Distillation with Explanations from Large Language Models

Hanyu Zhang^{1,2,3}, Xiting Wang^{4,*}, Xiang Ao^{1,2,3}, Qing He^{1,2,3,*}

¹Key Lab of AI Safety of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

²Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),

Institute of Computing Technology, CAS, Beijing 100190, China

³University of Chinese Academy of Sciences, CAS, Beijing 100049, China

⁴ Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

{zhanghanyu20z, aoxiang, heqing}@ict.ac.cn, xitingwang@ruc.edu.cn

Abstract

Free-text explanations are crucial for enhancing the interpretability of AI models. However, training models to generate high-quality free-text explanations is challenging, primarily due to the requirement of a substantial amount of human-written explanations, which can be expensive. Recently, Large language models (LLMs) like ChatGPT and GPT-4 have made remarkable progress in various NLP tasks while also providing explanations alongside their answers. Leveraging LLMs for data labeling offers a more cost-effective alternative. However, a key concern arises from the fact that the answers provided by LLMs are not entirely accurate, potentially introducing noise to both task outputs and explanation generation. To remedy this, we propose a new mechanism, *Distillation with Explanations* from LLMs. we observe that despite the incorrectness in LLMs-generated answers, their explanations are consistent with their answers. Leveraging this consistency, our method combines the ground truth labels and answers-explanations generated by LLMs, to simultaneously generate more accurate answers and the corresponding free-text explanations. Experimental results demonstrate that our approach achieves improved predictive performance and also generates explanations that exhibit greater alignment with the model's task outputs.

Keywords: Large language model, Explainability, Question Answering

1. Introduction

Model explanations play a crucial role in establishing trust for natural language processing (NLP). Among various types of explanation, free-text explanations have drawn substantial attention, due to their comprehensibility, rich information content, and intuitive nature for humans. (Narang et al., 2020; Rajani et al., 2019; Kumar and Talukdar, 2020; Brahman et al., 2021). Typically, free-text explanations are generated alongside task answers, following a paradigm known as the selfrationalization model (Alvarez Melis and Jaakkola, 2018). However, training models to generate highquality free-text explanations is challenging, primarily due to the requirement of a substantial amount of human-written explanations, which can be expensive (Narang et al., 2020; Marasović et al., 2022). Despite the numerous datasets containing task labels for various NLP tasks, datasets annotated with free-text explanations for the answers are scarce. This scarcity of explanation annotations poses a hindrance to the training of self-rationalization models.

Recently, Large language models (LLMs) such as ChatGPT and GPT-4 have made remarkable progress in various NLP tasks (Zhang et al., 2022; Bang et al., 2023; Qin et al., 2023). In addition

	WinoGrande	CQA	COPA
Accuracy	81.46%	76.36%	96.12%

Table 1: Task accuracy by ChatGPT.

to their outstanding performance for various tasks, LLMs have also demonstrated the ability to generate high-quality explanations along with answers when provided with appropriate prompts. Notably, explanations produced by LLMs are both linguistically well-formed and informative, making them comparable to explanations written by humans. Moreover, the use of LLMs for generating explanations offers several advantages in terms of convenience, time efficiency, and cost-effectiveness.

Despite the outstanding performance of LLMs, directly manipulating these large models presents challenges. As a result, distillation techniques have emerged as an effective solution for transferring knowledge from LLMs to smaller models (Hsieh et al., 2023). In the distillation process, smaller models are trained using labels generated by LLMs. However, an issue arising with LLMs for labeling is that the answers by LLMs are not entirely accurate. Table 1 presents the accuracy results of Chat-GPT answers on three commonsense reasoning datasets, revealing that LLMs still have limitations. Both the answers and explanations generated by

^{*} Corresponding authors.

Samples with correct answers by ChatGPT



Figure 1: We provide examples where ChatGPT answers correctly and incorrectly. Green represents the correct answer and explanation, and red represents the incorrect answer and explanation. When ChatGPT provides an incorrect answer, in the first model $I \rightarrow O^L, I \rightarrow R^L$ (middle), an incorrect task label and the corresponding explanation are used, which introduces noise to the supervised learning of question answering. The second model $I \rightarrow O^*, I \rightarrow R^L$ (down) combines the correct task label with the incorrect explanation, leading to inconsistency.

LLMs may contain incorrect information, introducing noise for the smaller model training. Therefore, how to deal with these incorrectly annotated samples poses a significant problem that requires careful consideration.

Figure 1 shows an example where ChatGPT provides an incorrect answer ¹. The question is "What can you do after learning about science? " and the ground truth label is 'invent'. Whereas, Chat-GPT gives an incorrect but conceptually related answer 'experiment' and provides the explanation that describes the relation between 'experiment' and 'learning about science'. In distillation, directly using the answers 'experiment' and explanations generated by LLMs without considering their correctness will introduce noise to both the task output and explanation generation. Alternatively, using the ground truth labels '*invent*' as task supervision and LLMs-generated explanations as rationale supervision may preserve the accuracy of task predictions but result in inconsistent explanations and answers in the distilled model. Also, we can simply drop the incorrectly annotated samples but this would reduce the available training data, leading to a waste of data resources. Therefore, when faced with the scenario of partially incorrect labeling by LLMs, determining how to effectively leverage these labels and explanations becomes a significant challenge.

In this work, we introduce a novel mechanism, **Distillation with Explanations**, for training smaller models using a combination of ground truth labels and annotations by LLMs. We observe that the explanations generated by LLMs consistently align with their own answers, even if the answers are incorrect (as depicted in Figure 3), along with previous studies on LLMs (Wei et al.; Kojima et al.). Based on this observation, we leverage the consistency between explanations and answers to develop a scoring model. Then, we employ ground truth labels as task supervision and LLMsgenerated explanations as explanation supervision. Simultaneously, we utilize the consistency scoring model to refine the generated explanations and ensure they align with the predicted outcomes, thus yielding more coherent and reasonable explana-

¹Answer 'experiment' and answer 'invent' both seem correct in Figure 1. Here is the illustration in ECQA dataset to explain why 'invent' is better. (ECQA dataset provides both reasons for the chosen answer and the rejected answer for CQA dataset.) For 'invent', it states "Invent is to create or design something that has not existed before; be the originator of. We can invent after learning about science." For 'experiment', it states "Experiment is a scientific procedure undertaken to make a discovery or demonstrate a known fact. Experiment is to be performed parallelly while learning and not after learning about science." There are indeed some confusing answers in CQA dataset. Both answers seem correct. However, it could be possible to identify the best answer through meticulous explanations.

tions.

Distillation with Explanations allows the smaller model to generate more reasonable explanations while simultaneously achieving improved prediction performance. Our experimental results on three datasets demonstrate our approach surpasses the base framework with 0.8-6% for task accuracy and the associated explanations consistency is improved meanwhile. Moreover, we conduct a comparison between ChatGPT-generated annotations and human-written explanations and validate that even in the presence of incorrect answers, the distilled smaller model, guided by explanations from LLMs, can still achieve superior performance. In summary, our main contributions are three-fold:

- We consider the inevitable issue of using LLMs for data labeling that LLMs will generate incorrect labels. And We show that despite the incorrect answers by LLMs, their explanations remain consistent with corresponding answers.
- 2. Based on the consistency between explanations and answers generated by LLMs, we propose a novel Distillation with Explanations approach, for smaller models distillation with guidance by the consistency.
- Experiments demonstrate that our approach effectively improves the task performance and the consistency between explanations and answers of smaller models.

2. Related work

Free-text explanation generation. Free-text explanations are particularly important among various types of explanations, including visualization(Kui et al., 2022; Chotisarn et al., 2023; Klaus et al., 2023), feature importances(Ribeiro et al., 2016; Lundberg and Lee, 2017; Gao et al., 2019), for their capacity to contain rich semantic information in an easily understandable format. Especially with the advancement of large language models, studies of generating free-text explanations have become common. (Rajani et al., 2019; Kumar and Talukdar, 2020; Latcinnik and Berant, 2020) adopt a pipeline architecture to generate explanations and answers. (Chen et al., 2019, 2021) generate free-text explanations for recommendations using the multi-task framework. Another common approach is self-rationalizing models (Alvarez Melis and Jaakkola, 2018), which generates both the answers and explanations in an output (Hancock et al., 2018; Ehsan et al., 2018; Liu et al., 2019a; Wu and Mooney, 2019; Narang et al., 2020; Tang et al., 2020). While this approach is user-friendly, it is important to note that it can result in a decrease

in task accuracy within this framework (Wiegreffe et al., 2021).

Distillation from large language models. Large language models (LLMs) have demonstrated outstanding performance across various NLP tasks. However, due to their excessive parameter size, these models face challenges when it comes to practical deployment. As a result, some studies have researched on knowledge distillation (Hinton et al., 2015) to transfer the knowledge from large models to smaller ones. (Wang et al.) use LLMs to generate explanations for all answers to train a smaller model. However, during inference, LLMs are required to generate explanations, making the smaller model unable to apply independently. (Ho et al., 2022) employ the LLMs to generate multiple explanations and then filter out the explanations for correct answers to fine-tune the smaller model. (Magister et al., 2023; Li et al., 2022) fine-tune a student model on the chain of thought (CoT) (Wei et al.) outputs generated by LLMs. (Hsieh et al., 2023) uses CoT prompting to extract labels and explanations from LLMs, and use them to train smaller models. Compared to these existing works, our approach takes into account the inevitable issue of LLMs: incorrect annotations. And we discuss how to handle and utilize incorrect labels and explanations.

3. Distillation with Explanations

We propose the Distillation with Explanations framework that leverages the consistency between explanations and answers to enable the smaller model to generate more reasonable explanations. Our overall framework is depicted in Figure 2. First, we use all the annotations by LLMs, irrespective of correctness, to train a consistency model. The consistency model assesses the probability of generating an answer based on an explanation. Subsequently, we utilize the probability derived from the consistency model as a reward to guide the generation of explanations during smaller model distillation.

3.1. Multi-task framework and the choice of task supervision

We first introduce the base framework employed in our model. Given the question input I, and the ground truth output O^* , we refer to the smaller model as f. Additionally, LLMs generate the answer O^L and explanation R^L .

A commonly used self-explaining framework is to train a generation model using supervised prediction labels and explanations. In our approach, we adopt a multi-task framework, represented as $I \rightarrow O, \ I \rightarrow R$, where predictions and explanations serve as separate supervision signals. This frame-



Figure 2: The overview framework of our Distillation with Explanations method. We first use all the explanations and labels from LLMs to train a consistency model, without considering the correctness. Then within the multi-task framework, $I \rightarrow O^*, I \rightarrow R^L$, we use the trained consistency model as a reward to guide the generation of explanations.

work has proven effective in generating predictions (Hsieh et al., 2023). The output predictions O and explanations R are conditioned solely on the task inputs I. By adding different prefix words before the input sentence, we guide the model to produce either the task answer or explanation. Here, we generate predictions through text generation, instead of training specific classification models or multiple-choice models. We opt for this because text generation aligns with the mechanism by LLMs to provide answers. Moreover, this method offers enhanced convenience in practical usage.

When considering the selection of the supervision signal, we choose O^* instead of O^L because O^L contains certain incorrect labels. This choice aims to mitigate the potential decrease in the accuracy of the distilled smaller model.

Hence, our framework could be rewritten as $I \rightarrow O^*, I \rightarrow R^L$. For $x \in I$ the task prediction loss is the text generation loss $L_{\text{task}} = l(f(x), y^*)$ for $y^* \in O^*$, and the explanation generation loss is the text generation loss $L_{\text{exp}} = l(f(x), r^L)$ for $r^L \in R^L$. The multi-task learning objective combines these two as follows:

$$L = L_{\text{task}} + \lambda L_{\text{exp}} \tag{1}$$

3.2. Consistency model

Although LLMs will occasionally produce incorrect answers sometimes, we notice a consistent alignment between the generated explanations and answers, even in cases where the answers are incorrect (shown in Figure 3). Based on this insight, we utilize all the generated predictions $r^L \in R^L$ and explanations $y^L \in O^L$, regardless of the correctness, to train a consistency model s(r) with loss: $L_{\text{consistency}} = l(s(r), y^L)$. With the consistency model, we can calculate the probability $P_s(y|r)$ of generating an answer given an explanation, by accumulating the probability of generating each text token.

3.3. Alignment through reinforcement learning

With the introduction of the consistency model, which measures the alignment between the generated explanations and answers, we now incorporate this consistency into the multi-task training process. To optimize the generated explanations to align with the generated answers, we employ the REINFORCE method (Williams, 1992). For a model-generated explanation r and a generated prediction y, the reward of r is calculated through the probability that consistency model s(r) outputs y when given input r. To ensure the efficiency of the optimization, we add a baseline reward which is the probability that the consistency model outputs y when given the input question x only. The RL loss is :

$$L_{\text{align}} = -(\operatorname{score}(r; y) - \operatorname{score}(x; y)) \log P_f(r)$$
 (2)

where $score(r; y) = \log P_s(y|r)$ is the explanation reward and $score(x) = \log P_s(y|x)$ is the baseline reward. $P_f(r)$ is the probability that distilled model f generate r.

Then the multi-task learning objective is :

$$L = L_{\text{task}} + \lambda L_{\text{exp}} + \gamma L_{\text{align}}$$
(3)

Automatic weight adjustment. To achieve better optimization, we employ an automatic weight adjustment method (Kendall et al., 2018) to finetune the weights among different tasks in multi-task learning. The specific optimization objective is defined as follows:



Figure 3: We conducted a comparative analysis to examine whether the answers can be inferred from explanations in two distinct scenarios: the answers from ChatGPT are correct and incorrect. We use two pre-trained NLI models: RoBERTa-base and DeBERTaV3-large. The first row is the result of the entailment proportion, which means the answer can be inferred from the explanation. Oppositely, the second row is the contradiction proportion. On CQA task, we also include two datasets with human-annotated explanations: ECQA and COS-E.

$$L = \frac{1}{2\sigma_1^2} L_{\text{task}} + \frac{1}{2\sigma_2^2} L_{\text{exp}} + \frac{1}{2\sigma_3^2} L_{\text{align}} + \log \sigma_1 \sigma_2 \sigma_3$$
(4)

where $\sigma_1, \sigma_2, \sigma_3$ are parameters to update.

4. Experiments

In this section, we first provide an overview of the experimental settings in our study. We then give an analysis of the consistency between predictions and explanations generated by LLMs. Subsequently, we present the main results of our Distillation with Explanations approach, for enhancements in both prediction accuracy and explanation quality. We complement our findings with a case study, providing further insights and illustrating the effectiveness of our method. Finally, we compare the model performance trained with ChatGPT and human-written explanations.

4.1. Experimental settings

Datasets. We conduct experiments on three commonsense reasoning datasets: WinoGrande (Sakaguchi et al., 2019), CQA (Talmor et al., 2019), COPA (Roemmele et al., 2011). Answers and explanations generated by ChatGPT are acquired by ChatGPT API². **Baselines.** We compare our model with several commonly used self-explaining frameworks. In all the frameworks, we use T5-base models (Raffel et al., 2020) as the base model to generate answers or explanations. First, we compare our method with three common frameworks:

- + $\mathbf{I} \rightarrow \mathbf{O}$. The framework maps task inputs to task answers.
- + $\mathbf{I} \to \mathbf{OR}.$ The framework maps task inputs to task answers followed by explanations.
- $I \rightarrow O^*, I \rightarrow R^L$. The framework maps task inputs to answers and explanations separately. To provide guidance for the model's output of answers and explanations, we introduce "task prefixes" [label] and [rationale] that precede the task inputs. (Raffel et al., 2020; Hsieh et al., 2023)

Then we additionally choose different task labels as task supervision in the framework $I\to O, I\to R$ as our baseline:

- $I \rightarrow O^L, I \rightarrow R^L$. During the training phase, use the answers and explanations generated by ChatGPT as supervision
- $I \rightarrow O^{C}, I \rightarrow R^{C}$. During the training phase, we only use samples in which ChatGPT provides correct answers.

Metrics. We assess the performance of models on two criteria: the accuracy of the generated task answers and the quality of the explanations. To

²https://openai.com/blog/

introducing-chatgpt-and-whisper-apis

	Model	Accuracy ↑	LAS ↑	Prob score ↑	BERTscore ↑
	$\mathrm{I} \to \mathrm{O}$	61.0277	-	-	-
	$\mathrm{I} \to \mathrm{OR}$	57.9447	-0.0266	0.8968	0.8210
WinoGrande	$\mathrm{I} \to \mathrm{O}^*, \mathrm{I} \to \mathrm{R}^\mathrm{L}$	62.6877	-0.1236	0.6338	0.8215
	Ours	63.2411	<u>-0.0868</u>	0.6824	0.8194
	$\mathrm{I} \to \mathrm{O}$	59.6201	-	-	-
	$\mathrm{I} \to \mathrm{OR}$	54.6656	0.2420	0.8604	0.7346
CQA	$\mathrm{I} \to \mathrm{O}^*, \mathrm{I} \to \mathrm{R}^\mathrm{L}$	59.2898	0.1905	0.6330	0.7151
	Ours	60.1982	<u>0.2013</u>	<u>0.6385</u>	<u>0.7185</u>
	$\mathrm{I} \to \mathrm{O}$	62.2	-	-	-
	$\mathrm{I} \to \mathrm{OR}$	58.4	0.1536	0.9301	0.7760
COPA	$\mathrm{I} \to \mathrm{O}^*, \mathrm{I} \to \mathrm{R}^\mathrm{L}$	67.4	-0.3301	0.4686	0.7492
	Ours	71.6	<u>-0.3214</u>	<u>0.6683</u>	<u>0.7731</u>

Table 2: Experiments comparing various baseline distillation frameworks and our Distillation with Explanations method. The <u>underline</u> denotes our method shows improvement over the framework $I \rightarrow O^*, I \rightarrow R^L$ we based on (highlighted in gray).

evaluate the quality of explanations, we employ the following metrics:

DeBERTaV3-large model ⁵.

- LAS. Leakage-Adjusted Simulatability (LAS) (Hase et al., 2020) measures how well explanations help a simulator predict the model's answers, averaging over groups based on whether the explanations leak labels or not.
- **Prob Score.** We use the probability model in Section 3.2 to evaluate how well the explanations can infer the answer.
- **BERTscore.** As one of the widely used metrics in text generation, we also include BERTscore (Zhang et al., 2020). BERTscore measures the correlation between the generated text and the ground truth text. In our experiments, BERTscore is computed only on samples where ChatGPT predicts correct answers. However, it is crucial that various explanations can explain a question, and the ground truth explanation represents only one possible explanation. Therefore, the BERTscore should be considered as a reference metric.

We conduct our experiments on one A800 GPU. We use T5-base models to generate answers and explanations in all the frameworks. T5-base models parameters are initialized from the public model ³. We train the models with learning rate = 5×10^{-5} , max epochs = 200, early stopping step = 3. We use batch size = 64, 16, 16 for Wino-Grande, CQA, COPA datasets respectively. We initialize loss weight $\lambda = 1.0$ and $\gamma = 0.1$ for Wino-Grande, $\gamma = 0.003$ CQA datasets and $\gamma = 0.03$ for COPA dataset. For NLI models, we use pretrained RoBERTa-base model ⁴ and pre-trained

4.2. Consistency between answers and explanations generated by ChatGPT

First, we employ natural language inference (NLI) models to assess whether the answers can be inferred from the explanations. We use two pretrained NLI models: RoBERTa-base (Liu et al., 2019b) and DeBERTaV3-large (He et al., 2023). Within the NLI framework, each sample consists of a premise (the explanation) and a hypothesis (comprised of the question and the answer). Figure 3 shows the NLI results, separately on where ChatGPT generates correct and incorrect answers.

Notably, even when ChatGPT produces incorrect answers, the proportion of answers entailed from explanations remains consistent compared to situations where the answers are correct. On the CQA dataset, the entailment proportion is even higher for incorrect answers. This observation suggests that the consistency between explanations and answers is not significantly influenced by the correctness of the answers. Although incorrect answers and their corresponding explanations may not directly contribute to distillation from ChatGPT, the consistency between incorrect answers and corresponding explanations can still provide valuable guidance for model training.

We also present the NLI results obtained from two sets of manually annotated explanation data on the CQA dataset: ECQA (Aggarwal et al., 2021) and COS-E (Rajani et al., 2019). It is worth noting that the entailment proportion from these manually annotated datasets is significantly lower, partially due to the brevity and limited information contained within the human-written explanations.

³https://huggingface.co/t5-base

⁴https://huggingface.co/cross-encoder/ nli-roberta-base

⁵https://huggingface.co/cross-encoder/ nli-deberta-v3-large

	Model	Accuracy ↑	LAS ↑	Prob score ↑	BERTscore ↑
	$I \to O^*, I \to R^L$	62.6877	-0.1236	0.6338	0.8215
	$\mathrm{I} \to \mathrm{O^L}, \mathrm{I} \to \mathrm{R^L}$	56.6798	-0.2592	0.6824	0.8194
WinoGrande	$I \rightarrow O^C, I \rightarrow R^C$	58.7352	-0.1867	0.6920	0.8217
	Ours	63.2411	-0.0868	0.6824	0.8194
	$\mathrm{I} \to \mathrm{O}^*, \mathrm{I} \to \mathrm{R}^\mathrm{L}$	59.2898	0.1905	0.6330	0.7151
	$\mathrm{I} \to \mathrm{O^L}, \mathrm{I} \to \mathrm{R^L}$	59.8679	0.1481	0.6701	0.7141
CQA	$I \rightarrow O^C, I \rightarrow R^C$	59.7027	0.1754	0.6866	0.7116
	Ours	60.1982	0.2013	0.6385	0.7185
	$\mathrm{I} \to \mathrm{O}^*, \mathrm{I} \to \mathrm{R}^\mathrm{L}$	67.4	-0.3301	0.4686	0.7492
	$\mathrm{I} \to \mathrm{O^L}, \mathrm{I} \to \mathrm{R^L}$	66.4	-0.3390	0.6287	0.7697
COPA	$I \rightarrow O^C, I \rightarrow R^C$	67.4	-0.3589	0.4746	0.7585
	Ours	71.6	-0.3214	0.6683	0.7731

Table 3: Results comparing various variants of $I \to O, I \to R$ and our Distillation with Explanations method.

Model	Consistency score ↑	Plausibility score ↑	$LAS \uparrow$	Prob core ↑	BERTscore ↑
$I \to O^*, I \to R^L$	0.4625	0.3375	-0.3301	0.4686	0.7492
$\mathrm{I} \to \mathrm{O^L}, \mathrm{I} \to \mathrm{R^L}$	0.6125	0.4	-0.339	0.6287	0.7697
$\mathrm{I} \to \mathrm{O^C}, \mathrm{I} \to \mathrm{R^C}$	0.4125	0.4	-0.3589	0.4746	0.7585
Ours	0.625	0.4875	-0.3214	0.6683	0.7731

Table 4: User study results of different models on COPA dataset.

4.3. Main results of Distillation with Explanations

We first compare our Distillation with Explanations method with two baseline frameworks: $I \rightarrow OR$ and $I \rightarrow O, I \rightarrow R$. Results are shown in Table 2. Results reveal that the prediction accuracy of $I \rightarrow OR$ is significantly lower compared to $I \rightarrow O, I \rightarrow R$, and even lower than the situation where only labels are utilized ($I \rightarrow O$). The drop in prediction accuracy can be attributed to the additional training requirement of optimizing explanation generation in the $I \rightarrow OR$ framework, which adversely affects the prediction performance. However, the explanation quality of $I \rightarrow OR$ is superior to that of $I \rightarrow O, I \rightarrow R$. This can be attributed to the that the generated answers serve as valuable guidance for better explanation generation.

In comparison to the frameworks $I \rightarrow O, I \rightarrow R$ we base on (the highlighted gray rows in Table 2), our method demonstrates improvements in both prediction accuracy and explanation quality, with the exception of only one anomaly observed in the WinoGrande dataset regarding BERTscore. While our method may not yield explanations as outstanding as the $I \rightarrow OR$ framework, our accuracy significantly surpasses $I \rightarrow OR$.

Furthermore, we conduct a comparative analysis between our method and various variants of the $I \rightarrow O, I \rightarrow R$ frameworks. The results are presented in Table 3. Remarkably, our model consistently outperforms the others in terms of accuracy.

For explanation quality, our model achieves the highest LAS score. Regarding the remaining explanation evaluation metrics, we hypothesize that under the $I \rightarrow O^L, I \rightarrow R^L$ and $I \rightarrow O^C, I \rightarrow R^C$, where the answers and explanations utilized for distillation are all consistent, this consistency within the training data contributes to achieving the highest Prob score.

User study We also conducted a human evaluation on 80 randomly sampled questions from the COPA dataset. We ask participants to label two scores for the four generated explanations respectively. *Consistency score*: whether the explanation is consistent with the answer (1) or not (0). Plausibility score: whether the explanation is reasonable (1) or not (0). Results are shown in Table 4. The first two columns are human evaluation metrics and the last three columns are the automatic metrics. There are some interesting correlations between human evaluations and automatic metrics. We note 1) a potential correlation between Prob score and consistency_score: both our method and the second $\mathrm{I} \to \mathrm{O^L}, \mathrm{I} \to \mathrm{R^L}$ method exhibit comparatively higher scores in these two metrics. 2) BERTscore may be correlated with plausibility score: on both metrics, the scores of the last three methods are not significantly different, and the first $I \rightarrow O^*, I \rightarrow R^L$ method has a slightly lower score.

Case study. We provide an illustrative example from the WinoGrande dataset in Table 5 to highlight the capabilities of our method. In this ques-

Question: The commodities trader decided to buy wool and sell cotton because the _ was priced high. Choice A: wool Choice B: cotton

ChatGPT	Answer: wool 🗡
	Explanation: The sentence states that the commodities trader decided to buy wool and sell cotton. The reason for this decision is given in the second part of the sentence, which states that something was priced high. Based on common sense, the commodities trader would buy a commodity that is priced low and sell a commodity that is priced high to make a profit. Therefore, the commodity that was priced high must be wool as it is the one that the trader decided to buy. Conversely, cotton must be priced lower since the trader chose to sell it instead. Therefore, the correct answer is option A, wool, as it is the commodity that was priced high and prompted the trader's decision to buy wool and sell cotton.
$I \rightarrow O^*, I \rightarrow R$	^L Answer: cotton \checkmark Explanation: The sentence states that the commodities trader decided to buy wool and sell cotton. The reason for this decision is given in the second part of the sentence, which states that the wool was priced high. Therefore, the correct answer is option A, wool, as it is the commodity that was priced high and the trader decided to buy it instead of selling it. Note: The answer of $I \rightarrow O^*, I \rightarrow R^L$ is correct but the generated explanation contradicts the answer.
Ours	Answer: cotton \checkmark Explanation: The sentence states that the commodities trader decided to buy wool and sell cotton. The reason for this decision is given in the second part of the sentence, which states that the cotton was priced high. Therefore, the correct answer is option B, cotton, as it is the commodity that was priced high and prompted the trader to switch to buying wool instead.

Table 5: A test example where ChatGPT provides an incorrect answer. $I \rightarrow O^*, I \rightarrow R^L$ also generates the correct answer, but with a contradicting explanation. Our model successfully corrects the answer and generates corresponding explanations.

tion, "The commodities trader decided to buy wool and sell cotton because the _ was priced high.", ChatGPT gives the wrong answer wool. In model $I \rightarrow O^*, I \rightarrow R^L$, the answer cotton is correct. However, the generated explanation explains why the answer is wool and states that the correct answer is wool, contradicting the generated answer itself. Our trained model is able to correct the answer and generate the corresponding explanation.

Ablation study Table 6 shows the ablation study results on auto weights (w/o AutoW) part and RL alignment part (w/o RL). In the ablation study removing AutoW, loss weights are fixed at the initialization values of AutoW. In general, there is some drop in accuracy when dropping these parts.

4.4. Comparing with human-written explanations

We also conduct a comparison between models distilled from LLMs and models trained using humanwritten explanations. The experiment maintained the same settings. The only difference is the annotated explanations to be used. The results are presented in Table 7. In both $I \rightarrow O, I \rightarrow R$ and our alignment methods, the model distilled from ChatGPT exhibited superior performance in terms of both task accuracy and explanation consistency, surpassing the other two manually annotated methods. This observation suggests that despite the annotation incorrectness, ChatGPT is capable of

		Accuracy	LAS score	Prob score	BERTscore
WinoGrande	Ours	63.2411	-0.0868	0.6824	0.8194
	w/o w/o AutoW	60.7115	-0.1575	0.6358	0.8223
	w/o RL	62.6877	-0.1237	0.6339	0.8216
CQA	Ours	60.1982	0.2013	0.6385	0.7185
	w/o w/o AutoW	59.7853	0.1759	0.5399	0.6839
	w/o RL	59.4550	0.1894	0.6355	0.7193
COPA	Ours	71.6	-0.3214	0.6683	0.7731
	w/o AutoW	71.4	-0.3040	0.6372	0.7729
	w/o RL	69.2	-0.3249	0.4550	0.7411

Table 6: Ablation study.

achieving commendable results, possibly due to the wealth of information contained in its explanations. Moreover, when comparing our alignment method with the conventional $I \rightarrow O, I \rightarrow R$ framework, we observed no significant improvement in explanation quality on the two manually annotated datasets. This lack of improvement can be attributed to the fact that the manually annotated explanations already demonstrate consistency with the label annotations, as these annotations are all based on the correct answers. Therefore, the additional consistency alignment method proves unnecessary in these cases.

5. Conclusion

Given the remarkable performance of LLMs across various NLP tasks, leveraging LLMs for label and explanation annotations offers convenience and cost-effectiveness. However, an inherent problem

Model	Explanation	Accuracy	LAS	Prob score
$I \to O, I \to R$	ECQA	58.8770	0.1271	0.1944
	COS-E	58.2989	0.1389	0.1177
	ChatGPT	59.2898	0.1905	0.6330
RL align	ECQA	58.6292	0.1278	0.1890
	COS-E	59.8679	0.1368	0.1239
	ChatGPT	60.7762	0.1923	0.6234

Table 7: Results of models trained on explanations generated by ChatGPT and human-written explanations.

arises from the annotation incorrectness of LLMs. To address this issue, we introduce the Distillation with Explanations method. Our proposed method effectively leverages these annotations by utilizing a consistency model. Through empirical evaluation, we demonstrate that Distillation with Explanations enhances task output accuracy and improves the consistency between explanations and answers.

6. Limitations

Firstly, the choice of LLM prompts for generating answers and explanations are important, as the quality of the prompt greatly influences the text generated by LLMs. We do not include discussions about this topic, as identifying the optimal prompt for each specific task can yield another challenge. There have been some studies and attempts on ChatGPT prompts⁶, which may provide insights for better prompting. Secondly, our method relies on the availability of ground truth task labels. While there are numerous datasets with task labels on various NLP tasks, applying our method to new specific tasks in real-world scenarios may prove challenging. Future research should focus on investigating alternative approaches to distilling directly from LLMs, possibly leveraging the confidence estimates provided by the LLMs themselves.

7. Acknowledgements

The research work is supported by National Key R&D Plan No. 2022YFC3303303, the National Natural Science Foundation of China under Grant No.61976204, and the CAAI Huawei MindSpore Open Fund. Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS, Beijing Nova Program 20230484430.

Ethics Statement

The data used to train LLMs, such as ChatGPT, encompasses a complex and diverse range of text, which may include content with ethical considerations. As our model is distilled from these LLMs, there is a possibility that our model may be influenced by the ethical aspects embedded within the LLMs. Therefore, it is crucial to enhance the ethical dimensions within LLMs.

- Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065.
- David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12592– 12601.
- Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2021. Towards explainable conversational recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2994–3000.
- Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-attentive multi-task learning for explainable recommendation. In *IJCAI*, volume 2019, pages 2137–2143.
- Noptanit Chotisarn, Sarun Gulyanon, Tianye Zhang, and Wei Chen. 2023. Vishien-maat: Scrollytelling visualization design for explaining siamese neural network concept to non-technical users. *Visual Informatics*, 7(1):18–29.

⁶https://github.com/f/awesome-chatgpt-prompts

- Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings* of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, page 81–87, New York, NY, USA. Association for Computing Machinery.
- Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. 2019. Explainable recommendation through attentive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3622–3629.
- Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.
- Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings* of the Association for Computational Linguistics: *EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradientdisentangled embedding sharing.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. Cite arxiv:1503.02531Comment: NIPS 2014 Deep Learning Workshop.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.
- Julien Klaus, Mark Blacher, Andreas Goral, Philipp Lucas, and Joachim Giesen. 2023. A visual analytics workflow for probabilistic modeling. *Visual Informatics*, 7(2):72–84.

- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models.*
- Xiaoyan Kui, Naiming Liu, Qiang Liu, Jingwei Liu, Xiaoqian Zeng, and Chao Zhang. 2022. A survey of visual analytics techniques for online education. *Visual Informatics*, 6(4):67–77.
- Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.
- Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.
- Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.
- Hui Liu, Qingyu Yin, and William Yang Wang. 2019a. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason.
- Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2022. Few-shot selfrationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424.
- Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan.

2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- 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.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA:
 A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zheng Tang, Gus Hahn-Powell, and Mihai Surdeanu. 2020. Exploring interpretability in event extraction: Multitask learning of a neural event classifier and an explanation decoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 169–175, Online. Association for Computational Linguistics.

- PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. Pinto: Faithful language reasoning using prompt-generated rationales. In Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.
- Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings* of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10266– 10284.
- Ronald J Williams. 1992. Simple statistical gradientfollowing algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32.
- Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy. Association for Computational Linguistics.
- 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*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.