3 (175B) model by up to 9.5% in accuracy. As a side benefit, human evaluation further shows that our method can generate high-quality explanations to justify its predictions, moving towards the goal of explainable AI.

1 Introduction

Large language models (LLM) have achieved impressive results with in-context learning; by adding a few demonstrations as the prompts, they can solve unseen tasks without any parameter update (Brown et al., 2020; Thoppilan et al., 2022; Chowdhery et al., 2022; Wei et al., 2022a). Recently, it is shown that adding explanation-augmented prompts can elicit strong performance in various reasoning tasks (Wei et al., 2022b; Lampinen et al., 2022), such as math word problem (Cobbe et al., 2021), symbolic reasoning (Wei et al., 2022b), numerical reasoning (Zhou et al., 2022) and commonsense reasoning tasks (Talmor et al., 2019). In addition, they also enable LLM to generate reasonable explanations to justify the reasoning outcomes.

In this paper, we consider the problem of leveraging these elicited explanations by LLM to improve the training of small reasoners. Small language models (SLM)[1] could be more favorable over LLM in many real situations due to their low cost in both storage and computation. Nevertheless, one important open question is how to close the performance gap with respect to LLM on complicated reasoning tasks, as is observed in Zelikman et al. (2022), especially in few-shot settings (Li et al., 2019). Surprisingly, Hase et al. (2020) shows that using human-annotated explanations does not improve the performance compared to standard finetuning on T5 (Raffel et al., 2019). One possible reason is that many human-annotated explanations collected via crowdsourcing (Wiegreffe and Marasović, 2021) could be logically inconsistent and grammatically incorrect (Narang et al., 2020), which restricts the amount of available high-quality explanations. On the other hand, using explanation-augmented prompts enables LLM to automatically generate descent explanations (Wiegreffe et al., 2021a), making it a plausible alternative to generate arbitrary amount of explanations. Therefore, a key question is: *Can the explanations generated by LLM improve the reasoning capability of SLM?*

In this paper, we show that explanations generated from LLM can consistently improve reasoning capability of SLM. Our framework is shown in Figure 1. Specifically, we first utilize several examples with human-written explanations as demonstrations of LLM and then generate explanations for *training* set. We systematically explore three approaches to generating explanations. The first approach utilizes explanations generated through chain of thought prompting and explanations are adopted if LLM have correct predictions and are rejected otherwise (Zelikman et al., 2022). The second one is to generate explanations by rationalization prompting conditioned on golden labels (Wiegreffe et al., 2021a).

---

[1] We argue that small and large models are relative concepts. For the same model, it can be small or large depending on the context.

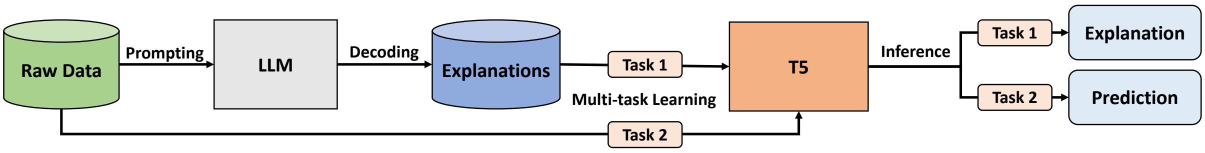


Figure 1: Overview of proposed framework.

Intuitively, the first approach may generate higher quality explanations than the second if LLM' predictions are correct as incorrect explanations tend to generate incorrect predictions (Wei et al., 2022b). However, the first approach will reject explanations on problems with incorrect predictions, leaving their explanations empty. On the other hand, the second one explicitly conditions on golden labels and may still generate useful explanations on problems where chain of thought prompting cannot predict correctly. Therefore, we propose a third hybrid approach: adopting explanations generated by chain of thought prompting if LLM have correct predictions and use rationalization prompting otherwise. As we will show in section 5, three explanation generation methods can consistently and significantly improve fine-tuning baselines without explanations and our hybrid approach achieves best results on two of three datasets.

We further adopt a multi-task (MT) learning framework shown in Figure 2 to utilize the LLM-generated explanations since (1) it can naturally allow training with partially generated explanations and (2) self-rationalizing model (Wiegreffe et al., 2021b), where golden label and the human-written explanation is linearly concatenated as the target, performs significantly worse than MT counterpart(Hase et al., 2020). Interestingly, we find that even with the same MT approaches (i.e., MT-Re (Hase et al., 2020) and MT-Ra (Camburu et al., 2018)) as Hase et al. (2020), we can consistently and significantly improve strong T5 standard finetuning baseline using LLM-generated explanations, which is in stark contrast to the results in Hase et al. (2020), where finetuning T5 with MT-Re and MT-Ra only achieves on par results using crowd-sourced ones. In addition, we further propose MT-CoT, where the small language model is trained to jointly solve two tasks: (i) directly generating the answer and (ii) generating an explanation and then the answer, as shown in Figure 2 (c). Unlike MT-Re and MT-Ra, MT-CoT positions the answer after the explanation, hoping the model can learn to derive it from the explanation like chain of thought (Wei et al., 2022b). Our results show that all three explanation generation approaches can improve reasoning capability of small language models with MT-Ra, MT-Re and MT-CoT setups. And MT-CoT can achieve the best results over MT-Re and MT-Ra on two of three datasets. In addition, our method can outperform standard finetuning baseline by up to 8.1% in accuracy and even perform better than finetuning/prompting a 60x larger GPT-3 model (175B) by up to 9.5% in accuracy on *CommonsenseQA*. Finally, as a side benefit, human evaluation further shows that our method can generate high-quality explanations to justify its predictions, moving towards the goal of explainable AI (Samek et al., 2019).

In a nutshell, we summarize our contribution as following:

- We show that multi-task learning with explanations from LLM can consistently and significantly improve strong T5 single-task fine-tuning baselines across various settings.
- We propose a hybrid prompting approach to generating explanations from LLM and MT-CoT to further improve our learning with explanations from LLM paradiam.
- We demonstrate that our method can perform better than finetuning/prompting a 60x larger GPT-3 model (175B) by up to 9.5% in accuracy on *CommonsenseQA* and generate high-quality explanations to justify its predictions towards the goal of explainable AI.

2 Related Work

Prompting with Explanations. Recently, a new learning paradigm, *in-context learning* where several training examples are used as demonstrations of LLM without any parameter update, has shown promising results in various NLP tasks (Brown et al., 2020). Although promising, LLM still struggle with tasks requiring strong reasoning capability (Wei et al., 2022b). To enable better few-shot in-context learning of LLM for reasoning tasks, Wei et al. (2022b) proposes chain of thought prompting, which provides intermediate reasoning steps as explanations in prompts before answers and has achieved state-of-the-art in arithmetic, symbolic and common sense reasoning tasks. Zhou et al. (2022) further extends chain of thought prompting with least-to-most prompting, which decomposes a complex problem into a list of subproblems with natural languages, and then sequentially solves these subproblems in a recursive fashion. Kojima et al. (2022) moves one step further and shows that LLM are zero-shot reasoners by simply adding "*Let's think step by step*" without any demonstration in prompts. Unlike these work, Lampinen et al. (2022) explores explanations after answers prompting for LLM, where answers are fed into LLM before providing their explanations in prompts, and also observes consistent gains.

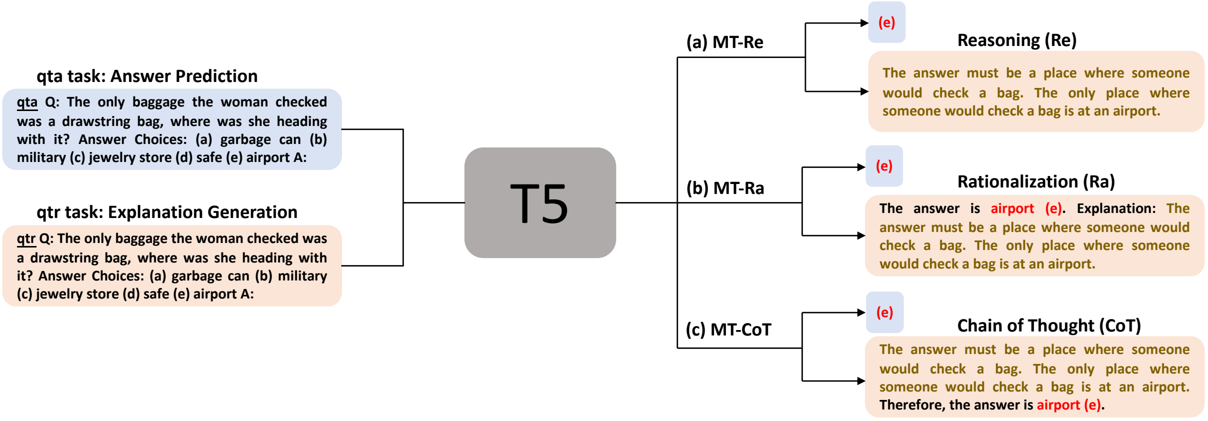
These also exist work to utilize explanations generated from LLM rather than focusing on their final predictions. Wiegreffe et al. (2021a) explores utilizing LLM to annotate explanations for existing datasets and proposes a sample-then-filter paradigm with human annotations. Ye and Durrett (2022) proposes to utilize a calibrator to calibrate GPT-3 as they find that GPT-3 tends to generate consistent but less factual explanations for textual reasoning tasks. However, none of these work explores if these noisy explanations generated from LLM without human-involved filtering can be used to improve SLM reasoning capability. The closest work to ours is STaR (Zelikman et al., 2022). STaR begins with prompting a descent large language model GPT-J with 6B parameters (Wang, 2021) possibly including answer hints via chain of thought prompting to generate explanations with incorrect answer rejection. After that, they utilize filtered training datasets with explanations to finetune GPT-J as a teacher model and then utilize the teacher model to generate explanations of training datasets to train a student GPT-J model iteratively with a self-training fashion until performance plateaus. However, STaR often requires dozens of iterations to converge, which is both time-consuming and compute-intensive to train a large 6B model. What's worse, their method may not be applicable to smaller language models, e.g. GPT-2 (Radford et al., 2019) and strong non-autoregressive models, e.g. T5, as they may not generate high-quality explanations with prompting. In addition, they only focus on chain of thought style prompting and finetuning while our approach can improve SLM across model sizes, explanation generation and multi-task finetuning methods.

Learning with Explanations. Learning with explanations has been commonly studied in robotics (Johnson, 1994) and computer vision (Hendricks et al., 2016). Recently, it has received increasing attention in NLP as well. Camburu et al. (2018) proposes MT-Ra for natural language inference task with LSTM and does not observe gains over single-task finetuning. Narang et al. (2020) utilizes MT-Ra setup on both T5-base and T5-11B models but mainly focuses on explanation generation. Instead, Rajani et al. (2019) observes improvements with two-stage finetuning using human-annotated explanations for common sense reasoning task, where the first stage is to train a model for explanation generations with GPT (Radford et al., 2018) and the second one utilizes explanations as input to train a classification model based on BERT (Devlin et al., 2019). However, Hase et al. (2020) finds that both two-stage finetuning and multi-task learning with MT-Re and MT-Ra setups only obtain comparable results over standard finetuning baselines on T5. We instead show that MT-Re, MT-Ra and our proposed MT-CoT with explanations from LLM can consistently and significantly outperform standard finetuning baselines without accuracy-explanation trade-off (Jain et al., 2020).

3 Explanation Generation from LLM

Problem setup. Denote $D = \{(x_i, y_i)\}^N$ to be a dataset with N training instances, where x_i is a problem and y_i is its answer. Also, we have a handful of human-written instances $E = \{(x_i^p, e_i^p, y_i^p)\}^M$, where e_i^p is a free-text explanation to explain why a problem x_i^p has y_i^p as its answer and $\{(x_i^p, y_i^p)\}^M \subset D$ with $M \ll N$ (we set $M = 7$ in our experiments). Our goal is to fully leverage LLM with E as demonstrations for in-context learning to generate explanation e_i for all (x_i, y_i), where $1 \leq i \leq N$, so that we can utilize these generated explanations from LLM to improve SLM reasoning capability.

COTE. A chain of thought is a series of intermediate reasoning steps before providing an answer of a problem, mimicking human deliberate thinking process to perform complicated reasoning tasks (Wei et al., 2022b). Chain of thought prompting provides intermediate reasoning steps as explana-



qta Q: The only baggage the woman checked was a drawstring bag, where was she heading with it? Answer Choices: (a) garbage can (b) military (c) jewelry store (d) safe (e) airport A:

qtr Q: The only baggage the woman checked was a drawstring bag, where was she heading with it? Answer Choices: (a) garbage can (b) military (c) jewelry store (d) safe (e) airport A:

The answer must be a place where someone would check a bag. The only place where someone would check a bag is at an airport.

The answer is airport (e). Explanation: The answer must be a place where someone would check a bag. The only place where someone would check a bag is at an airport.

The answer must be a place where someone would check a bag. The only place where someone would check a bag is at an airport. Therefore, the answer is airport (e).

Figure 2: The comparison among (a) MT-Re (Hase et al., 2020), (b) MT-Ra (Camburu et al., 2018) and (c) our proposed MT-CoT for multi-task learning with explanations under text-to-text format using T5. Left parts are inputs of T5 and right parts are targets for different multi-task learning setups. Task *qta* (question to answer) is trained to directly generate answers for all three modes while *qtr* (question to reason) task is trained to generate reasoning, rationalization and chain of thought for (a) MT-Re, (b) MT-Ra and (c) MT-CoT, respectively.

tions before answers in prompts. Formally, for $1 \leq i \leq N$, we first concatenate all instances in E and x_i as prompt $\hat{p}_i$ = (x_1^p, e_1^p , y_1^p, ..., x_M^p, e_M^p, y_M^p, x_i). We then feed prompt $\hat{p}_i$ into LLM and greedily decode until a stop token is generated. After that, we parse the decoded sentence as explanation part $\hat{e}_i$ and prediction part $\hat{y}_i$. Intuitively, if $\hat{y}_i \neq y_i$, $\hat{e}_i$ may not have high quality as incorrect explanations tend to generate incorrect predictions (Wei et al., 2022b). Thus, we utilize ***C**hain **O**f **T**hought prompting with incorrect answer r**E**jection* (COTE) (Zelikman et al., 2022) by only adopting $e_i := \hat{e}_i$ if $\hat{y}_i = y_i$; otherwise, we reject $\hat{e}_i$ and set e_i as *none*.

RP. Since COTE uses the answers in original datasets to reject explanations with incorrect predictions, these instances will no longer have explanations. To alleviate this issue, an alternative is apply ***R**ationalization **P**rompting* (RP) (Wiegreffe et al., 2021a) to generate explanations for every instance in training sets. Unlike COTE, RP provides explanations given golden answers. Specifically, for $1 \leq i \leq N$, we concatenate all instances in E and (x_i, y_i) as prompt $\bar{p}_i$ = (x_1^p, y_1^p, e_1^p , ..., x_M^p, y_M^p, e_M^p, x_i, y_i). We then feed prompt $\bar{p}_i$ into LLM and greedily decode until a stop token is generated. The decoded sentence $\bar{e}_i$ is cast as explanation $\hat{e}_i$, i.e. $e_i := \bar{e}_i$, without filtering.

CROP. COTE will possibly generate relatively high-quality explanations if LLM give correct predictions of problems at hand as incorrect explanations tend to generate incorrect predictions (Wei et al., 2022b). However, for problems with incorrect predictions, COTE casts their explanations as *none*. On the other hand, RP can generate explanations for every instance in the dataset, but we cannot easily assess their quality without human annotation. Therefore, we propose ***C**hain of Thought with **R**ationalization Pr**O**mpting backu**P*** (CROP), where when COTE generates *none* as explanations, we will utilize RP as a backup approach. Intuitively, if LLM cannot predict a problem correctly under chain of thought prompting, the problem may be difficult (Zelikman et al., 2022) and RP may provide a meaningful explanation as it can access golden label during explanation generation process.

4 Multi-task Learning with Explanations

In this section, we elaborate how to utilize explanations generated from LLM to improve SLM reasoning capability with a multi-task learning framework. We detail three multi-task learning with explanations methods in the following.

MT-Re. Multi-task Learning with Reasoning (MT-Re) is introduced by Hase et al. (2020) (see Figure 2 (a)). MT-Re is trained to directly generate predictions for *qta* (question to answer) task the same as standard finetuning without explanations and generate explanations without explicitly providing answers in *qtr* (question to reason) task. The training objective of MT-Re is to mix loss $\mathcal{L}_{\text{qta}}$ for *qta* task and $\mathcal{L}_{\text{qtr}}$ for *qtr* task:

$$\mathcal{L}_{\text{mt}} = \alpha\mathcal{L}_{\text{qta}} + (1-\alpha)\mathcal{L}_{\text{qtr}}, \tag{1}$$

where α weights $\mathcal{L}_{\text{qta}}$ and $\mathcal{L}_{\text{qtr}}$ loss, and is tuned on development set.

MT-Ra. Multi-task Learning with Rationalization (MT-Ra) is first proposed by Camburu et al. (2018) for natural language inference task using LSTM-based models (Hochreiter and Schmidhuber, 1997) and we adopt it with a more powerful T5 model for other reasoning tasks. As shown in Figure 2 (b), models are trained to generate predictions for *qta* task the same as MT-Re and also trained to generate rationalization for *qtr* task. This is different from MT-Re as MT-Ra allows explanations to be explicitly conditioned on predictions. For MT-Ra, we use the same training objective as Equation 1 and tune α on development set.

MT-CoT. MT-Re does not explicitly model interactions between explanations and answers during training, which may make models hard to capture their relations. While MT-Ra is explicitly trained to generate explanations conditioned on answers, it may still have difficulty in understanding their causal effects as answers are never trained to explicitly access their explanations. To bridge this gap, we propose Multi-task Learning with Chain of Thought (MT-CoT), where models are trained to generate answers for *qta* task and generate chain of thought for *qtr* task, as shown in Figure 2 (c). For MT-CoT, we use the same training objective as Equation 1 and tune α on development set.

In MT-CoT training paradigm, models not only know answers from *qta* task but also are explicitly shown how answers are derived with intermediate reasoning steps before knowing them from *qtr* task. As we will show in experiments, this training paradigm is a supplement to MT-Re and MT-Ra, and can consistently improve small language model reasoning capability and also outperform MT-Re and MT-Ra on two datasets.

5 Experiments

5.1 Experimental setup

We evaluate our methods on three reasoning tasks.

(1) **CommonsenseQA** (Talmor et al., 2019) is a 5-way multi-choice question answering dataset that requires common sense reasoning with 9741/1221/1140 for training/development/test set questions, respectively. Since its test set is not publicly available, we report results on its development set following previous work (Zelikman et al., 2022; Li et al., 2019).

(2) **StrategyQA** is a binary yes/no question answering dataset requiring implicit multi-hop reasoning steps and should be inferred using a strategy (Geva et al., 2021). It has 2290 training set and 490 test set questions. Since its test set is not publicly available, we utilize their split in GitHub [2], where original training set is randomly split into 90% for training and 10% for development set. In our experiments, we report results on their Github development set and utilize their Github training set for training without utilizing explanations from their original annotations.

(3) **OpenbookQA** is a 4-way multi-choice question answering dataset requiring open book facts with broad common knowledge and multi-hop reasoning (Mihaylov et al., 2018). It has 4957/500/500 questions for training/development/test set split, respectively and we report results on its test set.

Explanation generation from LLM We utilize GPT-3 *text-davinci-002* engine with official OpenAI API [3] to generate explanations through greedy decoding (by setting temperature as 0) following in-context learning paradigm. In each dataset, we have the same 7-shot examples with human-written explanations for COTE, RP and CROP detailed in section 2. We defer details of prompts into Appendix A.

Multi-task learning with explanations. After obtaining explanations by COTE, RP and CROP, we utilize MT-Re, MT-Ra and MT-CoT introduced in section 4 to train models with explanations based on T5. We implement multi-task learning framework with Huggingface *transformers* library (Wolf et al., 2020). For baselines, we utilize single-task finetuning (ST) without explanations. For fair comparison with ST, we keep hyper-parameters of multi-task learning the same as its corresponding ST except weight α which we tune with grid search $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ on development sets. When training on *none* explanations generated by COTE, we mask their loss for *qtr* task. For both ST and multi-task finetuning, we directly generate predictions from *qta* task for fair comparisons.

5.2 Main results

In this section, we compare results between multi-task learning with explanations and its single-task

---

[2] `https://github.com/eladsegal/strategyqa`

[3] `https://beta.openai.com/docs/models/gpt-3`

	CommonsenseQA			StrategyQA			OpenbookQA		
	COTE	RP	CROP	COTE	RP	CROP	COTE	RP	CROP
ST		$63.05_{0.50}$			$58.60_{1.36}$			$58.08_{0.65}$	
MT-Re	$63.78_{0.43}$	$63.78_{0.20}$	$64.05_{0.22}$	$60.26_{0.92}$	$60.52_{0.81}$	$60.26_{0.62}$	$59.48_{0.93}$	$60.44_{1.49}$	$59.04_{1.63}$
MT-Ra	$\underline{64.05}_{0.60}$	$\underline{64.14}_{0.22}$	$\underline{\mathbf{64.50}}_{0.22}$	$\underline{60.52}_{0.86}$	$\underline{60.79}_{0.43}$	$60.61_{0.64}$	$58.68_{2.11}$	$59.52_{0.20}$	$\underline{60.40}_{0.59}$
MT-CoT	$63.88_{0.14}$	$63.69_{0.30}$	$63.75_{0.51}$	$60.26_{1.46}$	$\underline{60.79}_{1.31}$	$\underline{\mathbf{61.05}}_{0.85}$	$\underline{\mathbf{60.68}}_{0.37}$	$\underline{60.64}_{0.66}$	$59.64_{0.90}$

Table 1: Accuracy comparison (%) of single-task finetuning baselines (ST) with MT-Re, MT-Ra and MT-CoT utilizing explanations generated by COTE, RP and CROP. Results are averaged over five runs with their standard deviation in the subscript. Best results for each column with the same explanations are underlined and best results for each **dataset** are bold.

	50	100	200	400
	CommonsenseQA			
ST	$21.92_{1.57}$	$27.06_{2.83}$	$28.04_{2.78}$	$44.49_{2.16}$
MT	$\mathbf{29.25}_{3.03}$	$\mathbf{33.28}_{3.53}$	$\mathbf{36.13}_{5.29}$	$\mathbf{46.55}_{1.53}$
α*	0.1	0.2	0.3	0.6
	OpenbookQA			
ST	$27.08_{2.96}$	$28.32_{2.88}$	$30.68_{2.10}$	$37.80_{4.64}$
MT	$\mathbf{29.76}_{3.74}$	$\mathbf{32.92}_{0.95}$	$\mathbf{34.84}_{1.27}$	$\mathbf{43.68}_{0.94}$
α*	0.1	0.1	0.2	0.2

Table 2: Accuracy comparison (%) between single-task finetuning (ST) and multi-task learning with explanations (MT) along with optimal α* in development sets under different training sample sizes. Results are averaged over five different training data splits with their standard deviation listed in the subscript.

finetuning counterpart using full training data on three datasets introduced in section 5.1. Specifically, we generate explanations for each dataset with COTE, RP and CROP, and for each explanation generation method, we train T5-base model under MT-Re, MT-Ra and MT-CoT setups with 5 different runs in each setting. For single-task finetuning baseline, we only keep *qta* task by removing *qtr* task in multi-task learning setup. Results are summarized in Table 1.

Three multi-task learning with three different explanation generation methods consistently and significantly outperform single-task finetuning baselines, showing the effectiveness of utilizing explanations from LLM. However, MT-CoT and MT-Ra have 4 and 6 underlined results, respectively, while MT-Re does not have any. We hypothesize it is because MT-CoT and MT-Ra *explicitly* mention answers by *the answer is* in *qtr* task, making it easier for T5 to model relations between explanations and answers. Considering best results for each dataset, two of three are obtained via CROP with the remaining one obtained by COTE, showing that chain of thought prompting generates better explanations for SLM finetuning when their predictions are correct and RP backup can possibly further improve SLM reasoning capability. In addition, two of these three best results are obtained by MT-CoT, demonstrating that our method MT-CoT can serve as a good candidate to improve SLM reasoning with explanations from the toolbox.

5.3 Few-shot learning results

We have shown the effectiveness of our method on full-training settings in section 5.2 and further explore if explanations can improve SLM reasoning capability under few-shot settings. We conduct few-shot learning experiments for both *CommonsenseQA* and *OpenbookQA* datasets with best settings in section 5.2. Specifically, we choose MT-Ra finetuning with explanations generated by CROP for *CommonsenseQA* dataset and MT-CoT finetuning with explanations generated by COTE for *OpenbookQA* dataset. We conduct experiments with $\{50, 100, 200, 400\}$ training sample sizes for both datasets on T5-base model and for each sample size, we randomly sample five data splits from its whole training set and each data split has a single run. Similar to previous experiments, we have single-task finetuning as our baselines and tune α using grid search on development sets for multi-task learning experiments. Besides accuracy, we also report optimal α on development sets, denoted as α*. Intuitively, if α* is small, $\mathcal{L}_{\text{qtr}}$ loss has more weight in the multi-task learning training objective listed in Equation 1 and hence, explanations are more important for correct prediction. We summarize our results in Table 2.

Multi-task learning with explanations (MT) consistently and significantly outperforms single-task finetuning baselines (ST). For *CommonsenseQA* dataset, when training sample sizes are in $\{50, 100, 200\}$, MT significantly improves over ST about 6%-8% absolute accuracy. For *OpenbookQA* dataset, when training sample sizes are in $\{100, 200, 400\}$, MT improves over ST about 4%-6% absolute accuracy. More interestingly, $\alpha*$ tends to be smaller when less training data is used on both datasets. Intuitively, when training data sizes are small, models may have difficulty in learning just from limited problem and answer pairs and hence, requires a small $\alpha*$ in the multi-task training objective 1, i.e. larger weight on $\mathcal{L}_{\text{qtr}}$ loss during multi-task learning process. These consistent and significant gains show that our method not only can improve results in full-training settings but also is very useful when training data is limited.

	T5-small	T5-base	T5-large	T5-3B
		CommonsenseQA		
ST	48.26	63.05	72.56	81.82
MT	**49.17**	**64.50**	**74.37**	**82.47**
		OpenbookQA		
ST	50.36	58.08	61.60	76.60
MT	**51.72**	**60.68**	**64.60**	**78.60**

Table 3: Accuracy comparison (%) between ST and MT across different model sizes.

5.4 Results across model sizes

Previous experiments utilize T5-base model and we further explore if explanations can improve language model reasoning capability across model sizes. We conduct full-training set experiments for both *CommonsenseQA* and *OpenbookQA* datasets with best settings for each dataset in section 5.2 across $\{\text{T5-small}, \text{T5-base}, \text{T5-large}, \text{T5-3B}\}$. For T5-small and T5-base, we have five different runs for each setting and their average results are reported. For T5-large and T5-3B, we only report a single run due to their intensive computational cost. Results are summarized in Table 3.

MT consistently improves its ST counterpart on both *CommonsenQA* and *OpenbookQA* across model sizes from T5-small (60 million parameters) to T5-3B. For *CommonsenQA*, MT improves ST about 0.7%-1.8% absolute accuracy and for *OpenbookQA*, MT improves ST about 1.4%-3.0% absolute accuracy. Even for T5-3B, MT can improve strong ST with 2% absolute accuracy. These consistent results show that our approach can work on both small and relatively large models.

	CSQA	OBQA
GPT-J Direct Finetuning (6B) ◇	60.0	-
STaR (6B) ◇	72.5	-
GPT-3 Direct Finetuning (175B)*	73.0	-
GPT-3 Direct Prompting (175B)	80.59	83.00
GPT-3 Chain of Thought Prompting (175B)	73.71	72.60
GPT-3 Explain. after Answers Prompting (175B)	80.84	**83.40**
T5 MT (3B)	**82.47**	78.60

Table 4: Accuracy comparison (%) between T5 multi-task learning with explanations with various state-of-the-art LLM on CommonsenseQA (CSQA) and OpenbookQA (OBQA), and model sizes are listed in the parenthesis. Results with ◇ and * are from Zelikman et al. (2022) and Xu et al. (2021), respectively.

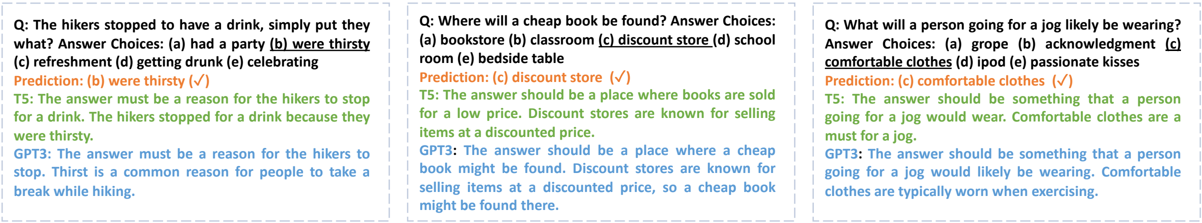
5.5 Comparison with Large Language Models

We further compare our method on T5-3B with state-of-the-art LLM. Specifically, we adopt GPT-J direct finetuning, its self-bootstrapping version (STaR) (Zelikman et al., 2022) and GPT-3 direct finetuning (Xu et al., 2021) as baseline methods with parameter update on downstream tasks. We also adopt GPT-3 direct prompting (Brown et al., 2020), GPT-3 chain of thought prompting (Wei et al., 2022b) and GPT-3 explanations after answers prompting (Lampinen et al., 2022) as prompting baselines. These three prompting methods utilize the same set of demonstrations for explanation generation in section 2 and we defer their prompts into Appendix A. Results are summarized in Table 4.

Our approach can outperform strong 60x larger GPT-3 finetuning and various GPT-3 prompting methods on *CommonsenseQA* up to about 9.5% absolute accuracy. Also, although STaR can outperform its GPT-J baseline with chain-of-thought style iterative finetuning, their result still has about 10% absolute accuracy gap with our method on *CommonsenseQA* even with doubled parameter size and more compute during iterative finetuning process. For *OpenbookQA*, our model underperforms GPT-3 direct prompting and explanations after answers prompting but can still outperform GPT-3 chain of thought prompting with 6% absolute accuracy. In short, our method can achieve strong performance even compared with 60x larger GPT-3.

5.6 Human evaluation on generated explanations

A side benefit of our model is to generate explanations towards more explainable AI to alleviate the notorious black box issue of deep neural networks (Koh and Liang, 2017). To evaluate quality of gen-



Q: The hikers stopped to have a drink, simply put they what? Answer Choices: (a) had a party (b) were thirsty (c) refreshment (d) getting drunk (e) celebrating
Prediction: (b) were thirsty (✓)
T5: The answer must be a reason for the hikers to stop for a drink. The hikers stopped for a drink because they were thirsty.
GPT3: The answer must be a reason for the hikers to stop. Thirst is a common reason for people to take a break while hiking.

Q: Where will a cheap book be found? Answer Choices: (a) bookstore (b) classroom (c) discount store (d) school room (e) bedside table
Prediction: (c) discount store (✓)
T5: The answer should be a place where books are sold for a low price. Discount stores are known for selling items at a discounted price.
GPT3: The answer should be a place where a cheap book might be found. Discount stores are known for selling items at a discounted price, so a cheap book might be found there.

Q: What will a person going for a jog likely be wearing? Answer Choices: (a) grope (b) acknowledgment (c) comfortable clothes (d) ipod (e) passionate kisses
Prediction: (c) comfortable clothes (✓)
T5: The answer should be something that a person going for a jog would wear. Comfortable clothes are a must for a jog.
GPT3: The answer should be something that a person going for a jog would likely be wearing. Comfortable clothes are typically worn when exercising.

(a) T5 wins (b) Tie (c) T5 loses

Figure 3: T5 and GPT-3 generated explanations used in human evaluation.

Preference			Agreement Level		
T5	Tie	GPT-3	Level 0	Level 1	Level 2
14%	44%	42%	7%	56%	37%

Table 5: Head-to-head human explanation preference comparison between T5 (3B) and GPT-3 (175B) on *CommonsenseQA* along with their agreement percentage on three levels.

erated explanations from *qtr* task for our model, we conduct human evaluation since automatic metrics are not highly correlated with human assessment (Clinciu et al., 2021; Kayser et al., 2021).

Specifically, we perform a head-to-head explanation comparison on *CommonsenseQA* dataset between T5-3B and GPT-3 175B few-shot explanations after answers prompting since these models achieve close performance on this dataset, as shown in Table 4. T5 model is trained with explanations generated by GPT-3 and we would like to know its generated explanation quality compared to that of GPT-3, which has been shown to be high-quality in Wiegreffe et al. (2021a). Therefore, we randomly sample 100 examples that are predicted correctly by both GPT-3 and T5, and for each example, we present a question, its ground truth answer and two randomly shuffle explanations as (a) and (b) generated by T5 and GPT-3 to three different human annotators with advanced NLP backgrounds and then ask them which explanation they prefer: (a), (b) or tie, similar to Wiegreffe et al. (2021a). Finally, we adopt a majority vote approach to decide preference on each example if at least two annotators have the same preference; otherwise, we cast that example's two explanations are tied. In addition, we report agreement percentage across three levels. Level 0 means all three annotators have different preferences, level 1 means only two annotators have the same preference and level 2 means all three annotators have the same preference. Results are summarized in Table 5.

As expected, explanations generated by T5 are less preferred over those from GPT-3 but there are still 58% (14%+44%) explanations having better or competitive quality over GPT-3. In addition, more than 60% explanations have disagreement (7% in level 0 + 56% in level 1). Given Wiegreffe et al. (2021a) finds GPT-3 can generate competitive explanations even compared to human-written ones, we argue that this high disagreement is because explanations generated by both T5 and GPT-3 are high-quality, making humans hard to choose. To verify this hypothesis, we choose three T5 and GPT-3 generated explanation examples used in our human evaluation experiments, as shown in Figure 3. Both T5 and GPT-3 can generate plausible explanations to justify their predictions and even though T5 loses to GPT-3 in example (c), its explanation is still reasonably good. We also provide examples with incorrect predictions in appendix B, some of which still have plausible predictions and explanations although different from golden labels. These results demonstrate explanations generated by our model have high quality even compared with strong GPT-3 with 60x larger size.

6 Conclusion

In this paper, we leverage explanations from LLM to improve small reasoners in a multi-task learning framework. Extensive experiments on multiple reasoning tasks show our method can consistently and significantly outperform single-task finetuning baselines across explanation generation method, multi-task learning setups, training sample and small reasoner sizes, and can outperform strong finetuning/prompting a 60x larger GPT-3 on *CommonsenseQA* by up to 9.5% in accuracy. In addition, our model can generate high-quality explanations even compared to strong GPT-3 towards more explainable AI according to human evaluation.

Limitations

Our approach requires a multi-task learning finetuning approach to integrate explanations into small language models and will require tuning weight α on a development set, which will require more compute during hyper-parameter tuning process. In addition, our work is constrained to textual reasoning problems, lacking more explorations in other reasoning tasks, e.g. symbolic reasoning and arithmetic reasoning, which we plan to leave as future work.

Ethics Statement

Our work is built on top of explanations generated from LLM, which have been observed to capture gender, race and religion biases (Brown et al., 2020; Lucy and Bamman, 2021; Abid et al., 2021). Generated explanations with these possible biases may be integrated into small models during finetuning process and be exposed when these small models generate explanations to justify their predictions. Therefore, our model could potentially share the same kinds of bias as the original LLM used for explanation generation. However, our multi-task learning framework naturally allows us to disable explanation generation and still enjoy performance gains by direct answer prediction without the risk of explicitly exposing these biases.

Acknowledgement

This research was sponsored in part by the DARPA PTG program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Abubakar Abid, Maheen Saleem Farooqi, and James Y. Zou. 2021. Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *EMNLP (Findings)*, pages 4351–4367.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *ECCV*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

W. Lewis Johnson. 1994. Agents that learn to explain themselves. *AAAI*.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1224–1234.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *ArXiv*, abs/1703.04730.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Andrew Kyle Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *ArXiv*, abs/2204.02329.

Shiyang Li, Jianshu Chen, and Dian Yu. 2019. Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach. *ArXiv*, abs/1909.09743.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *ArXiv*, abs/2004.14546.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/ language understanding paper.pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, Luan David, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray

Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark O. Riedl, and Yejin Choi. 2021a. Reframing human-ai collaboration for generating free-text explanations. *ArXiv*, abs/2112.08674.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *NeurIPS Datasets and Benchmarks*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021b. Measuring association between labels and free-text rationales. In *EMNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2021. Human parity on commonsenseqa: Augmenting self-attention with external attention. *ArXiv*, abs/2112.03254.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning. *ArXiv*, abs/2205.03401.

E. Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *ArXiv*, abs/2203.14465.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *ArXiv*, abs/2205.10625.

Appendix

A Prompt details

Here we provides prompts we use in our experiments. Our prompts on *CommonsenseQA* and *StrategyQA* datasets are based on (Zelikman et al., 2022) and (Wei et al., 2022b), respectively. Explanations in prompts for *OpenbookQA* are based on science facts in *OpenbookQA* dataset Github repository https://github.com/allenai/OpenBookQA.

Q: What do people use to absorb extra ink from a fountain pen?
Answer Choices:
(a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A: blotter (e)

Q: What home entertainment equipment requires cable?
Answer Choices:
(a) radio shack (b) substation (c) television (d) cabinet (e) desk

A: television (c)

Q: The fox walked from the city into the forest, what was it looking for?
Answer Choices:
(a) pretty flowers (b) hen house (c) natural habitat (d) storybook (e) dense forest

A: natural habitat (c)

Q: Sammy wanted to go to where the people were. Where might he go?
Answer Choices:
(a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

A: populated areas (a)

Q: Where do you put your grapes just before checking out?
Answer Choices:
(a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market

A: grocery cart (b)

Q: Google Maps and other highway and street GPS services have replaced what?
Answer Choices:
(a) united states (b) mexico (c) countryside (d) atlas (e) oceans

A: atlas (d)

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices:
(a) harder (b) anguish (c) bitterness (d) tears (e) sadness

A: bitterness (c)

Table 6: Standard prompt for CommonsenseQA.

Q: What do people use to absorb extra ink from a fountain pen?
Answer Choices:
(a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A: The answer must be used to absorb extra ink. Blotters are designed to absorb liquids. Therefore, the answer is blotter (e).

Q: What home entertainment equipment requires cable?
Answer Choices:
(a) radio shack(b) substation (c) television (d) cabinet (e) desk

A: The answer must require cable. Cable is used to provide satellite channels to televisions. Therefore, the answer is television (c).

Q: The fox walked from the city into the forest, what was it looking for?
Answer Choices:
(a) pretty flowers (b) hen house (c) natural habitat (d) storybook (e) dense forest

A: The answer must be a reason for a fox to go into the forest. The forest is a fox's natural habitat. Therefore, the answer is natural habitat (c).

Q: Sammy wanted to go to where the people were. Where might he go?
Answer Choices:
(a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

A: The answer must be a place with many people. Populated areas, by definition, have a lot of people. Therefore, the answer is populated areas (a).

Q: Where do you put your grapes just before checking out?
Answer Choices:
(a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market

A: The answer should be the place where grocery items are placed before checking out. Of the above choices, grocery cart makes the most sense for holding grocery items. Therefore, the answer is grocery cart (b).

Q: Google Maps and other highway and street GPS services have replaced what?
Answer Choices:
(a) united states (b) mexico (c) countryside (d) atlas (e) oceans

A: The answer must be something that used to do what Google Maps and GPS services do, which is give directions. Atlases were also used to give directions. Therefore, the answer is atlas (d).

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices:
(a) harder (b) anguish (c) bitterness (d) tears (e) sadness

A: The answer should be a feeling which would cause someone who was doing all the work to get divorced. If someone feels bitter towards their spouse, they are likely to want a divorce. Therefore, the answer is bitterness (c).

Table 7: Chain of Thought prompt for CommonsenseQA.

Q: What do people use to absorb extra ink from a fountain pen?
Answer Choices:
(a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A: blotter (e)
Explanation: the answer must be used to absorb extra ink. Blotters are designed to absorb liquids.

Q: What home entertainment equipment requires cable?
Answer Choices:
(a) radio shack (b) substation (c) television (d) cabinet (e) desk

A: television (c)
Explanation: the answer must require cable. Cable is used to provide satellite channels to televisions.

Q: The fox walked from the city into the forest, what was it looking for?
Answer Choices:
(a) pretty flowers (b) hen house (c) natural habitat (d) storybook (e) dense forest

A: natural habitat (c)
Explanation: the answer must be a reason for a fox to go into the forest. The forest is a fox's natural habitat.

Q: Sammy wanted to go to where the people were. Where might he go?
Answer Choices:
(a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

A: populated areas (a)
Explanation: the answer must be a place with many people. Populated areas, by definition, have a lot of people.

Q: Where do you put your grapes just before checking out?
Answer Choices:
(a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market

A: grocery cart (b)
Explanation: the answer should be the place where grocery items are placed before checking out. Of the above choices, grocery cart makes the most sense for holding grocery items.

Q: Google Maps and other highway and street GPS services have replaced what?
Answer Choices:
(a) united states (b) mexico (c) countryside (d) atlas (e) oceans

A: atlas (d)
Explanation: the answer must be something that used to do what Google Maps and GPS services do, which is give directions. Atlases were also used to give directions.

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices:
(a) harder (b) anguish (c) bitterness (d) tears (e) sadness

A: bitterness (c)
Explanation: the answer should be a feeling which would cause someone who was doing all the work to get divorced. If someone feels bitter towards their spouse, they are likely to want a divorce.

Table 8: Rationalization prompt for CommonsenseQA.

Q: Yes or no: Do hamsters provide food for any animals?

A: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is yes.

Q: Yes or no: Could the members of The Police perform lawful arrests?

A: The members of The Police were musicians, not law enforcement officers. Only law enforcement officers can perform lawful arrests. Thus, the members of The Police could not perform lawful arrests. So the answer is no.

Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?

A: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5. So the answer is no.

Q: Yes or no: Is it common to see frost during some college commencements?

A: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. So the answer is yes.

Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)?

A: The War in Vietnam was 6 months. The gestation period for a llama is 11 months, which is more than 6 months. Thus, a llama could not give birth twice during the War in Vietnam. So the answer is no.

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.

Q: Yes or no: Is shrimp scampi definitely free of plastic?

A: Shrimp scampi is a dish made with shrimp. Shrimp have been found to contain microplastics. Microplastics are plastic material. Thus, shrimp scampi may contain plastic. So the answer is no.

Table 9: Chain of Thought prompt for StrategyQA.

Q: Yes or no: Do hamsters provide food for any animals?

A: yes
Explanation: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals.

Q: Yes or no: Could the members of The Police perform lawful arrests?

A: no
Explanation: The members of The Police were musicians, not law enforcement officers. Only law enforcement officers can perform lawful arrests. Thus, the members of The Police could not perform lawful arrests.

Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?

A: no
Explanation: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5.

Q: Yes or no: Is it common to see frost during some college commencements?

A: yes
Explanation: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements.

Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)?

A: no
Explanation: The War in Vietnam was 6 months. The gestation period for a llama is 11 months, which is more than 6 months. Thus, a llama could not give birth twice during the War in Vietnam.

Q: Yes or no: Would a pear sink in water?

A: no
Explanation: The density of a pear is about 0.6 g/cm^3, which is less than water. Objects less dense than water float. Thus, a pear would float.

Q: Yes or no: Is shrimp scampi definitely free of plastic?

A: no
Explanation: Shrimp scampi is a dish made with shrimp. Shrimp have been found to contain microplastics. Microplastics are plastic material. Thus, shrimp scampi may contain plastic.

Table 10: Rationalization prompt for StrategyQA.

Q: What is the most likely to be an effect of acid rain on an aquatic environment?
Answer Choices:
(a) decrease in plant life (b) increase in fish population (c) increase in plant growth (d) cleaner and clearer water

A: (a) decrease in plant life

Q: The moon's surface
Answer Choices:
(a) is smooth on the entire surface (b) contains large cavities cause by explosions (c) contains an internal core of cheese (d) is filled with lakes

A: (b) contains large cavities cause by explosions

Q: As a car approaches you in the night
Answer Choices:
(a) the headlights become more intense (b) the headlights recede into the dark (c) the headlights remain at a constant (d) the headlights turn off

A: (a) the headlights become more intense

Q: When the weather changes as it does from Christmas to Easter,
Answer Choices:
(a) the air may chill (b) the ground may freeze (c) the plants may die (d) the ground may warm

A: (d) the ground may warm

Q: Heat and moisture in the ocean is a good recipe for
Answer Choices:
(a) a violent storm (b) violent sea animals (c) condensation (d) inland storms

A: (a) a violent storm

Q: Hummingbirds take what with them
Answer Choices:
(a) Bees (b) energy (c) Pollen (d) Honey

A: (c) Pollen

Q: What covers over 90% of the Earth's surface and 0% of the moon's surface
Answer Choices:
(a) a magnesium iron silicate mineral (b) chemical element with the symbol S (c) the element with the symbol Fe (d) that which contains 2 hydrogen and 1 oxygen molecules

A: (d) that which contains 2 hydrogen and 1 oxygen molecules

Table 11: Standard prompt for OpenbookQA.

Q: What is the most likely to be an effect of acid rain on an aquatic environment?
Answer Choices:
(a) decrease in plant life (b) increase in fish population (c) increase in plant growth (d) cleaner and clearer water

A: Acid rain has a negative impact on water quality and decrease in water quality leads to a decrease in aquatic life. So the answer is (a) decrease in plant life.

Q: The moon's surface
Answer Choices:
(a) is smooth on the entire surface (b) contains large cavities cause by explosions (c) contains an internal core of cheese (d) is filled with lakes

A: The moon's surface contains many craters and craters are large cavities caused by explosions. So the answer is (b) contains large cavities cause by explosions.

Q: As a car approaches you in the night
Answer Choices:
(a) the headlights become more intense (b) the headlights recede into the dark (c) the headlights remain at a constant (d) the headlights turn off

A: Headlights of a car are source of light. As a source of light becomes closer , that source will appear brighter. So the answer is (a) the headlights become more intense.

Q: When the weather changes as it does from Christmas to Easter,
Answer Choices:
(a) the air may chill (b) the ground may freeze (c) the plants may die (d) the ground may warm

A: Christmas is in the winter season while Easter is in the beginning of spring. Weather change from winter to spring is to become warm. So the answer is (d) the ground may warm.

Q: Heat and moisture in the ocean is a good recipe for
Answer Choices:
(a) a violent storm (b) violent sea animals (c) condensation (d) inland storms

A: An ocean is a source of heat and moisture for a hurricane and a hurricane is a violent storm. So the answer is (a) a violent storm.

Q: Hummingbirds take what with them
Answer Choices:
(a) Bees (b) energy (c) Pollen (d) Honey

A: When hummingbirds dip their long bills into flowers to drink nectar, they also pick up grains of pollen. Pollen that sticks to a hummingbird's feathers and bill gets carried to the next flower it visits. So the answer is (c) Pollen.

Q: What covers over 90% of the Earth's surface and 0% of the moon's surface
Answer Choices:
(a) a magnesium iron silicate mineral (b) chemical element with the symbol S (c) the element with the symbol Fe (d) that which contains 2 hydrogen and 1 oxygen molecules

A: Water covers over 90% of the Earth's surface and 0% of the moon's surface, and contains 2 hydrogen and 1 oxygen molecules. So the answer is (d) that which contains 2 hydrogen and 1 oxygen molecules.

Table 12: Chain of Thought prompt for OpenbookQA.

Q: What is the most likely to be an effect of acid rain on an aquatic environment?
Answer Choices:
(a) decrease in plant life (b) increase in fish population (c) increase in plant growth (d) cleaner and clearer water

A: (a) decrease in plant life
Explanation: Acid rain has a negative impact on water quality and decrease in water quality leads to a decrease in aquatic life.

Q: The moon's surface
Answer Choices:
(a) is smooth on the entire surface (b) contains large cavities cause by explosions (c) contains an internal core of cheese (d) is filled with lakes

A: (b) contains large cavities cause by explosions
Explanation: The moon's surface contains many craters and craters are large cavities caused by explosions.

Q: As a car approaches you in the night
Answer Choices:
(a) the headlights become more intense (b) the headlights recede into the dark (c) the headlights remain at a constant (d) the headlights turn off

A: (a) the headlights become more intense
Explanation: Headlights of a car are source of light. As a source of light becomes closer , that source will appear brighter.

Q: When the weather changes as it does from Christmas to Easter,
Answer Choices:
(a) the air may chill (b) the ground may freeze (c) the plants may die (d) the ground may warm

A: (d) the ground may warm
Explanation: Christmas is in the winter season while Easter is in the beginning of spring. Weather change from winter to spring is to become warm.

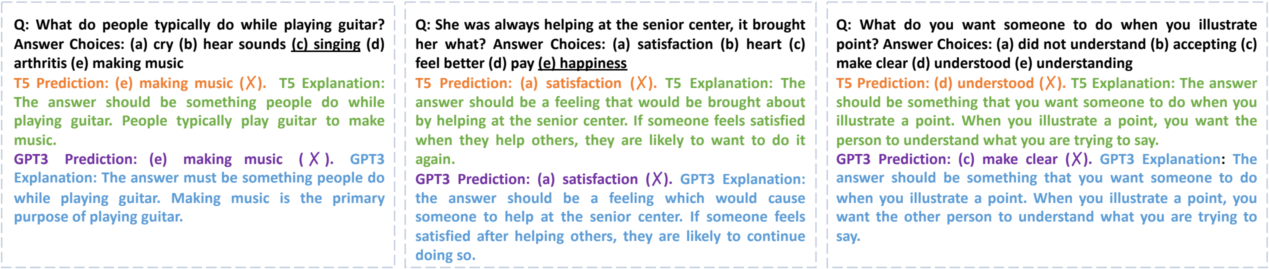
Q: Heat and moisture in the ocean is a good recipe for
Answer Choices:
(a) a violent storm (b) violent sea animals (c) condensation (d) inland storms A: (a) a violent storm
Explanation: An ocean is a source of heat and moisture for a hurricane and a hurricane is a violent storm.

Q: Hummingbirds take what with them
Answer Choices:
(a) Bees (b) energy (c) Pollen (d) Honey A: (c) Pollen
Explanation: When hummingbirds dip their long bills into flowers to drink nectar, they also pick up grains of pollen. Pollen that sticks to a hummingbird's feathers and bill gets carried to the next flower it visits.

Q: What covers over 90% of the Earth's surface and 0% of the moon's surface
Answer Choices:
(a) a magnesium iron silicate mineral (b) chemical element with the symbol S (c) the element with the symbol Fe (d) that which contains 2 hydrogen and 1 oxygen molecules

A: (d) that which contains 2 hydrogen and 1 oxygen molecules
Explanation: Water covers over 90% of the Earth's surface and 0% of the moon's surface, and contains 2 hydrogen and 1 oxygen molecules.

Table 13: Rationalization prompt for OpenbookQA.



Q: What do people typically do while playing guitar? Answer Choices: (a) cry (b) hear sounds (c) singing (d) arthritis (e) making music

T5 Prediction: (e) making music (✗). T5 Explanation: The answer should be something people do while playing guitar. People typically play guitar to make music.

GPT3 Prediction: (e) making music (✗). GPT3 Explanation: The answer must be something people do while playing guitar. Making music is the primary purpose of playing guitar.

Q: She was always helping at the senior center, it brought her what? Answer Choices: (a) satisfaction (b) heart (c) feel better (d) pay (e) happiness

T5 Prediction: (a) satisfaction (✗). T5 Explanation: The answer should be a feeling that would be brought about by helping at the senior center. If someone feels satisfied when they help others, they are likely to want to do it again.

GPT3 Prediction: (a) satisfaction (✗). GPT3 Explanation: the answer should be a feeling which would cause someone to help at the senior center. If someone feels satisfied after helping others, they are likely to continue doing so.

Q: What do you want someone to do when you illustrate point? Answer Choices: (a) did not understand (b) accepting (c) make clear (d) understood (e) understanding

T5 Prediction: (d) understood (✗). T5 Explanation: The answer should be something that you want someone to do when you illustrate a point. When you illustrate a point, you want the person to understand what you are trying to say.

GPT3 Prediction: (c) make clear (✗). GPT3 Explanation: The answer should be something that you want someone to do when you illustrate a point. When you illustrate a point, you want the other person to understand what you are trying to say.

(a) (b) (c)

Figure 4: T5 and GPT-3 generated explanations with incorrect predictions.

B Explanation examples

Here we further provide three examples as shown in Figure 4, where both T5 and GPT-3 have incorrect predictions. We observe that in both example (b) and (c), T5 and GPT-3 have plausible predictions and explanations although their predictions are different from golden labels.