SCOTT: Self-Consistent Chain-of-Thought Distillation

Peifeng Wang[1][*] **Zhengyang Wang**[2], **Zheng Li**[2], **Yifan Gao**[2], **Bing Yin**[2], **Xiang Ren**[1]

[1]Department of Computer Science, University of Southern California, [2]Amazon.com Inc

{peifengw,xiangren}@usc.edu,
{zhengywa,amzzhe,yifangao,alexbyin}@amazon.com

Abstract

Large language models (LMs) beyond a certain scale, demonstrate the emergent capability of generating free-text rationales for their predictions via chain-of-thought (CoT) prompting. While CoT can yield dramatically improved performance, such gains are only observed for sufficiently large LMs. Even more concerning, there is little guarantee that the generated rationales are consistent with LM's predictions or faithfully justify the decisions. In this work, we propose a faithful knowledge distillation method to learn a small, self-consistent CoT model from a teacher model that is orders of magnitude larger. To form better supervision, we elicit rationales supporting the gold answers from a large LM (teacher) by contrastive decoding, which encourages the teacher to generate tokens that become more plausible only when the answer is considered. To ensure faithful distillation, we use the teacher-generated rationales to learn a student LM with a counterfactual reasoning objective, which prevents the student from ignoring the rationales to make inconsistent predictions. Experiments show that, while yielding comparable end-task performance, our method can generate CoT rationales that are more faithful than baselines do. Further analysis suggests that such a model respects the rationales more when making decisions; thus, we can improve its performance more by refining its rationales.

1 Introduction

Large language models (LMs) elicit strong reasoning capabilities through chain-of-thought (CoT) prompting (Wei et al., 2022b), which asks LMs to generate free-text rationale for explaining their multi-step reasoning. However, CoT prompting does not guarantee that the rationale is consistent with the prediction, rendering the rationale useless for justifying the model's behavior. In this work,
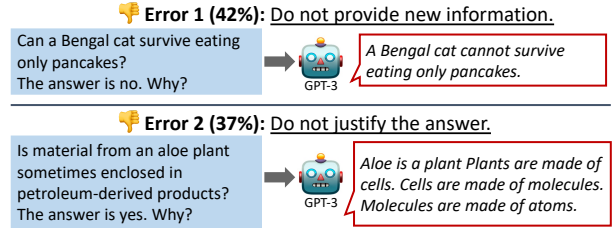

Figure 1: Vacuous rationales generated by a prompted LM (GPT-3) for StrategyQA. In both types of error cases, LM fails to give rationales consistent with the answers due to hallucination.

we present a knowledge distillation (KD) method for eliciting faithful CoT reasoning, where a small student model learns from a large teacher model to generate CoT rationales that are consistent to its own predictions.

Existing works (Shridhar et al., 2022; Li et al., 2022a) propose learning to reason from large LMs mainly for computation efficiency or task performance. They prompt a large LM (the teacher) to generate rationales for a downstream dataset, which is then used to train a small LM (the student). However, these works neglect the following two issues which could undermine the faithfulness of the rationales. First, LMs are prone to hallucination, meaning they often generate text that is not grounded by the input (Maynez et al., 2020; Ji et al., 2022). Therefore, the teacher may not generate on-topic rationales, which fully support the answer. In our pioneer study (Figure 1) over 100 random rationales generated by GPT-3, we found 42% of them not providing new information that is not stated in the task input and 37% of them not justifying the answer[1]. This inconsistency between the rationale and answer would then be inherited by the student. Second, the student may treat rationale generation and answer prediction as two independent processes. This is due to the spuri-

---

[1]Wiegreffe et al. obtains a similar observation on the rationales generated by GPT-3 for the CommonsenseQA dataset.
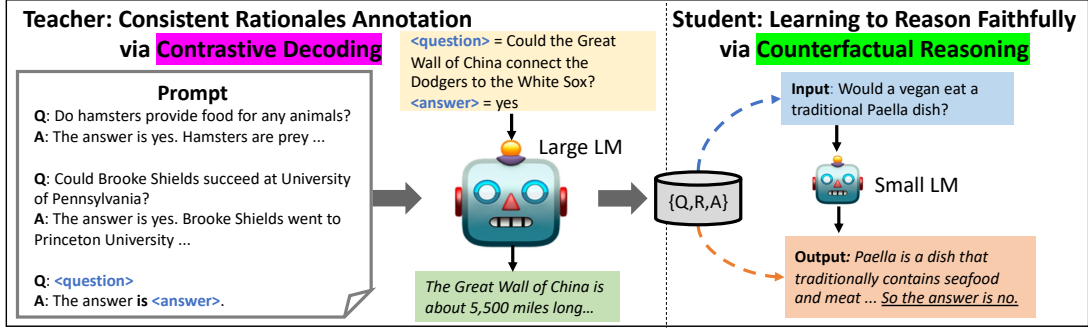
Figure 2: **Overview of our knowledge distillation framework for faithful reasoning.** (a) Teacher: A large LM prompted to generate a consistent rationale given a question and the gold answer in the training set via contrastive decoding. (b) Student: A small LM fine-tuned to generate a rationale and then answer via counterfactual reasoning.

ous correlations between the question and answer, which is exploited as a reasoning shortcut by the student (Branco et al., 2021). The two issues together would lead to an unfaithful student which learns to generate vacuous rationales and may make predictions inconsistent with the rationales.

To address these issues, we propose to enhance the vanilla KD process from two ends respectively. To elicit more on-topic rationales from the teacher, we propose to leverage contrastive decoding which aims to ground each rationale to the answer (§ 3.1). This technique encourages the teacher to generate tokens that are more plausible only when the answer is considered instead of the ones that are fairly plausible even without the answer during decoding. To train a faithful student, we ask the student to conduct counterfactual reasoning, i.e., predicting accordingly when the rationales are leading to different answers (§ 3.2). We obtain the training data by asking the teacher to generate a rationale for a sampled incorrect answer. The reasoning shortcut between the question and the gold answer is thus removed since now the student needs to give a different answer for the same question, according to the rationales provided during training.

We conduct experiments on several open-domain question answering tasks that require knowledge-intensive reasoning. Experiments show that: (1) Contrastive decoding can lead to a more consistent teacher which generates rationales that are more supportive of the gold answers. (2) Trained on the more consistent rationale-answer pairs, the student learns to better associate the answer prediction with the rationale generation. (3) With counterfactual reasoning as an auxiliary training objective, the student learns not to take the reasoning shortcut and instead respect the rationale more. (4) Despite being more faithful, our model performs comparably to the baselines. (5) Ablation study shows that

although performing better, larger student models are more prone to being inconsistent. Our method robustly remedies the inconsistency regardless of the size of the student model. (6) With a more faithful student, we can better improve its performance by correcting its rationale, demonstrating the utility of our method in model refinement.

2 Chain-of-Thought Distillation

Our goal is to 1) elicit consistent rationales, i.e., those well justifying the gold answers, from a large LM as supervision, and then 2) train a self-consistent student model to reason faithfully, i.e., answer accordingly to its generated rationale. We consider the task of language-based reasoning where the required knowledge is not provided in the task input. Specifically, we focus on open-domain question answering (QA) which is the most general setting adopted by prior works: given a question q, a QA system is asked to predict the gold answer a^*. For interpretability, we also require the model to provide a free-text rationale r, which justifies its prediction. Below we describe the overview of a vanilla KD framework as illustrated in Figure 2. We then discuss the limitations and propose our method in § 3.

2.1 Generating Rationale Annotation

Instead of asking humans to annotate a rationale for each question-answer tuple $\{q, a^*\}$, we obtain the rationale from a teacher model automatically using in-context learning. The idea is to prompt a frozen LM as the teacher with only a few annotated examples as demonstration before a new instance is provided. Each example consists of a question q randomly sampled from the training set, the gold answer a^* and a human-annotated rationale r which justifies why a^* is correct. The prompt p is structured in the format as shown in

Figure 2 (the Prompt in the left part). To obtain the rationale for a new question q, one basic strategy could be greedy decoding, which selects the most plausible token at each step:

$$t_i^* = \arg\max P(t_i | p, q, a^*, t_{<i}). \qquad (1)$$

2.2 Training a Student Model

Now with the annotated training data $\{q, r, a^*\}$, we can train a smaller model as the student. There are many ways to implement a QA model that can make a prediction as well as generate a rationale. In this work, we focus on the self-rationalization paradigm, where the student firstly generates a rationale and then predicts the answer conditioning on the generated rationale. This is in contrast to related works which conduct post-rationalization, i.e., generating the rationale after the answer is predicted, or multi-task learning, which treats rationale generation as an auxiliary task besides answer prediction. The reason is that the generation of the rationale for the latter two paradigms does not affect the decision making by design, and therefore the faithfulness of the rationale is not guaranteed in the first place.

Given a question q, the student model is trained to output a sequence of rationale tokens concatenated with the answer tokens as shown in Figure 2 (the output in the right part). One straightforward implementation is simply fine-tuning a text-to-text LM over the silver training data generated by the teacher using standard language modeling loss:

$$\mathcal{L}_{factual} = -\sum_i \log P(t_i | q, t_{<i}), \qquad (2)$$

which we refer as factual reasoning loss.

3 Distilling a Self-Consistent Student

There are two vital issues with the vanilla KD process described in the previous section. Firstly, neural LMs are known to suffer from the issue of hallucination, meaning they often generate text that is not grounded by the input (Maynez et al., 2020; Ji et al., 2022). This would lead to the generated rationale not supporting the given answer. The inconsistency between the rationale and the answer would then be inherited by the student, which is misled to think that the answer prediction is independent of the rationale generation. Secondly, the student model would learn to predict the answer by taking a reasoning shortcut (Branco et al.,

Figure 3: **Contrastive decoding** for obtaining rationales that are more grounded by the gold answers, by preferring tokens that are more plausible only when the answer is considered.

2021), without taking into account the generated rationale (even though the answer prediction is conditioned on the rationale). This is due to the spurious correlations between the question and the answer which are found in various implicit reasoning task datasets (Gururangan et al., 2018; Zellers et al., 2019; Blodgett et al., 2020).

The two issues mentioned above would result in an untrustworthy student whose generated rationales do not consistently justify its answers. To mitigate this, we propose two corresponding techniques as detailed below.

3.1 A Consistent Teacher: Contrastive Decoding

To encourage the teacher to generate a more on-topic rationale that supports the answer, our proposed method extends a prior technique called contrastive decoding for open-ended text generation (Li et al., 2022b). The core idea is to search rationale tokens that are more plausible only when the answer is considered instead of the ones that are fairly plausible even without the answer during decoding. To implement this idea, we firstly model the hallucinating behavior by providing a perturbed answer a' to the same teacher and then obtain the plausibility growth of any token t_i given the answer a^* as

$$G(t_i | a^*) = \log \frac{P(t_i | p, q, A, a^*, t_{<i})}{P(t_i | p, q, A, a', t_{<i})}. \qquad (3)$$

We investigate two ways of perturbing the answer: setting a' as an empty string or an incorrect answer

other than a^{*}[2]. The first way (with an empty string) punishes tokens that are generally plausible when the gold answer a^* is not considered by a hallucinated LM. The second way (with an incorrect answer) takes a step further by encouraging the teacher to generate a rationale that is more distinctive between gold and wrong answers. Figure 3 shows the generations for an example question from greedy decoding and contrastive decoding.

To strike a balance between language fluency and the grounding with a^*, we incorporate the plausibility growth into Eq. 1 by aggregation as our final contrastive decoding strategy:

$$t_i^* = \arg\max P(t_i | p, q, A, a^*, t_{<i}) + G(t_i | a^*) \quad (4)$$

3.2 A Faithful Student: Counterfactual Reasoning

To encourage the student to reason faithfully towards its generated rationale, we train the student to conduct counterfactual reasoning (Roese, 1997), i.e., answer accordingly when the rationale is leading to a different answer. This would help remove the reasoning shortcut between a question and the gold answer (Figure 4) since now the student is asked to answer differently for the same question. To implement this idea, we firstly replace the gold answer fed to the teacher in Eq. 4 with a wrong answer a' randomly (with the same sampling strategy as in § 3.1) as if a' is correct. We thus obtain a counterfactual rationale r' that leads to the wrong answer a'. We then train the model to generate a' when r' is directly fed to the decoder as teacher-forcing:

$$\mathcal{L}_{counterfactual} = -\sum_i \log P(t_i | q, r', t_{<i}). \quad (5)$$

To avoid confusing the student about the task, we indicate the training objective Eq. 2 (or Eq. 5) to the student by appending the keyword `[Factual]` (or `[Counterfactual]`) at the beginning of both the input sequence to the encoder and the output sequence to the decoder (see Figure 4 for an example input and output). The overall training loss is the sum of Eq. 2 and Eq. 5.

4 Experiments

We aim to answer the following research questions in our experiments: (1) Can our contrastive de-
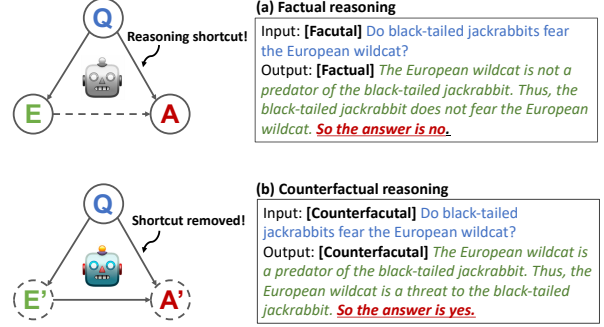

Figure 4: **Counterfactual reasoning** for teaching the student to reason faithfully, i.e., answer differently according to the rationale.

coding strategy lead to a more consistent teacher? (2) Can a more consistent teacher and the counterfactual reasoning objective lead to a student that reasons more faithfully? (3) Can we have more control over a self-consistent student's predictions by modifying its generated rationales?

4.1 Datasets

We experiment with several language-based reasoning tasks that are knowledge-intensive: (1) CSQA (Talmor et al., 2018) is a five-choice QA dataset that tests general commonsense about the daily concepts. (2) StrategyQA (Geva et al., 2021) is a binary (yes/no) QA dataset where the required reasoning steps are implicit in the question. (3) CREAK (Onoe et al., 2021) is a fact-checking (true/false) dataset which tests commonsense reasoning about entity knowledge. (4) QASC (Khot et al., 2020) is an eight-choice QA dataset which requires both knowledge facts retrieval and the common sense for composing the facts. Since the test labels for these datasets are not publicly available, we treat the official development set as our test set, while randomly splitting the official training set into a new training set and development set.

4.2 Evaluation Metrics

(1) To evaluate the consistency between the rationales generated by the teacher and the gold answers provided as input, we use the LAS metric (Hase et al., 2020), whose core idea is to measure how well the rationales assist a simulator to predict the *gold answers* a^*, computed as the difference between the task performance when the rationale is provided as input vs. when it is not, namely $Acc(qr \to a^*) - Acc(q \to a^*)$. (2) To evaluate the faithfulness of the rationales generated by the student, we use LAS to measure how well the ra-
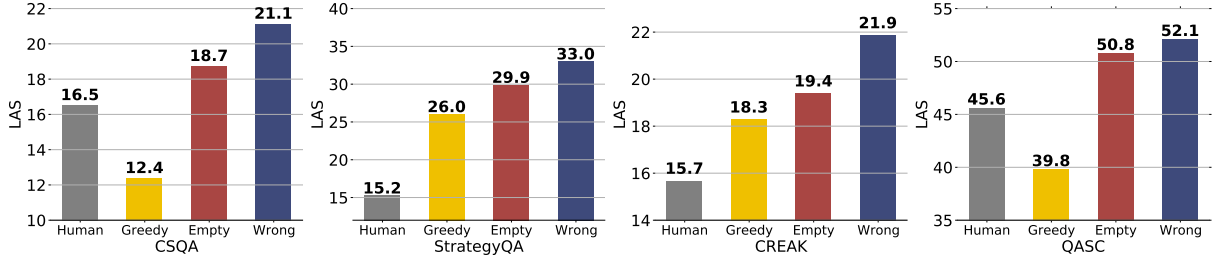
Figure 5: Simulatability (LAS) of the rationales generated from different teacher models as a measurement the consistency between the rationales and the gold answers. {*Greedy, Empty, Wrong*} refer respectively to using greedy decoding, contrastive decoding with empty/wrong answer to obtain rationale tokens from the teacher.

tionales help a simulator to predict *a student's predictions* a', namely $Acc(qr \rightarrow a') - Acc(q \rightarrow a')$. We implement each simulators with a fine-tuned T5-large model (Raffel et al., 2020) respectively. (3) To evaluate how well the student preserves its task performance on the downstream datasets, we use accuracy as the metric.

4.3 Implementation Details

We use GPT-neox (Black et al., 2022), a LM with 20B parameters as the teacher since the model checkpoint is publicly available, which allows us to host it offline and have access to token-wise probabilities as required in our contrastive decoding. We then implement two teacher variants by using an empty string or a wrong answer as the perturbed answer a' in Eq. 4 respectively. The obtained rationales are then used to fine-tune two T5-3b LMs as the students respectively. For both variants, we train the student using the sum of factual training loss Eq. 2 and counterfactual training loss Eq. 5.

4.4 Baselines

Chain-of-Thought (CoT) Since we elicit the rationales from GPT-neox (with 20b parameters) (Black et al., 2022) to train the student, we prompt the same model (GPT-neox) to firstly explain and then predict using CoT prompting (Wei et al., 2022b).

Learn from Human To demonstrate the advantage of our automatic way of generating rationale annotations, we implement this baseline as a fine-tuned T5-3b LM over human-annotated rationales, which are expensive to obtain and could be noisy.

Learn from Greedy Decoding We implement this baseline as a fine-tuned T5-3b LM over the rationales obtained by greedy decoding using the same LM as our main method. We also implement another variant by adding the counterfactual reasoning loss when fine-tuning the student, where the rationales for the wrong answers are obtained by greedy decoding.

Table 1: Human evaluation on the rationales generated by different teacher models for StrategyQA. A fair level of agreement measured by Fleiss Kappa (κ=0.26) is obtained among three annotators.

Teacher Model	Grammaticality	New Info	Supports Answer
Greedy	**0.99**	0.65	0.48
Contrast.-Empty	0.97	0.77	0.58
Contrast.-Wrong	0.97	**0.82**	**0.63**

We also implement two baselines of our method by training the student with the rationales obtained by contrastive decoding with empty/wrong answers based on factual reasoning only. We run all the experiments for 5 times using a fixed set of random seeds and report the average results.

4.5 Main Results

Can contrastive decoding lead to a more consistent teacher? Figure 5 shows the consistency between the rationales generated by different teachers and the gold answers measured by LAS. Across four datasets, contrastive decoding with either empty or wrong answers yield more consistent rationales compared to human annotation and greedy decoding. This demonstrates the effectiveness of our contrastive decoding strategy in encouraging the teacher to generate more on-topic rationales. Moreover, using wrong answers is better than using empty strings for contrastive decoding. This shows that by contrasting with the wrong answers, the teacher can generate more distinguishable rationales that lead to the gold answers, thus obtain higher consistency. Greedy decoding yields less consistent rationales compared to human annotation, verifying our claim that LMs are prone to generating text not grounded by the gold answers.

We also conduct a human evaluation over 100 rationales generated by different decoding strategies for StrategyQA. Annotators are asked to judge the rationales by 3 dimensions: 1) Grammaticality (Is the rationale grammatical?) 2) New Info (Does the
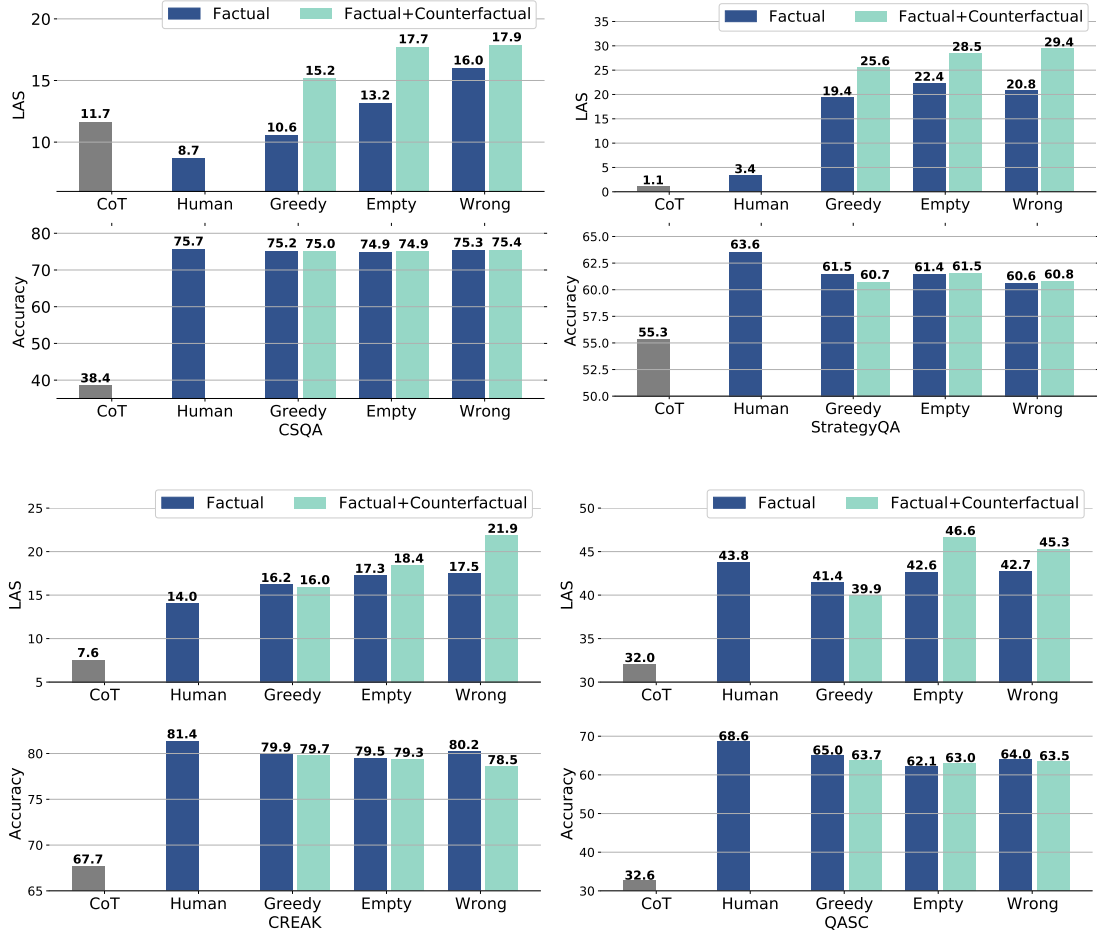
Figure 6: Faithfulness (LAS) and task performance (accuracy) of the compared methods on the experimented datasets. The x-axis represents the CoT baseline and knowledge distillation methods that use Human Annotation, Greedy Decoding, and Contrastive Decoding with empty strings/wrong answers as the teachers. For knowledge distllation methods, we apply Factual training and Factual+Counterfactual training to train the students.

rationale provide new information not expressed in the question?) 3) Supports Answer (Does the rationale justify the answer?). Table 1 confirms that our two contrastive decoding strategies yield more informative and on-topic rationales than greedy decoding, with a slightly worse grammaticality. We list examples in Table 2 (appendix) to showcase how rationales from contrative decoding are more consistent with gold answers than greedy decoding.

Can a more consistent teacher train a more faithful student? Figure 6 (upper parts of each sub-figure) shows the faithfulness of the students measured in LAS on the experimented datasets. First, the CoT method often achieves much lower LAS compared to the KD methods across four datasets, showing that the generated rationales do not faithfully reflect the decision making in CoT. Second, we observe that students trained with the rationales from contrastive decoding with either empty strings or wrong answers generally achieve higher LAS scores compared to the baselines. To-

gether with the observation on the consistency of the teacher (Figure 5), this validates that a more consistent teacher train a more faithful student and the inconsistency in the training data generated by the teacher will be inherited by the student.

Can couterfactual reasoning loss further improve the faithfulness? Figure 6 shows the students fine-tuned additionally with counterfactual training loss achieve higher faithfulness than their counterparts which are fine-tuned with factual training only. This validates that counterfactual reasoning can further improve the student's faithfulness, as it may still treat rationale generation and answer prediction as two independent processes.

Can a faithful student still preserve its performance? Figure 6 (lower parts of each sub-figure) shows the performance of the students measured in accuracy. First, CoT methods achieve lower accuracy compared to the KD methods, showing the benefit of combining the supervision from the teacher (the rationales) and the labeled datasets
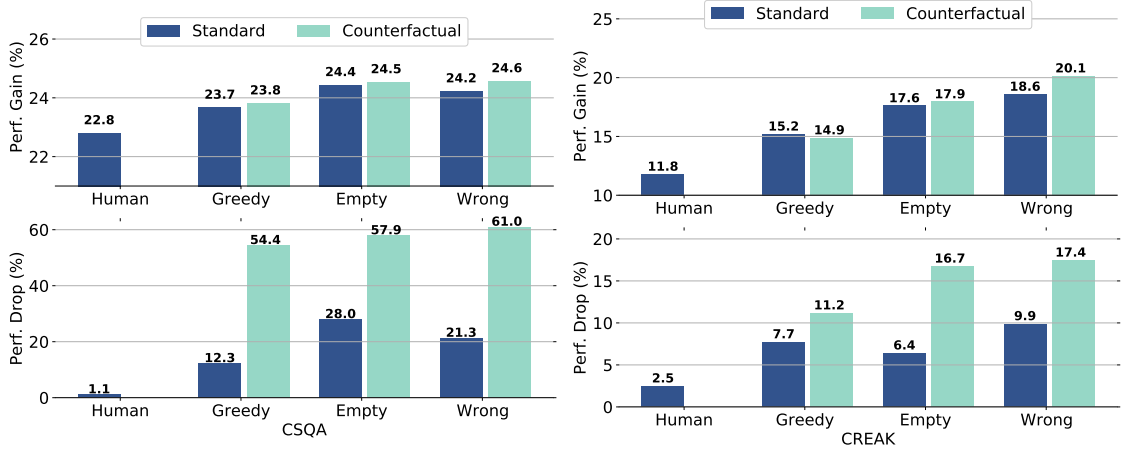
Figure 7: Performance gain (drop) of the compared methods when the oracle (perturbed) rationales are fed to the decoder of the model on CSQA and CREAK.

(the answers). Second, all the KD methods achieve comparable performance. Together with the observation over faithfulness, this demonstrates our method can improve faithfulness of the model while not hurting its performance. Note that the student which learns from human annotation achieves slightly better results compared to other students. This is because the human rationales are less consistent with the answers (as evidenced in Figure 5). Therefore, the student learns to generate the rationales and predict the answers more independently, which allows it to exploit the spurious correlation and achieve better performance. Our further analysis (§ 4.7) shows that such performance gain is suspicious as changing the rationales does not change the student's predictions mostly.

4.6 Ablation on the student model size

We ablate the student model size to see how its faithfulness and performance are affected. From Figure 8, we observe that larger student models achieve higher performance but lower faithfulness. This confirms that it requires sufficient capacity for storing knowledge necessary for reasoning (Wei et al., 2022a), but larger models are also better at answering the questions independently of the rationales. Still, our models are more faithful than baselines and comparable in performance with different model sizes.

4.7 Controlling the behavior of the Student

One important utility of faithful rationales is that we can have more control over the behavior of the student via changing its rationales. If the model can make predictions consistent with its rationales, we can either impair or improve the its performance by perturbing or refining its rationales. To verify this,
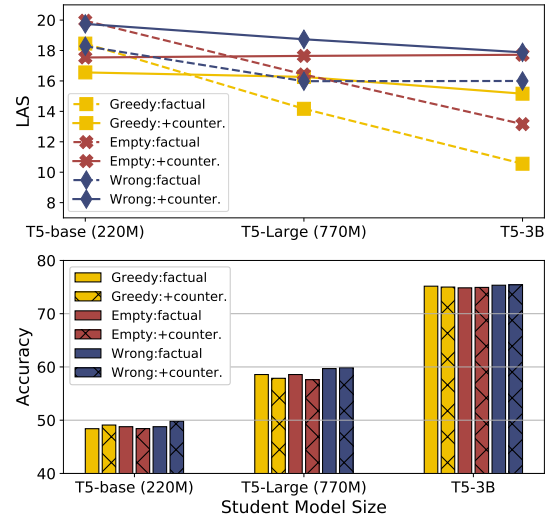

Figure 8: Faithfulness (LAS) and task performance (accuracy) of the compared methods with different student model sizes. Each model is named by the teacher it learns from and the training objective as `teacher model:training objective`.

we conduct two types of edition to the rationales generated by the student, namely perturbation and refinement as described below. We then feed the edited rationales to the decoder of the student directly (as teacher forcing) and see if the student will act accordingly, i.e., predict more badly (or accurately) due to the worse (or better) rationales.

Rationales Perturbation For perturbing the rationales, we randomly replace 50% of the tokens in the generated rationales from the student and then feed the perturbed rationales r' back to the decoder of the student. We finally calculate the performance drop (or sensitivity), i.e., $Acc(qr \rightarrow a^*) - Acc(qr' \rightarrow a^*)$. Figure 7 (the lower parts) shows the results on CSQA and CREAK. First, perturbing the rationales from the student that is fine-

tuned with human-annotation has little (down to 1.1% on CSQA) impact on its performance, meaning that the student largely ignores the rationales when making prediction. Second, learning from rationales obtained by contrastive decoding with empty or wrong answers leads to a student that is more sensitive to the rationale perturbation compared to learning from greedy decoding. This again verifies the necessity of having a consistent teacher in order to train a faithful student. Lastly, our counterfactual training loss further improves the sensitivity of the student, demonstrating that the student is more faithful towards the rationales.

Rationales Refinement As a proxy refinement, we obtain the oracle rationales r^* automatically by asking the teacher to rationalize for gold answers using each compared decoding strategy. For the student trained with human annotation, we directly use the annotated rationales as the oracle. We then calculate the performance gain, i.e., $Acc(qr^* \rightarrow a^*) - Acc(qr \rightarrow a^*)$. Figure 7 (the upper parts) shows the results on CSQA and CREAK. First, we observe that oracle human-annotated rationales do not bring as much performance gain as machine-generated rationales do. This demonstrates that even trained with human annotation, the student is still prone to being unfaithful to its rationales. Second, we observe that contrastive decoding (with either empty strings or wrong answers) leads to higher performance gains from the student. By adding counterfactual training, the performance gains are further increased. This demonstrates the advantage brought by our method, which is that we can have more success in debugging a reasoning model by refining its rationales.

5 Related Works

Free-text Rationales A variety of datasets have been proposed to collect human-annotated rationales alongside each task instance (Camburu et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021), aiming to train the downstream models to explain their predictions in natural language. However, human annotation is expensive and the resulting rationales are reported to be of poor quality (Aggarwal et al., 2021; Sun et al., 2022). Our work leverages a prompted LM to obtain rationales automatically for supporting both correct and incorrect answers, using only a few annotated examples as demonstration. The rationales for supporting the incorrect answers further enable the student to conduct coun-

terfactual reasoning, which is not available from existing human annotation.

Prompted Self-Rationalization Models Recent works have been proposed to prompt large LMs to generate a free-text rationale before making the prediction (Nye et al., 2021; Wei et al., 2022b). However, this technique relies on extremely large LMs (with over 100B parameters) to work effectively (Wei et al., 2022b,a), which requires significant computation resources or expensive API calls (Shridhar et al., 2022). Meanwhile, the rationales generated by such models are shown to contradict the context (Ye and Durrett, 2022) and fail to faithfully represent the underlying reasoning process (Wang et al., 2022). In contrast, our student is trained to be more faithful towards its generated rationales using a smaller LM.

Knowledge Distillation There exist some works that explore the idea of distilling rationales from a large LM to a small LM as the student. Eisenstein et al. proposed to train the student to extract the sentence containing the answer, which is not applicable to reasoning tasks that require background knowledge. Shridhar et al. proposed to train the student to ask and answer sub-questions necessary for decomposing the main question, which is tailored to solve math word problems (Cobbe et al., 2021) with an equation generator for guiding the student while we do not have such a constraint. Li et al. proposed to train the student on the joint task of generating the answers and the rationales, which only act as a regularization and do not affect the student's prediction during inference. More importantly, both Shridhar et al. and Li et al. do not consider the faithfulness of the rationales, which is critical for examining the behavior of the student.

6 Conclusion

This work presents a faithful KD framework for learning a small, self-consistent CoT model from a large teacher model. To ensure the student reason faithfully, we propose (1) contrastive decoding for obtaining a consistent teacher and (2) counterfactual reasoning for teaching a faithful student. Experiments show that these two techniques jointly lead to a more faithful student compared to the baselines, while preserving much performance accuracy. Our further analysis shows that changing the rationales has a larger impact on the student's behavior and thus we can have more success in debugging the model by refining its rationales.

Limitations

Compared to a standard knowledge distillation process, our method requires additional computation when preparing training data and training the student. First, our contrastive decoding needs to perform forward pass in the teacher model one time more than greedy decoding does to obtain the perturbed plausibility for each token generated (Eq. 4). Second, our KD process introduces additional training data for training the student with the counterfactual reasoning objective (Eq.5). Besides computation cost, this work focuses on improving faithfulness of the rationales rather than performance, which is complementary to prior works which leverages rationales for improving the performance only.

Ethics Statement

Our KD process leverages large LMs to obtain rationale annotation, which may expose social bias encoded in these models (Lucy and Bamman, 2021). The bias may be further inherited by the student model. Nevertheless, our method improves the faithfulness of the rationales, making the predictions from the student accountable. Without the faithful rationales, it would be unclear to users about whether the model is making predictions based on some unintended bias.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Workshop on Commonsense Reasoning and Knowledge Bases*.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Ruben Branco, António Branco, Joao Rodrigues, and Joao Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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

Jacob Eisenstein, Daniel Andor, Bernd Bohnet, Michael Collins, and David Mimno. 2022. Honest students from untrusted teachers: Learning an interpretable question-answering pipeline from a pretrained language model. *arXiv preprint arXiv:2210.02498*.

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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

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? *arXiv preprint arXiv:2010.04119*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022a. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022b. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. *OpenReview*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Neal J Roese. 1997. Counterfactual thinking. *Psychological bulletin*, 121(1):133.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions. *arXiv preprint arXiv:2212.00193*.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. *arXiv preprint arXiv:2206.11083*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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 preprint arXiv:2201.11903*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning. *arXiv preprint arXiv:2205.03401*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

A Appendix

A.1 Example generations from teachers

Table 2 shows examples about how rationales generated by contrastive decoding with empty/wrong answers are more on-topic than rationales generated by greedy decoding.

Table 2: Examples where rationales generated by contrastive decoding with empty/wrong answers are rated higher than rationales generated by greedy decoding in our human evaluation.

Q: Do workers at Nissan's headquarters eat with chopsticks?
A: Yes.
Greedy: Workers at Nissan's headquarters eat with chopsticks. Chopsticks are used in China. Thus, workers at Nissan's headquarters eat with chopsticks.
Contrastive-Empty: Workers at Nissan's headquarters are Japanese. Japanese eat with chopsticks.
Contrastive-Wrong: Nissan's headquarters is in Japan. Japan is a country where people eat with chopsticks. Thus, workers at Nissan's headquarters eat with chopsticks.

Q: Would Constitution of the United States paper offend PETA?
A: Yes.
Greedy: PETA is an animal rights group. The Constitution of the United States is a document that protects the rights of animals. Thus, PETA would be offended by the Constitution of the United States.
Contrastive-Empty: PETA is against the use of animal products. The Constitution of the United States is written on animal products. Thus, PETA would be offended by the Constitution of the United States.
Contrastive-Wrong: PETA is against the use of animal products. The Constitution of the United States is written on animal products. Thus, PETA would be offended by the Constitution of the United States.

Q: Does Sam Harris worship Shiva?
A: No.
Greedy: Sam Harris is a Hindu. Shiva is a Hindu god. Thus, Sam Harris does not worship Shiva.
Contrastive-Empty: Shiva is a Hindu god. Sam Harris is an atheist. Thus, Sam Harris does not worship Shiva.
Contrastive-Wrong: Shiva is a Hindu god. Sam Harris is an atheist. Thus, Sam Harris does not worship Shiva.

Q: Are a dozen pickles good for easing hypertension?
A: No.
Greedy: A dozen pickles have about 1,000 calories. A person needs about 2,000 calories a day. Thus, a dozen pickles would not be good for easing hypertension.
Contrastive-Empty: Pickles are acidic. Acidic foods are not good for easing hypertension.
Contrastive-Wrong: The recommended daily intake of sodium is 2,400 mg. A dozen pickles have about 0.5 g of sodium. Thus, a dozen pickles would not be good for easing hypertension.

Q: Would it be unusual to see frost in September in Texas?
A: Yes.
Greedy: Frost can occur in September in Texas. However, it is unusual.
Contrastive-Empty: Frost occurs in the winter. Thus, it would be unusual to see frost in September in Texas.
Contrastive-Wrong: Frost is usually seen in the winter. Thus, it would be unusual to see frost in September in Texas.