# Teaching Small Language Models Reasoning through Counterfactual Distillation

**Tao Feng**[1], **Yicheng Li**[1], **Chenglin Li**[1], **Hao Chen**[2], **Fei Yu**[2], **Yin Zhang**[1]*
[1]Zhejiang University, Hangzhou, China
[2]Ant Group, Hangzhou, China
{tao_feng, yichengli, chenglinli, zhangyin98}@zju.edu.cn
chuhu.ch@antgroup.com, feiyu.fyyu@gmail.com

## Abstract

With the rise of large language models (LLMs), many studies are interested in transferring the reasoning capabilities of LLMs to small language models (SLMs). Previous distillation methods usually utilize the capabilities of LLMs to generate chain-of-thought (CoT) samples and teach SLMs via fine-tuning. However, such a standard distillation approach performs poorly when applied to out-of-distribution (OOD) examples, and the diversity of the generated CoT samples is insufficient. In this work, we propose a novel counterfactual distillation framework. Firstly, we leverage LLMs to automatically generate high-quality counterfactual data. Given an input text example, our method generates a counterfactual example that is very similar to the original input, but its task label has been changed to the desired one. Then, we utilize multi-view CoT to enhance the diversity of reasoning samples. Experiments on four NLP benchmarks show that our approach enhances the reasoning capabilities of SLMs and is more robust to OOD data. We also conduct extensive ablations and sample studies to understand the reasoning capabilities of SLMs.

## 1 Introduction

Large language models (LLMs) have demonstrated remarkable performance in a wide range of downstream tasks (Brown et al., 2020; Wei et al., 2021; Hoffmann et al., 2022). Recent works have shown that chain-of-thought (CoT) can elicit reasoning capabilities in LLMs by asking the model to incorporate intermediate reasoning steps while solving a problem (Kojima et al., 2022; Wang et al., 2022). However, the efficacy of prompt-based CoT methods is restricted to very large models (beyond 10B parameters) (Wei et al., 2022). Due to the substantial computational resources or expensive API calls required for accessing LLMs that support CoT, various studies have delved into distilling the reasoning ability of LLMs into SLMs (Li et al., 2023b; Ho et al., 2022). Existing works focus on CoT distillation, a method that utilizes the CoT rationales of LLMs as supervision for training SLMs, rather than just labels (Shridhar et al., 2023; Wang et al., 2023b; Chen et al., 2023a; Zhao et al., 2023). The SLMs improve their reasoning capabilities by imitating the reasoning process of the LLMs. Although the CoT distillation method has been proven to be effective, it has the following weaknesses:

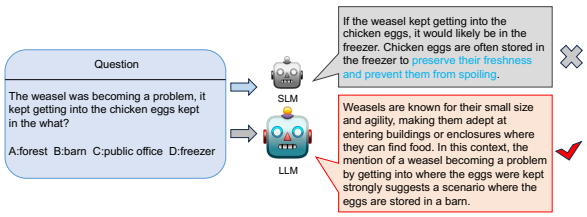


Figure 1: Rationales generated respectively by the LLM and the SLM with CoT distillation on common sense reasoning data.

(1) SLM is constrained by its own capacity and the scale of the annotated data. In the process of imitating LLM reasoning, it may learn spurious correlations, which leads SLM to merely remember the keywords and patterns in the training data, without understanding causal features. Therefore, this approach is typically effective for data from the same distribution, but it struggles when dealing with OOD data. In Figure 1, the rationales generated by the SLM indicate that it has learned the correlation between "egg kept" and "freezer" through CoT distillation. However, during inference, it completely ignores the context and directly links these two phrases together, lacking a deep understanding of the text semantics.

(2) Large language models typically only generate rationales for options they consider worth attention, rather than from the perspective of

* Corresponding author: Yin Zhang.



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842

small language models. As shown in Figure 1, the LLM explains why "barn" is the correct answer but does not mention the other options. Given the rich knowledge and reasoning capabilities of LLMs, they may perceive that the other three options are not substantial enough to provoke controversy. However, SLMs with limited capacity do not possess this ability. Furthermore, this knowledge is crucial for SLMs because they can utilize it through the process of elimination to determine the answer. In the process of answering questions, humans also adopt a similar approach. When individuals are confident that a particular option is highly accurate, they typically tend to choose directly. However, in situations where there is uncertainty about the answer, employing the method of exclusion becomes an effective strategy.

To address these issues, we propose to enhance the standard CoT distillation process from two aspects respectively. In order to mitigate the reliance of SLMs on spurious correlations (Calderon et al., 2022; Li et al., 2023a; Deng et al., 2023), we propose counterfactual data augmentation to encourage SLMs to better learn the causal structure of the text. Specifically, we utilize LLMs to edit task instances by adjusting causal relationships in the instances or modifying text segments that are crucial for label assignments. First, we use off-the-shelf language processing tools and LLMs to obtain the parts of the text that need to be edited. Then, prompt engineering and in-context learning are applied with LLMs to generate and filter counterfactual data, which closely resembles the original input but with altered answers. Additionally, we employ the multi-view CoT to introduce more diversified knowledge for SLMs, including both positive view CoT (PVC) and negative view CoT (NVC). PVC denotes the standard CoT generated by LLMs when answering questions, while NVC requires LLMs to generate refutational rationales for each incorrect option, explaining why the current option is considered erroneous.

We conduct experiments on four question answering tasks that require knowledge-intensive reasoning. Experiments demonstrate that: (1) Both counterfactual data augmentation and multi-view CoT are beneficial to improving model performance. (2) On the same distribution dataset, the proposed approach significantly outperforms standard CoT distillation, with an average improvement of 11.43%. (3) Across various parameter scales (ranging from 120M to 770M) and model structures (from decoder-only to encoder-decoder) for small models, our method consistently shows improved performance. (4) Compared to the baseline model, our approach demonstrates robust generalization on OOD data. Furthermore, extensive experiments indicate that our method enhances the reasoning capabilities of SLMs.

## 2 Related Work

### 2.1 Counterfactual Data Augmentation

Augmenting models with counterfactual data is a popular recent approach for improving model robustness (Kaushik et al., 2019; Bitton et al., 2021; Khashabi et al., 2020; Wu et al., 2021; Ross et al., 2021). With the rise of LLMs, some research works have utilized them to generate counterfactual data to improve the performance of text classification or reasoning (Dixit et al., 2022; Sachdeva et al., 2023; Li et al., 2023e). The work most similar to ours is DISCO (Chen et al., 2023b), which constructs counterfactual data by prompting LLMs to generate phrase perturbations. However, DISCO is not a general counterfactual augmentation strategy, as it is only applicable to natural language inference tasks. Moreover, DISCO modifies only a single span of text at a time, resulting in counterfactual data with limited semantic diversity, which may lead to the risk of model overfitting. Differently, our method attempts to break these causal relationships and reconstruct them using LLMs, thereby increasing the semantic diversity of the text. This is beneficial to reduce the risk of overfitting of the augmented model and enhance its performance on OOD data.

### 2.2 Chain-of-Thought Distillation

LLMs have demonstrated outstanding performance across various downstream tasks, using in-context exemplars or human instructions (Wang et al., 2023a; Si et al., 2023; Gu et al., 2023; Li et al., 2023c). Recent research indicates that CoT prompts can enhance the reasoning capabilities of LLMs for complex problems (Wei et al., 2022; Wang et al., 2022). However, such benefits are only observable in language models of substantial scale. Therefore, migrating CoT capabilities into SLMs through distillation has attracted much attention. The approach typically employs CoT prompts to

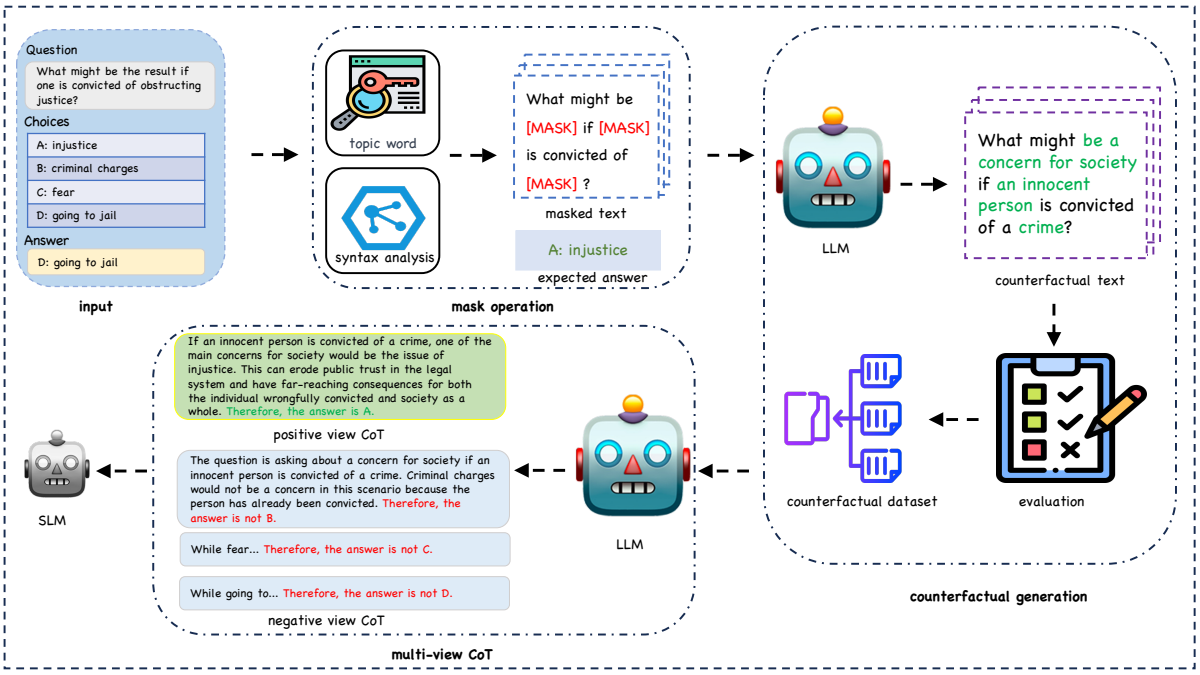


Figure 2: Overview of the counterfactual distillation with multi-view CoT. Our approach consists of three steps: mask operation, counterfactual generation, and multi-view CoT. Initially, through the topic word and syntactic analysis, we identify phrases in the text that may involve causal features, and replace them with the special character [MASK]. Then, given the expected answer, LLM is utilized to complete the masked text and evaluate the generated text to obtain a high-quality counterfactual dataset. In the end, LLM generates rationales for each option, providing supporting evidence for the correct answer and refuting evidence for the incorrect ones.

generate rationales from very large teacher models, and uses them to fine-tune small student models (Huang et al., 2022; Yang et al., 2023; Pezeshkpour et al., 2023; Li et al., 2023d; Chae et al., 2023). By establishing a feedback mechanism between LLMs and SLMs, small models can learn and improve their own shortcomings in a targeted manner (Jiang et al., 2023). Furthermore, the rationales generated by LLMs are closely related to the answers by contrastive decoding, thereby effectively reducing the reasoning errors and hallucinations that SLMs inherit from LLMs (Wang et al., 2023b). For complex multi-step reasoning data, semantic decomposition is beneficial for SLMs reasoning(Shridhar et al., 2023). However, in these approaches, LLMs typically only generate rationales for options they deem worthy of attention, rather than from the perspective of SLMs. Our method uses multi-view CoT to increase the diversity of rationales, allowing small models to learn different types of knowledge, which improves their performance in solving complex reasoning problems. The experimental results show that our method improves performance across multiple QA reasoning datasets, effectively enhancing the reasoning capabilities of SLMs.

## 3 Method

The core idea of our method is to disrupt the causal relationship in the original text and reconstruct it using LLM. Since the answers to the newly generated questions change, SLM is able to learn the causal differences between similar texts through multi-view CoT distillation, thereby enhancing its reasoning capabilities. The overview of our approach is illustrated in Figure 2. In this section, we elaborate on our method and discuss the motivation behind it.

### 3.1 Mask Operation

In order to disrupt the causal relationships in the original text, we first need to identify phrases associated with causal features. To address this issue, we propose a method involving the topic word and syntactic analysis to extract phrases in the text that are relevant to causal features.

- **Topic Word** In general, humans tend to have a clear topic when making statements to ensure that the discussion has a clear focus and accurately conveys information. Similarly, we believe that there are "topic words" in the text, and the

majority of causal features are associated with them. As shown in Figure 2, the topic word of "*What might be the result if one is convicted of obstructing justice?*" is "*obstructing justice*". In order to get the topic word in the text, a prompt goes like this:

**Question:** James was looking for a good place to buy farmland. Where might he look?

**Answer:** The topic word for the sentence is "farmland."

Figure 3: The prompt of topic word

In Figure 3, the underlined texts are completed by LLMs with in-context learning examples.

- **Syntactic Analysis** In QA, the set of possible causal features is large, and relying solely on the topic word cannot fully capture them. As an important part of language structure, noun phrases play a key role in language expression and communication. They are employed to introduce, describe, and connect concepts, thereby enhancing the expressiveness and accuracy of the text. Therefore, noun phrases are highly likely to constitute elements of causal features. In this paper, we use the Stanford syntactic analysis tool [1] to obtain noun phrases in the text, but in order to preserve the elements of the original text as much as possible, personal pronouns and possessive pronouns will be retained.

While maintaining the original linguistic structure of the text, we replace the topic word and noun phrases with the special character [MASK] to disrupt the causal relationships in the original text. Compared with the method of using LLM to perturb a single span in the text (Chen et al., 2023b), our method aims to reduce the difficulty of counterfactual data generation and improve the semantic diversity of the generated text.

### 3.2 Counterfactual Generation

As illustrated in Figure 4, to prompt LLMs for counterfactual text generation, we utilize demonstrations and instructions to construct the prompt. Each demonstration consists of four parts: questions, choices, expected answers, and question completion. In the context of given options, LLMs need to complete the masked question to align it with

[1]https://stanfordnlp.github.io/CoreNLP/

**Question:** Aside from [MASK] what does your [MASK] need ?
**Choices:** A: bone B: charm C: lots of attention D: walked
**Expected answer:** lots of attention
**Completion Question:** Aside from water and nourishment what does your dog need?

**Instruction:** Multiple choice questions consist of questions, options and answers. Based on the above example, please complete the [MASK] part of the question to make it a multiple-choice question with smooth semantics and clear logic.

**Question:** What might be [MASK] if [MASK] is convicted of [MASK] ?
**Choices:** A: injustice B: criminal charges C: fear D: going to jail
**Expected answer:** injustice
**Completion Question:** What might be a concern for society if an innocent person is convicted of a crime?

Figure 4: The prompt of counterfactual generation

the expected answer. Since QA data are usually multiple-choice questions, a request needs to be made to LLM for each option except the answer, so each original QA generates multiple counterfactual examples. To obtain high-quality counterfactual data, we design an evaluation strategy aimed at selecting the most promising counterfactual examples, thereby eliminating potential errors. The strategy is used to verify whether the generated completion questions match the new answers. Based on the self-consistency principle of LLMs (Wang et al., 2022), we sample five reasoning paths for each counterfactual example. If more than three paths are consistent with the new answer, the data is retained; otherwise, it is discarded.

In order to preserve the elements of the original text, our method only modifies the [MASK] part of the text while leaving the rest of the text unchanged. This enables SLMs to focus on the differences between factual and counterfactual texts, thereby allowing it to gain a profound understanding of how different causal features lead to diverse answers in similar texts.

### 3.3 Multi-view CoT

Based on the powerful knowledge base and reasoning capability of LLMs, they often ignore some incorrect options that are obvious to them when answering questions. However, these options may be difficult for SLMs to discern. Therefore, we propose the multi-view CoT, which enables SLMs to learn different types of knowledge by allowing LLMs to generate diverse reasoning paths from different perspectives. As shown in Figure 2, the multi-view CoT consists of two parts: the positive view CoT (PVC) and the negative view CoT (NVC), both generated by LLMs through in-context learning. The goal of PVC is to explore relevant information for the correct answer, while NVC focuses more on generating negative reasoning paths for each option except the answer. Our

method can help small models distinguish and eliminate inadequate options among multiple choices, enabling them to make more targeted selections of the correct answer.

### 3.4 Training

The original data and counterfactual data are mixed together to form the training set. Given an input instance $x = (q, o, a)$ in this set consisting of a question, a set of options and the corresponding answer, the generated PVC and NVC are $x_{pvc} = (q, o, r_{pvc}, a)$ and $x_{nvc} = (q, o, \{(o_k, r^k_{nvc})\}^N_{k=1})$ respectively. In $x_{nvc}$, $N$ represents the number of options apart from the answer. Since the knowledge learned by SLMs from $x_{pvc}$ and $x_{nvc}$ is different and to avoid model confusion, we use special strings to construct two types of data formats during model training.

- $input_{pvc} = q \oplus o \oplus$ *[Direct election]* $\oplus\ r_{pvc} \oplus$ *Therefore the answer would be* $\oplus\ a$
- $input_{nvc} = q \oplus o \oplus$ *[Elimination method]* $\oplus\ r^k_{nvc} \oplus$ *Therefore the answer would not be* $\oplus\ o_k$

The $\oplus$ represents text concatenation. We use the special strings "*[Direct election]*" and "*[Elimination method]*" to guide the small model to generate reasoning paths that either support or refute a certain option.

Given a question, options, and the special string(*st*), the small model is trained to output a sequence of rationale tokens concatenated with the label tokens. In this paper, our approach involves fine-tuning a text-to-text language model using standard language modeling loss on the training data.

$$L = -\sum_i \log P(t_i | q, o, st, t_{<i})$$

## 4 Experiments

We study how small models can learn to reason better on four multi-step reasoning datasets: CommonsenseQA (CSQA) (Talmor et al., 2018), QuaRel (Tafjord et al., 2019), ARC (Clark et al., 2018)and QASC (Khot et al., 2020). The ARC dataset is divided into two parts: the challenge set and the easy set. In our experiments, we combine these two subsets for model training. Since the test labels of CSQA and QASC datasets are not public, we use the official development set as our test set.

In this paper, we utilize the gpt-3.5-turbo API [2] to generate reasoning paths. In the experiment, GPT-2(Radford et al., 2019), OPT(Zhang et al., 2022), and GPT-Neo(Black et al., 2021) are used as small models, and the batch size is set to 64 during training, with a total of 20 epochs. We use HuggingFace transformers and Pytorch for the implementation.

### 4.1 Models

#### 4.1.1 Baseline models

- **Fine-tune(FT)** The small model is fine-tuned on just the label, instead of also with CoT.
- **Fine-tune-CoT(FT-CoT)** The standard CoT distillation (Li et al., 2023b; Ho et al., 2022) method where the small model is trained on the CoT generated by gpt-3.5-turbo.

#### 4.1.2 Our models

In order to verify the effectiveness of multi-view CoT (MVC) and counterfactual data (CD) respectively, we design the following experiments

- **Fine-tune-MVC(FT-MVC)** In this experiment, we use original factual data to generate the multi-view CoT to train a small model, which can validate the effectiveness of the multi-view CoT.
- **Fine-tune-CD-CoT(FT-CD-CoT)** We use the original factual data and the generated counterfactual data to generate normal CoT to fine-tune the small model.
- **Our method** We merge the above two methods together, that is, using original factual and counterfactual data to generate the multi-view CoT to train small models.

### 4.2 Main Results

Table 1 summarizes the accuracy of small models using the proposed method compared to existing FT and FT-CoT. In different scales of models with parameter counts ranging from 120 million to 770 million, our method outperforms FT-CoT. Specifically, our method achieves performance improvements ranging from 4.17% to 23.22% across these scales, with an average improvement of 11.43%.

In the QASC and ARC, the distillation performance of our method on models with fewer parameters exceeds that of models trained with more

[2] https://api.openai.com/v1/chat/completions

| SLMs | Methods | CSQA | QuaRel | QASC | ARC |
|---|---|---|---|---|---|
| GPT2-base(120M) | FT | 23.34% | 53.62% | 15.66% | 25.79% |
| | FT-CoT | 36.94% | 53.26% | 13.28% | 26.86% |
| | FT-MVC | 41.25% | 53.80% | 26.35% | 31.74% |
| | FT-CD-CoT | 43.57% | 55.07% | 32.72% | 37.91% |
| | **Our method** | **46.68%** | **58.70%** | **35.85%** | **40.42%** |
| OPT-base(125M) | FT | 20.56% | 53.26% | 14.04% | 25.99% |
| | FT-CoT | 40.21% | 59.78% | 20.63% | 33.12% |
| | FT-MVC | 45.45% | 60.14% | 30.67% | 35.40% |
| | FT-CD-CoT | 48.24% | 61.23% | 36.29% | 37.91% |
| | **Our method** | **50.61%** | **63.95%** | **39.09%** | **41.09%** |
| GPT-neo(125M) | FT | 21.21% | 52.27% | 15.33% | 26.80% |
| | FT-CoT | 29.81% | 53.44% | 16.52% | 29.68% |
| | FT-MVC | 35.46% | 54.36% | 22.68% | 31.48% |
| | FT-CD-CoT | 34.81% | 54.89% | 24.95% | 33.34% |
| | **Our method** | **40.38%** | **58.15%** | **28.62%** | **37.91%** |
| GPT2-medium(350M) | FT | 38.29% | 56.34% | 17.06% | 26.30% |
| | FT-CoT | 47.13% | 58.15% | 25.27% | 36.70% |
| | FT-MVC | 52.66% | 62.14% | 32.34% | 39.26% |
| | FT-CD-CoT | 55.28% | 60.33% | 38.98% | 43.86% |
| | **Our method** | **58.15%** | **64.86%** | **48.49%** | **47.18%** |
| GPT2-large(770M) | FT | 40.95% | 56.52% | 24.51% | 28.10% |
| | FT-CoT | 52.33% | 58.51% | 32.83% | 38.61% |
| | FT-MVC | 54.22% | 60.87% | 39.96% | 39.60% |
| | FT-CD-CoT | 57.33% | 60.51% | 45.68% | 44.25% |
| | **Our method** | **60.69%** | **67.93%** | **52.48%** | **50.42%** |

Table 1: In experiments, we test the accuracy of different methods on four reasoning datasets. And in order to fully verify the impact of small model size on distillation performance, we conduct experiments on GPT2-base, OPT-base, GPT-neo, GPT2-medium, and GPT2-large models respectively.

parameters using FT-CoT. For instance, in the ARC dataset, our method achieves a performance of 40.42% on GPT2-base, which is 3.27% higher than that of GPT2-medium trained with FT-CoT. Additionally, our method outperforms GPT2-large trained with FT-CoT by 8.57% on GPT2-medium.

- **FT-MVC** *vs* **FT-CoT** As shown in Table 1, the reasoning performance of FT-MVC on models of various sizes exceeds that of FT-CoT. The maximum and minimum improvements are 13.07% and 0.36% respectively, and the average improvement reaches 4.33%. The experimental results indicate that the diversified reasoning paths generated by multi-view CoT enable small models to not only learn reasoning knowledge that supports correct answers but also acquire refutational knowledge regarding incorrect options, thereby effectively enhancing their own reasoning capabilities.

- **FT-CD-CoT** *vs* **FT-CoT** Compared to FT-CoT, FT-CD-CoT shows an average improvement of 7.20% across four datasets. This phenomenon shows that since models with fewer parameters have insufficient capacity, they may not really learn how to reason during the fine-tuning process of FT-CoT, but only remember some patterns and keywords in the training set. In contrast, our approach forces small models to focus on the causal relationships of text, thereby enhancing the accuracy and generalization capability of reasoning. The case study in the appendix provides a more intuitive explanation.

### 4.3 OOD data

To compare the generalization abilities of different methods on OOD data, we select one of the four datasets as the training set and the remaining three as test sets in the experiment. From Figure 5, it can be observed that when CSQA, QASC, and ARC are used as the training sets, our method demonstrates a significant improvement compared to FT-CoT, with enhancements ranging from 3.6% to 26.6%, and an average improvement of 14.9%.

However, when QuaRel is used as the training set, the improvement of our method is not signifi-

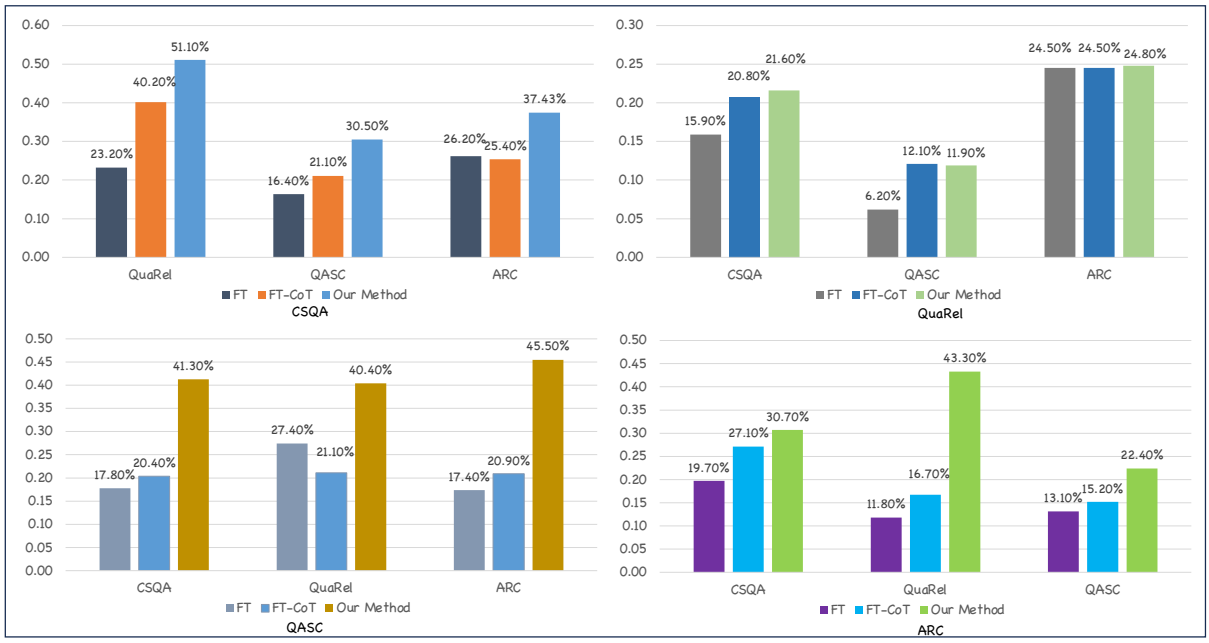


Figure 5: The performance of different methods in out-of-distribution (OOD) scenarios. Specifically, for these four datasets, we select one as the training set and the other three as the test sets to validate the generalization performance of the proposed method.

cant. There are mainly two reasons for this. Firstly, the training samples in the QuaRel dataset are relatively limited, consisting of only 1941 samples. Secondly, more importantly, there is a mismatch in task formats. QuaRel is a QA task with only two options, while the other three datasets have no fewer than four options. Therefore, we find that in this scenario, both the baseline model and our method exhibit very low performance.

### 4.4 Small model architectures

The small models used in previous experiments are all decoder-only language models. To validate the performance of the proposed method on small models with different architectures, we conduct distillation experiments on encoder-decoder language models. Figure 6 illustrates the results of different methods when the small model adopts Bart (Lewis et al., 2019) and T5 (Raffel et al., 2020). Compared to the standard supervised fine-tuning baseline (FT-CoT), our method improves by 13.2% and 11.6% on BART and T5 respectively. This result fully demonstrates that our method can consistently improve performance across small models of different architectures, and highlights its universality and effectiveness in distillation tasks.


Figure 6: The performance of BART and T5 models on four datasets

### 4.5 Data Quality Evaluation

To verify the quality of the generated counterfactual data, we input the original data and its corresponding counterfactual data into GPT-4. We inform GPT-4 that one of the entries is manually annotated and the other is AI-generated. Then, GPT-4 is used to determine the category of these data. The ap-

| | GPT2-base | | | GPT2-medium | | | GPT2-large | | |
|---|---|---|---|---|---|---|---|---|---|
| | FT | FT-COT | FT-MVC | FT | FT-COT | FT-MVC | FT | FT-COT | FT-MVC |
| CSample | 22.52% | 33.01% | 36.76% | 25.63% | 42.83% | 46.32% | 30.06% | 48.98% | 51.76% |
| CMix | 30.14% | 36.28% | 39.39% | 37.01% | 46.36% | 49.55% | 44.23% | 50.12% | 53.38% |

Table 2: Comparison of the performance of CSample and CMix across various models.

| CSQA | QuaRel | QASC | ARC |
|---|---|---|---|
| 48% | 42% | 44% | 58% |

Table 3: Classification accuracy of GPT-4 on four datasets.

| | FT | FT-CoT | Our Method |
|---|---|---|---|
| SVAMP | 8.67% | 10.11% | 16.00% |
| bAbI | 67.40% | 69.20% | 78.23% |

Table 4: The performance of different models on SVAMP and bAbI.

pendix includes specific prompt examples. We randomly select 100 data samples from each dataset for data quality evaluation. As demonstrated in Table 3, the average classification accuracy of GPT-4 across four datasets is 48%. Given that this is a binary classification task, this accuracy rate is comparable to that of random selection, suggesting that GPT-4 is unable to effectively distinguish between original data and their corresponding counterfactual data. This result further substantiates the high quality of the counterfactual data generated by our method.

### 4.6 Causal Features

We aim to demonstrate that the enhancement in model performance stems from the learning of causal features, not merely from the growth in data volume. To validate this hypothesis, we have designed the following experiment.

Specifically, we randomly select *N* samples from the CSQA dataset, which we name *CSample*. Subsequently, we randomly extract *N/2* samples from the *CSample*, which we call *CHalf*. Based on *CHalf*, we construct *N/2* pieces of counterfactual data, which we label as *CCD*. Finally, we merge the *CHalf* and *CCD* data to form a new dataset, which we name *CMix*. Therefore, the *CSample* and *CMix* datasets each contain *N* samples, and we then validate the performance of these two datasets on different models.

As shown in Table 2, *CMix* outperforms *CSample* across various models. This phenomenon reveals that by contrasting the differences between factual and counterfactual data, the model can effectively suppress the interference of spurious correlations, thereby enhancing the learning of causal features. It is worth noting that the improvement of *CMix* in the FT-MVC model is relatively small compared to its performance improvement in the FT and FT-CoT models. This shows that the diversified reasoning paths enable the small model to maintain stable performance under different data qualities.

### 4.7 Non-multiple choice tasks

In the aforementioned experiments, we mainly focus on multiple choice questions. To validate the performance of the proposed method on non-multiple choice tasks, we select the SVAMP(Patel et al., 2021) and bAbI(Weston et al., 2015) datasets for in-depth analysis. SVAMP is a mathematical reasoning dataset, while bAbI is a reading comprehension dataset consisting of 20 subtasks. To construct the training set, we extract 50 samples from each subtask of bAbI, resulting in a total of 1000 samples. Notably, since both the SVAMP and bAbI datasets do not include options, we utilize LLM to generate possible options for these two datasets during the training phase, thereby transforming them into multiple-choice tasks. As shown in Table 4, our method outperforms the baseline on both datasets. This result not only verifies the effectiveness of our method on multiple-choice tasks but also proves that it is also applicable to open-ended tasks.

## 5 Conclusion

In this paper, we propose a distillation approach based on counterfactual data augmentation and multi-view CoT, aiming to enable LLMs to instruct small language models in reasoning. The experimental results demonstrate that our method outperforms baseline models across various scales and

architectures of small models, and exhibits strong generalization and robustness on OOD data. Moreover, our method shows broad application potential across a variety of tasks.

## Limitations

In our research, we find that the text generated by LLMs does not always align with the expected answers, leading us to discard some data during the evaluation phase. This indicates that further optimization is needed in terms of data alignment and answer consistency. Additionally, our experiments are limited to English datasets and single-task settings. To better compare with the few-shot settings of large language models, future research can explore other languages and multi-task settings.

The use of closed-source large language models in this study incurs additional costs. Future research should explore the distillation effects on more open-source large models to reduce costs and improve the generalizability of the method. Additionally, the current experiments only involve question-answering reasoning tasks. Exploring how to validate the method's effectiveness on more task types is an important direction for future research.

## Acknowledgments

This work was supported by the NSFC project (No. 62072399), the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ23F020009, MoE Engineering Research Center of Digital Library, China Research Centre on Data and Knowledge for Engineering Sciences and Technology, the Fundamental Research Funds for the Central Universities (No. 226-2024-00170), and Ant Group. We also express our sincere gratitude to anonymous reviewers for their invaluable feedback and constructive comments.

## References

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. 2021. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. *arXiv preprint arXiv:2103.09591*.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, 58:2.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. Docogen: Domain counterfactual generation for low resource domain adaptation. *arXiv preprint arXiv:2202.12350*.

Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. *arXiv preprint arXiv:2310.09343*.

Hailin Chen, Amrita Saha, Steven Hoi, and Shafiq Joty. 2023a. Personalised distillation: Empowering open-sourced llms with adaptive learning for code generation. *arXiv preprint arXiv:2310.18628*.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023b. Disco: distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. 2023. Counterfactual active learning for out-of-distribution generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11362–11377.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Core: A retrieve-then-edit framework for counterfactual data generation. *arXiv preprint arXiv:2210.04873*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context. *arXiv preprint arXiv:2305.09137*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. 2022. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models. *arXiv preprint arXiv:2212.10670*.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of closed-source large language model. *arXiv preprint arXiv:2305.12870*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. *arXiv preprint arXiv:2004.04849*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023a. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. *arXiv preprint arXiv:2305.16572*.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023b. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023c. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

Xinwei Li, Li Lin, Shuai Wang, and Chen Qian. 2023d. Unlock the power: Competitive distillation for multi-modal large language models. *arXiv preprint arXiv:2311.08213*.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2023e. Large language models as counterfactual generator: Strengths and weaknesses. *arXiv preprint arXiv:2305.14791*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Pouya Pezeshkpour, Hayate Iso, Thom Lake, Nikita Bhutani, and Estevam Hruschka. 2023. Distilling large language models using skill-occupation graph context for hr-related tasks. *arXiv preprint arXiv:2311.06383*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *arXiv preprint arXiv:2107.07150*.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. 2023. Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. *arXiv preprint arXiv:2309.07822*.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring inductive biases of in-context learning with underspecified demonstrations. *arXiv preprint arXiv:2305.13299*.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7063–7071.

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

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2023. Effective distillation of table-based reasoning ability from llms. *arXiv preprint arXiv:2309.13182*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md Arafat Sultan, Jay-Yoon Lee, Mohit Iyyer, and Andrew McCallum. 2023. Multistage collaborative knowledge distillation from large language models. *arXiv preprint arXiv:2311.08640*.

# Appendix

## A Case Study

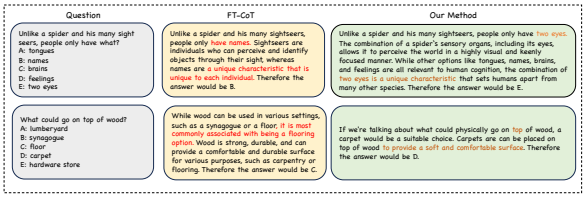


Figure 7: Comparison of the results of our method and FT-CoT on commonsense reasoning data.

To better illustrate the differences between our approach and FT-CoT, we present some examples in commonsense reasoning. As shown in Figure 7, FT-CoT only captures the surface correlation of semantics, such as "people" and "names," "wood" and "floor". Small models typically focus on word-level associations during reasoning, rather than the complete semantic of the context. Differently, our method focuses on deeper semantic representations, which helps small models better understand the causal relationships within the text.

## B Prompt

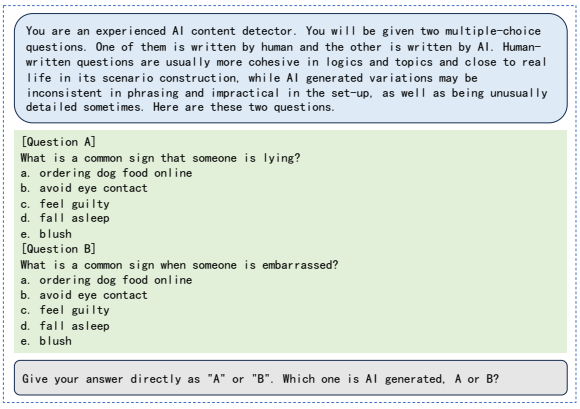


Figure 8: The prompts of data quality evaluation.

The prompt of data quality evaluation is shown in Figure 8.

## C More Results

To further verify the performance of FT-MVC and FT-CD-CoT on OOD data, we design an experiment similar to that in Figure 5, where one of the four datasets is selected as the training set, and the remaining datasets are used as the test set. As shown in Figure 9, FT-MVC and FT-CD-CoT outperform FT-CoT in most cases. Additionally, in

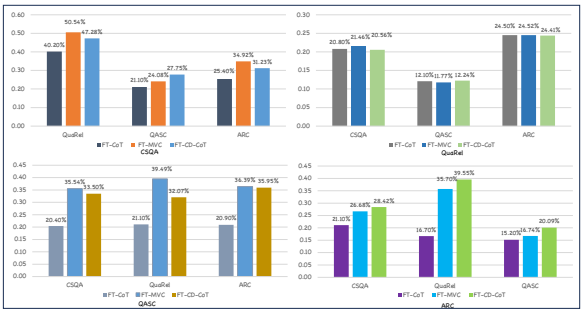


Figure 9: The performance of FT-MVC and FT-CD-CoT in out-of-distribution (OOD) scenarios.

combination with the results from Figure 5, it is evident that FT-MVC and FT-CD-CoT are complementary in terms of performance, and their integration can further enhance the reasoning ability of the small model.