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 (LLMs) is shown to elicit strong reasoning capabilities along with reasonable explanations. However, deploying them at scale is costly expensive in real-world applications, limiting their usage. In this paper, we propose a framework 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 standard finetuning baselines especially in few-shot settings by up to 8.1% accuracy, and even perform better than finetuning/prompting a 60x larger GPT-3 (175B) model [1] by up to 9.5% in accuracy. As a side benefit, human evaluation further shows that our method can generate competitive explanations to justify its predictions compared to strong GPT-3, moving towards the goal of explainable AI.

Introduction

Large language models (LLMs) have achieved impressive results with in-context learning; by adding a few demonstrations in 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; Sanh et al. 2022; Shi et al. 2022; Anil et al. 2023; OpenAI 2023). Recently, it is shown that explanation-augmented prompts can elicit strong performance in various reasoning tasks (Wei et al. 2022c; Lampinen et al. 2022), such as math word problem (Cobbe et al. 2021), symbolic reasoning (Wei et al. 2022c), numerical reasoning (Zhou et al. 2022) and commonsense reasoning tasks (Talmor et al. 2019). In addition, they also enable LLMs to generate reasonable explanations to justify the reasoning outcomes (Wiegreffe et al. 2021). However, these strong few-shot reasoning abilities only emerge when models scale to dozens or hundreds of billion of parameters (Wei et al. 2022b), making it costly expensive to deploy them at scale in real-world applications (Wei et al. 2022c).

Small language models (SLMs)[2] provide an alternative and could be more favorable over LLMs in many real-world applications due to their low cost in both storage and computation. Nevertheless, one important open question is how to close the performance gap between LLMs and SLMs on complicated reasoning tasks, as is observed in Wei et al. (2022b); Zelikman, Wu, and Goodman (2022), especially in few-shot settings (Li, Chen, and Yu 2019).

An intuitive way is to utilize explanations written by human as additional training signals to improve SLM reasoning capability. 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), and sometimes even irrelevant and meaningless, which restricts the amount of available high-quality explanations. As an example, explanations of CommonsenseQA (Talmor et al. 2019) collected by Rajani et al. (2019) include many meaningless and irrelevant explanations, e.g. "rivers flow trough valleys." and "this word is most relavant", that appear hundred of times in the training set.[3] On the other hand, using explanation-augmented prompts enables LLMs to automatically generate reasonable explanations (Wiegreffe et al. 2021), making it a plausible alternative to generate arbitrary amount of explanations quickly and cheaply. Therefore, a key question is: *If we utilize high-quality explanations generated by LLMs rather than the ones from human, can they improve the reasoning capability of SLMs?*

In this paper, we propose a framework leveraging explanations generated from LLMs to improve the reasoning capability of SLMs. Our framework is shown in Figure 1. Specifically, we first utilize several examples with human-written explanations as demonstrations for LLM and then generate explanations for the whole *training* set. After that, we adopt a

[1] We denote all GPT davinci series as GPT-3 in this paper and assume that their model sizes are 175B following Zhang, Gutiérrez, and Su (2023).

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

[3] https://github.com/salesforce/cos-e/issues/2

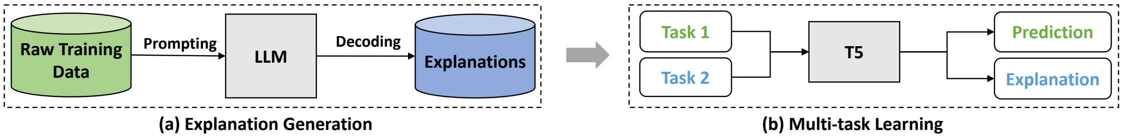


Figure 1: Overview of our framework. We first utilize examples with human-written explanations as demonstrations for LLM to generate explanations for the whole *training* set. We then adopt a multi-task learning framework to utilize the LLM-generated explanations, where one task is trained to generate predictions while the other is trained to generate explanations as additional supervison signals. During inference, models can generate both predictions and explanations with different task prompts.

multi-task learning setup to utilize the LLM-generated explanations to facilitate SLMs to acquire strong reasoning power together with explanation generation capabilities. Under this setup, one task is training SLMs to generate predictions the same as standard reasoning models while the other is training them to generate explanations as additional supervision signals. Such a setup enables the models not only to generate predictions but also to generate explanations to justify their predictions during inference. Experimental results show that our framework can consistently improve the reasoning capability of SLMs with multiple explanation generation approaches as well as different multi-task learning setups. 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* dataset. 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 building more explainable AI systems (Samek et al. 2019).

Related Work

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) propose multi-task learning with explanations for natural language inference tasks with LSTM and does not observe gains over standard single-task finetuning, i.e., direct predictions. Narang et al. (2020) utilizes a similar 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 explanation setups only obtain comparable results over standard finetuning baselines on T5. We instead show that our approach can improve SLMs across model sizes, explanation generation methods from LLMs, multi-task finetuning setups, and training data size consistently and significantly without accuracy-explanation trade-off (Jain et al. 2020).

Explanation-augmented Prompting. Recently, *in-context learning* has shown promising results in various NLP tasks (Brown et al. 2020). Although promising, LLMs still struggle with tasks requiring strong reasoning capability (Wei et al. 2022c). To enable better few-shot in-context learning of LLMs for reasoning tasks, Wei et al. (2022c) 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 (Cobbe et al. 2021), symbolic (Zhou et al. 2022) and common sense reasoning tasks (Geva et al. 2021). 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 LLMs 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 LLMs, where answers are fed into LLMs before providing their explanations in prompts, and also observes consistent gains. There is also existing work to utilize explanations generated from LLMs rather than focusing on their final predictions. Wiegreffe et al. (2021) explores utilizing LLMs 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 (Brown et al. 2020) as they find that GPT-3 tends to generate consistent but less factual explanations for textual reasoning tasks. However, these work do not explore if these noisy explanations generated from LLMs without human-involved filtering can be used to improve SLM reasoning capability.

Knowledge Distillation from LLMs. Knowledge distillation aims to transfer the knowledge from a large model to a small one that is easier for deployment with lower latency (Hinton, Vinyals, and Dean 2015). As LLMs become more capable (Brown et al. 2020; OpenAI 2023), there has been growing interests to distill knowledge from them to improve small models (Li et al. 2022; Sahu et al. 2022; Taori et al. 2023; Xu et al. 2023) or for self-improvement (Zelikman, Wu, and Goodman 2022; Wang et al. 2022; Huang et al. 2022). Recent work (Shridhar, Stolfo, and Sachan 2022; Ho, Schmid, and Yun 2022; Magister et al. 2022; Hsieh et al. 2023) also share similar motivation with ours to improve the reasoning capabilities of SLMs from LLMs. Shridhar, Stolfo, and Sachan (2022) trains two distilled models, one for problem decomposition and another for solving subproblems while our work uses a single model to solve problems directly.

Ho, Schmid, and Yun (2022); Magister et al. (2022); Hsieh et al. (2023) utilize single prompting methods to generate explanations along with single fine-tuning method to train student models, while our work systematically explores three different explanation generation methods along with three different fine-tuning setups. We further compare explanations from student model with its teacher through human evaluation, which is not done by Ho, Schmid, and Yun (2022); Magister et al. (2022); Hsieh et al. (2023).

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. 2022c). Chain of thought prompting provides intermediate reasoning steps as explanations 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. 2022c). Thus, we utilize ***C**hain **O**f **T**hought prompting with incorrect answer r**E**jection* (COTE) (Zelikman, Wu, and Goodman 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. 2021) 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 parsed and cast as explanation $\hat{e}_i$ without filtering.

CROP. COTE will possibly generate relatively high-quality explanations if LLM gives correct predictions of problems at hand as incorrect explanations tend to generate incorrect predictions (Wei et al. 2022c). 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, Wu, and Goodman 2022) and RP may provide a meaningful explanation as it can access golden label during explanation generation process.

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 utilize a multi-task learning with explanations since (1) it can naturally allow training with partially generated explanations and (2) Wiegreffe, Marasović, and Smith (2021) shows that self-rationalizing model, where golden label and human-written explanation is linearly concatenated as the target, performs significantly worse than MT counterpart (Hase et al. 2020). 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.

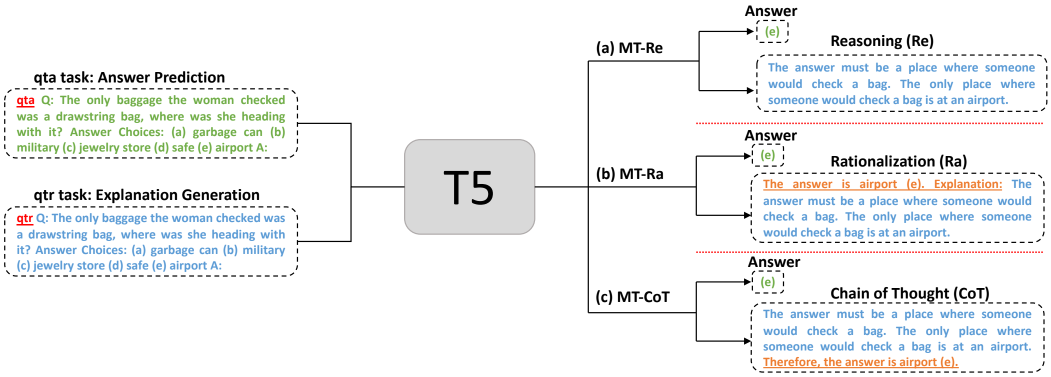

qta task: Answer Prediction
qta Q: The only baggage the woman checked was a drawstring bag, where was she heading with it? Answer Choices: (a) garbage can (b) military (c) jewelry store (d) safe (e) airport A:
qtr task: Explanation Generation
qtr Q: The only baggage the woman checked was a drawstring bag, where was she heading with it? Answer Choices: (a) garbage can (b) military (c) jewelry store (d) safe (e) airport A:
T5
(a) MT-Re
Answer
(e)
Reasoning (Re)
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.
(b) MT-Ra
Answer
(e)
Rationalization (Ra)
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.
(c) MT-CoT
Answer
(e)
Chain of Thought (CoT)
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.

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.

Experiments

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 requiring common sense reasoning with 9741/1221/1140 questions for training/development/test set, respectively. Since its test set is not publicly available, we report results on its development set following Zelikman, Wu, and Goodman (2022).

(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 [4], 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 OpenAI API [5] 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. 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 to train models with explanations based on T5 on NVIDIA RTX A6000. We implement multi-task learning (MT) 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 MT, we directly generate predictions from *qta* task for fair comparisons.

Main results

In this section, we compare results between multi-task learning with explanations and its single-task finetuning counterpart using full training data on three datasets. 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

[4] https://github.com/eladsegal/strategyqa

[5] 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.

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.

	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.

Few-shot learning results

We have shown the effectiveness of our method on full-training settings in *Main results* section 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 *Main results* section. 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, α* 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 α* 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.

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 *Main results* section 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 LLMs on CommonsenseQA (CSQA) and OpenbookQA (OBQA), and model sizes are listed in the parenthesis. Results with ◇ and * are from Zelikman, Wu, and Goodman (2022) and Xu et al. (2021), respectively.

Results comparison with LLMs

We further compare our method on T5-3B with state-of-the-art LLMs. Specifically, we adopt GPT-J direct finetuning, its self-bootstrapping version (STaR) (Zelikman, Wu, and Goodman 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. 2022c) 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 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.

Explanation comparison with LLM

A side benefit of our model is to generate explanations to alleviate the notorious black box issue of deep neural networks (Koh and Liang 2017). Our 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 competitive even compared to human-written ones in Wiegreffe et al. (2021).

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* with their agreement percentage on three levels.

Specifically, we perform a head-to-head explanation comparison on *CommonsenseQA* 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. We randomly sample 100 examples that are predicted correctly by both GPT-3 and T5 from *qta* task, and for each example, we present a question, its golden answer and two randomly shuffled and parsed explanations as (a) and (b) generated by GPT-3 and T5 from *qtr* task with greedy decoding 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. (2021). Finally, we adopt majority voting 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, only 37% explanations are in level 2 agreement and more than 60% explanations have disagreement (7% in level 0 + 56% in level 1). Given Wiegreffe et al. (2021) finds that 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. We provide several examples with analysis into Appendix B. Therefore, these results demonstrate that explanations generated by our model are competitive even compared to strong GPT-3 with 60x larger size.

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 various settings. In addition, human evaluation show that our model can generate competitive explanations even compared to GPT-3 175B towards more explainable AI.

Acknowledgements

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

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *NeurIPS*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *ArXiv*, abs/2110.14168.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Hase, P.; Zhang, S.; Xie, H.; and Bansal, M. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In *EMNLP (Findings)*, 4351–4367.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating Visual Explanations. In *ECCV*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, N.; Schmid, L.; and Yun, S.-Y. 2022. Large Language Models Are Reasoning Teachers. *arXiv preprint arXiv:2212.10071*.

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

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Huang, J.; Gu, S. S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4459–4473. Online: Association for Computational Linguistics.

Johnson, W. L. 1994. Agents that Learn to Explain Themselves. *AAAI*.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. *ArXiv*, abs/1703.04730.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *ArXiv*, abs/2205.11916.

Lampinen, A. K.; Dasgupta, I.; Chan, S. C. Y.; Matthewson, K.; Tessler, M. H.; Creswell, A.; McClelland, J. L.; Wang, J. X.; and Hill, F. 2022. Can language models learn from explanations in context? *ArXiv*, abs/2204.02329.

Li, S.; Chen, J.; and Yu, D. 2019. Teaching Pretrained Models with Commonsense Reasoning: A Preliminary KB-Based Approach. *ArXiv*, abs/1909.09743.

Li, Z.; Chen, W.; Li, S.; Wang, H.; Qian, J.; and Yan, X. 2022. Controllable Dialogue Simulation with In-context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4330–4347. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Magister, L. C.; Mallinson, J.; Adamek, J.; Malmi, E.; and Severyn, A. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 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*, 2381–2391. Brussels, Belgium: Association for Computational Linguistics.

Narang, S.; Raffel, C.; Lee, K.; Roberts, A.; Fiedel, N.; and Malkan, K. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. *ArXiv*, abs/2004.14546.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever., I. 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*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv*, abs/1910.10683.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4932–4942. Florence, Italy: Association for Computational Linguistics.

Sahu, G.; Rodriguez, P.; Laradji, I. H.; Atighehchian, P.; Vazquez, D.; and Bahdanau, D. 2022. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*.

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K.-R. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Raja, A.; Dey, M.; Bari, M. S.; Xu, C.; Thakker, U.; Sharma, S. S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N.; Datta, D.; Chang, J.; Jiang, M. T.-J.; Wang, H.; Manica, M.; Shen, S.; Yong, Z. X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Fevry, T.; Fries, J. A.; Teehan, R.; Scao, T. L.; Biderman, S.; Gao, L.; Wolf, T.; and Rush, A. M. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; Das, D.; and Wei, J. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. *ArXiv*, abs/2210.03057.

Shridhar, K.; Stolfo, A.; and Sachan, M. 2022. Distilling Reasoning Capabilities into Smaller Language Models. *arXiv preprint arXiv:2212.00193*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 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)*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N. M.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y.; Lee, H.; Zheng, H.; Ghafouri, A.; Menegali, M.; Huang, Y.; Krikun, M.; Lepikhin, D.; Qin, J.; Chen, D.; Xu, Y.; Chen, Z.; Roberts, A.; Bosma, M.; Zhou, Y.; Chang, C.-C.; Krivokon, I. A.; Rusch, W. J.; Pickett, M.; Meier-Hellstern, K. S.; Morris, M. R.; Doshi, T.; Santos, R. D.; Duke, T.; Søraker, J. H.; Zevenbergen, B.; Prabhakaran, V.; Diaz, M.; Hutchinson, B.; Olson, K.; Molina, A.; Hoffman-John, E.; Lee, J.; Aroyo, L.; Rajakumar, R.; Butryna, A.; Lamm, M.; Kuzmina, V. O.; Fenton, J.; Cohen, A.; Bernstein, R.; Kurzweil, R.; Aguera-Arcas, B.; Cui, C.; Croak, M.; Chi, E.; and Le, Q. 2022. LaMDA: Language Models for Dialog Applications. *ArXiv*, abs/2201.08239.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. *arXiv preprint arXiv:2212.10560*.

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022b. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022c. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*, abs/2201.11903.

Wiegreffe, S.; Hessel, J.; Swayamdipta, S.; Riedl, M. O.; and Choi, Y. 2021. Reframing Human-AI Collaboration for Generating Free-Text Explanations. *ArXiv*, abs/2112.08674.

Wiegreffe, S.; and Marasović, A. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *NeurIPS Datasets and Benchmarks*.

Wiegreffe, S.; Marasović, A.; and Smith, N. A. 2021. Measuring Association Between Labels and Free-Text Rationales. In *EMNLP*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; and Jiang, D. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. arXiv:2304.12244.

Xu, Y.; Zhu, C.; Wang, S.; Sun, S.; Cheng, H.; Liu, X.; Gao, J.; He, P.; Zeng, M.; and Huang, X. 2021. Human Parity on CommonsenseQA: Augmenting Self-Attention with External Attention. *ArXiv*, abs/2112.03254.

Ye, X.; and Durrett, G. 2022. The Unreliability of Explanations in Few-Shot In-Context Learning. *ArXiv*, abs/2205.03401.

Zelikman, E.; Wu, Y.; and Goodman, N. D. 2022. STaR: Bootstrapping Reasoning With Reasoning. *ArXiv*, abs/2203.14465.

Zhang, K.; Gutiérrez, B. J.; and Su, Y. 2023. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. *arXiv preprint arXiv:2305.11159*.

Zhou, D.; Scharli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; and Chi, E. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *ArXiv*, abs/2205.10625.

Appendix A

Here we provides prompts we use in our experiments. Our prompts on *CommonsenseQA* and *StrategyQA* datasets are based on (Zelikman, Wu, and Goodman 2022) and (Wei et al. 2022c), 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/cmˆ3, 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

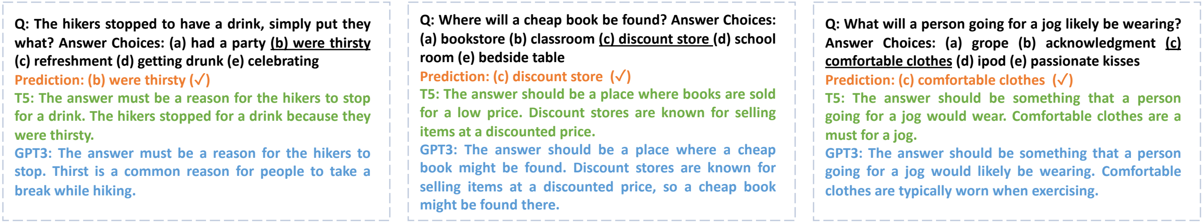
A: (d) that which contains 2 hydrogen and 1 oxygen molecules
Explanation: Water covers over 90% of the Earth's surface and 0% of the moon's surface, and contains 2 hydrogen and 1 oxygen molecules.

Table 13: Rationalization prompt for OpenbookQA.

Appendix B

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 in all three examples. Even though T5 loses to GPT-3 in example (c), its explanation is still reasonably good.

We also provide examples with incorrect predictions of both T5 and GPT-3 in 4. 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.



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.

(a) T5 wins

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.

(b) Tie

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

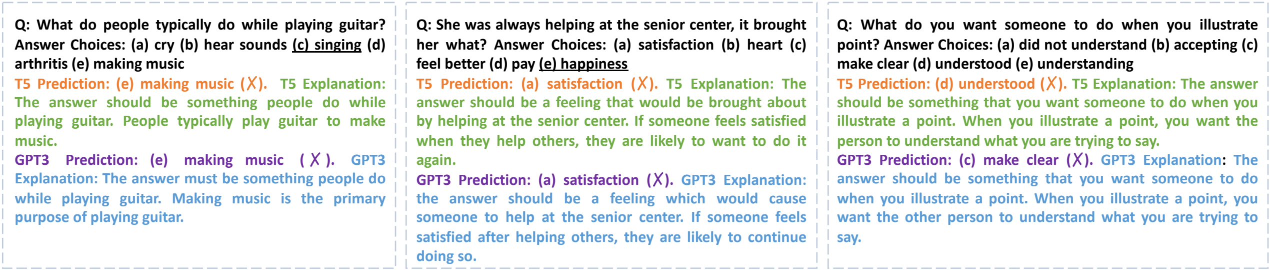
Prediction: (c) comfortable clothes (✓)

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.

(c) T5 loses

Figure 3: T5 and GPT-3 generated explanations used in human evaluation. In example (a), (b) and (c), T5 wins over, is tied with, loses to GPT-3, respectively.



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.

(a)

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.

(b)

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.

(c)

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