

Learning to Correct for QA Reasoning with Black-box LLMs

Anonymous ACL submission

Abstract

An open challenge in recent machine learning is about how to improve the reasoning capability of large language models (LLMs) in a black-box setting, *i.e.*, without access to detailed information such as output token probabilities. Existing approaches either rely on accessibility (which is often unrealistic) or involve significantly increased train- and inference-time costs. This paper addresses those limitations or shortcomings by proposing a novel approach, namely COBB (Correct for improving QA reasoning of Black-Box LLMs). It uses a trained adaptation model to perform a seq2seq mapping from the often-imperfect reasonings of the original black-box LLM to the correct or improved reasonings. Specifically, the adaptation model is initialized with a relatively small open-source LLM and adapted over a collection of sub-sampled training pairs. To select the representative pairs of correct and incorrect reasonings, we formulated the dataset construction as an optimization problem that minimizes the statistical divergence between the sampled subset and the entire collection, and solved it via a genetic algorithm. We then train the adaptation model over the sampled pairs by contrasting the likelihoods of correct and incorrect reasonings. Our experimental results demonstrate that COBB significantly improves reasoning accuracy across various QA benchmarks, compared to the best-performing adaptation baselines.¹

1 Introduction

Large language models (LLMs) have achieved significant advancements in various NLP tasks, demonstrating exceptional capabilities in understanding and generating text (Anthropic, 2024; OpenAI, 2023; Gemini et al., 2023; Touvron et al., 2023). Nevertheless, LLMs still present notable limitations, such as biased opinions toward specific groups (Santurkar et al., 2023) or inaccurate predictions for infrequent topics (Kandpal et al., 2023),

¹We will release the code and data upon acceptance.

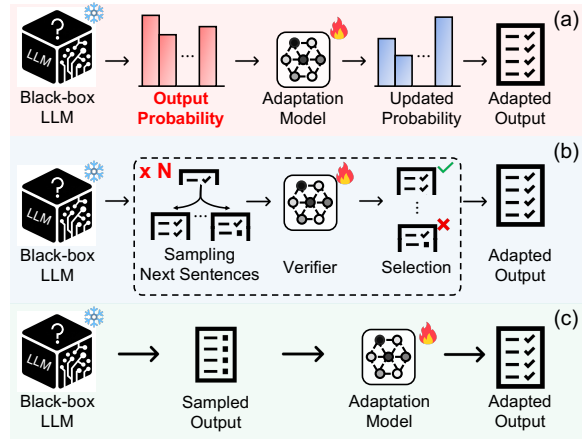


Figure 1: **Different black-box LLM adaptation methods.** (a) a model relying on the availability of output token probabilities; (b) a model with increased train- and inference-time costs; (c) COBB (proposed), not requiring output probabilities and is cost-efficient.

primarily due to the imperfections in the knowledge acquired during pre-training (Yao et al., 2023). Consequently, it is essential to control and adapt the responses of LLMs to achieve optimal performance for specific use cases. Representative methods include fine-tuning on supervised training datasets (Roziere et al., 2023; Azerbayev et al., 2024) and input-level optimization through prompt engineering and retrieval augmentation (Yang et al., 2024; Kim et al., 2024). However, these approaches require huge training costs or exhibit limited adaptation performance, respectively.

To address these challenges, prior works have focused on training relatively smaller models using responses from LLMs and human supervision, then generating adapted responses while assuming that the LLM parameters are fixed or inaccessible (*i.e.*, black-box). One approach assumes that the output token probabilities are available (Sun et al., 2022; Ormazabal et al., 2023), but this is often unrealistic. Although Sun et al. (2024) recently proposed training a verifier and employing beam search to

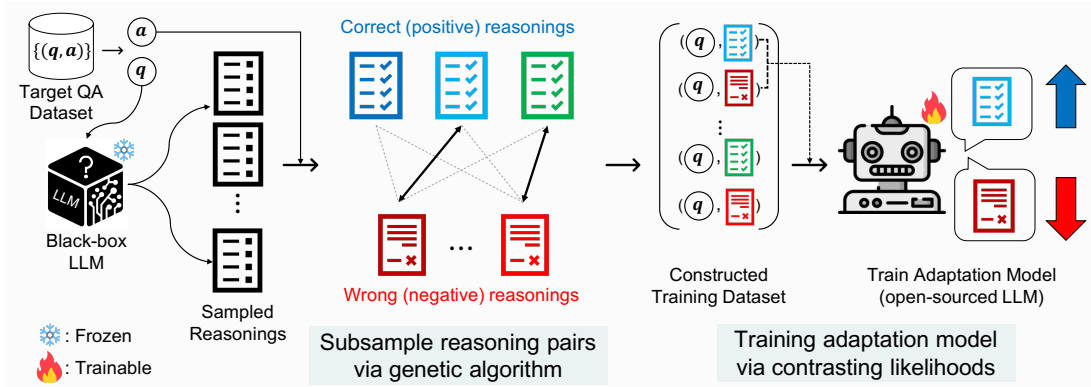


Figure 2: **An overview of COBB.** COBB first collects the multiple reasonings from black-box LLM, and labels them based on the correctness. Among all possible pairs of correct (positive) and incorrect (negative) reasonings, COBB subsample a few pairs that can maintain the characteristic of the entire set. Then, the adaptation model, initialized with an open-sourced LLM, is trained to increase/decrease the likelihood of positive/negative reasonings.

064 obtain adapted responses without this assumption, 100
 065 this method results in computationally expensive 101
 066 training and inference pipelines. Alternatively, Ji 102
 067 et al. (2024) introduced a straightforward seq2seq 103
 068 learning framework to enhance the alignment of 104
 069 black-box LLMs. However, extending this frame- 105
 070 work to other tasks is challenging, particularly in 106
 071 terms of constructing the training dataset and en- 107
 072 suring the effectiveness of the training method. 108

073 In this paper, we propose a simple yet efficient 109
 074 framework, learning to **Correct** for QA reasoning of 110
 075 **Black-Box LLMs** (COBB). Our key idea is to learn 111
 076 a seq2seq mapping from the original reasoning of 112
 077 black-box LLM to correct and improved reasoning, 113
 078 by training an adaptation model initialized with 114
 079 a relatively small open-source LLM. After training, 115
 080 the adaptation model can be easily deployed 116
 081 during inference as a single additional module, as 117
 082 illustrated in Figure 1.

083 Specifically, we firstly sample multiple chain- 118
 084 of-thought reasonings from black-box LLM and 119
 085 label their correctness using ground-truth human 120
 086 labels. Then, from all possible pairs of correct and 121
 087 incorrect reasonings, we subsample a few repre- 122
 088 sentative pairs that preserve the characteristics of 123
 089 the entire set. To identify such a subset, we formulate 124
 090 an optimization problem that minimizes the statisti- 125
 091 cal divergence between the subset and the entire 126
 092 set, solving it via a genetic algorithm. Finally, using 127
 093 this optimized subset, we train the adaptation 128
 094 model to simultaneously increase the likelihood of 129
 095 correct reasoning and decrease the likelihood of in- 130
 096 correct reasoning for the given input and reasoning. 131
 097 An overview of COBB is presented in Figure 2.

098 We demonstrate the effectiveness of COBB in 132
 099 improving QA reasoning with black-box LLMs 133

through extensive evaluations on four different QA 100
 datasets. For instance, COBB achieved average 101
 accuracy improvements of 6.2% and 2.2%, com- 102
 pared to the original black-box gpt-3.5-turbo 103
 and previous state-of-the-art adaptation methods, 104
 respectively. Furthermore, we found that the adap- 105
 tation model trained for a specific black-box LLM 106
 could generalize to adapt other LLMs, including 107
 both API-based and open-source models, which is 108
 crucial for efficient deployment in practice. Addi- 109
 tionally, our in-depth analyses reveal how COBB 110
 improves and corrects the reasoning from the black- 111
 box LLMs. We hope our work provides valuable 112
 insights into LLM adaptation research, which is 113
 becoming increasingly important for the future suc- 114
 cess of LLMs in real-world applications. 115

2 Related Works 116

2.1 Steering and adapting LLMs' responses 117

While recent LLMs have demonstrated remark- 118
 able success in various tasks, steering and adapt- 119
 ing their responses for the specific domain or 120
 user is still essential for achieving optimal perfor- 121
 mance (Santurkar et al., 2023; Salemi et al., 2023; 122
 Kandpal et al., 2023). Fine-tuning on human- or 123
 machine-labeled datasets is a straightforward ap- 124
 proach (Roziere et al., 2023; Azerbayev et al., 2024; 125
 Tan et al., 2024), but this method incurs significant 126
 costs due to the need to update the vast number of 127
 trainable model parameters, particularly for large- 128
 scale LLMs like GPT-4 (OpenAI, 2023) (>100B 129
 parameters). Consequently, prompt engineering 130
 (Kojima et al., 2022; Yang et al., 2024) and retrieval 131
 augmentation (Kim et al., 2024; Shi et al., 2024) 132
 are often preferred, as these methods only require 133

modifying the inputs to LLMs. However, recent observations indicate that input-level modifications alone are insufficient for adequately steering LLMs’ responses in the desired direction (Santurkar et al., 2023), likely due to the absence of learnable parameters and learning from human supervision. In this work, we propose an alternative way to steer and adapt LLMs using a trainable model and human supervision, without updating the target LLMs.

2.2 Learning to adapt black-box LLMs

As the scale of LLMs continues to increase, and their parameters often remain inaccessible (*i.e.*, black-box), the need to adapt their responses without updating their parameters has gained significant attention. A common approach involves introducing a relatively small trainable model to learn adaptation from the original responses of the black-box LLM. One line of work focuses on learning to adapt output probabilities (Sun et al., 2022; Ormazabal et al., 2023; Lu et al., 2023; Liu et al., 2024), but this method is impractical when the output probabilities of black-box LLMs are inaccessible. To address this limitation, Sun et al. (2024) propose a verification-based approach, generating the adapted responses in multiple steps via beam search, where scores are calculated using a learned verifier. However, this method increases the costs of training and inference due to the iterative computation between the black-box LLM and the verifier, and deploying the beam search. On the other hand, Ji et al. (2024) demonstrate that a simple seq2seq modeling approach can effectively improve the alignment of black-box LLMs. Despite its effectiveness, this method is limited for the other tasks, in terms of constructing the training dataset and ensuring the effectiveness of the training method. To overcome these limitations, we propose a novel approach to construct an effective training dataset, along with an improved training objective.

3 COBB: Learning to Correct for QA Reasoning with Black-box LLMs

In this section, we introduce our framework for learning to **Correct** for improving QA reasoning with **Black-Box** LLMs (COBB). We begin with an overview of the problem setup in Section 3.1. Next, in Section 3.2, we present how to construct an effective dataset for training the adaptation model. This dataset is created by solving an optimization problem using a genetic algorithm, to preserve the char-

acteristics of the entire set of correct and incorrect reasoning pairs from black-box LLM. Finally, we describe a training scheme in Section 3.3, where the adaptation model is trained by contrasting the likelihoods of positive and negative reasonings. The full procedure of COBB is outlined in Algorithm 1, and an overview is provided in Figure 2.

3.1 Preliminaries

Let denote black-box LLM as \mathcal{M} , which generates an original output sequence (*e.g.*, reasoning) \mathbf{y}_o for a given input sequence (*e.g.*, question) \mathbf{x} , *i.e.*, $\mathbf{y}_o \sim \mathcal{M}(\cdot|\mathbf{x})$. Then, our goal is to obtain an adaptation model π_θ , that can generate the adapted output (*e.g.*, improved reasoning) \mathbf{y}_a from given \mathbf{x} and \mathbf{y}_o :

$$\mathbf{y}_a \sim \pi_\theta(\cdot|\mathbf{x}, \mathbf{y}_o). \quad (1)$$

For example, Ji et al. (2024) initialize π_θ with a pre-trained open-sourced LLM, and fine-tune it by minimizing a supervised cross-entropy:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\log \pi_\theta(\mathbf{y}_a|\mathbf{y}_o, \mathbf{x}), \quad (2)$$

where $\mathbf{x}, \mathbf{y}_o, \mathbf{y}_a \sim \mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}_o^i, \mathbf{y}_a^i)\}_{i=1}^N$. To improve the LLM’s alignment regarding helpfulness and harmlessness, Ji et al. (2024) construct \mathcal{D} using weaker LLMs (*e.g.*, Alpaca-7B) for \mathbf{y}_o and stronger LLMs (*e.g.*, GPT-4) or human annotations for \mathbf{y}_c , respectively. In our case, we assume that a human-annotated QA dataset $\mathcal{Q} = \{(\mathbf{q}^i, \mathbf{a}^i)\}_{i=1}^M$ is available, where \mathbf{q} is question of target task and \mathbf{a} is the ground-truth answer. Then, our goal is to train adaptation model π_θ , which is also initialized with open-sourced LLM, using \mathcal{Q} and obtain the improved reasoning with \mathcal{M} for this task.

3.2 Optimizing dataset to learn from effective reasoning pairs via genetic algorithm

Collecting and labeling of training pairs. To train adaptation model π_θ using \mathcal{Q} , we first collect positive and negative reasonings from \mathcal{M} . Specifically, for each $\mathbf{q}, \mathbf{a} \sim \mathcal{Q}$, we sample K different reasonings $\{\mathbf{y}_{\text{cot},k}\}_{k=1}^K$ using few-shot chain-of-thought prompt \mathbf{p}_{cot} (Wei et al., 2022):

$$\mathbf{y}_{\text{cot},k} \sim \mathcal{M}(\cdot|\mathbf{q}, \mathbf{p}_{\text{cot}}). \quad (3)$$

Then, if the prediction by \mathbf{y}_{cot} is *correct* (*i.e.*, equal to the answer \mathbf{a}), we assign this reasoning to the *positive* reasoning set, \mathcal{Y}_{pos} . *If not*, we assign this reasoning to the *negative* reasoning set, \mathcal{Y}_{neg} . Remarkably, we denote that there are some cases

whether (1) \mathcal{M} can't generate any correct reasoning (i.e., $\mathcal{Y}_{\text{pos}} = \emptyset$) or (2) there is no incorrect reasoning (i.e., $\mathcal{Y}_{\text{neg}} = \emptyset$). For (1), we utilize answer-augmented prompting (Zelikman et al., 2022) to generate the reasoning to support the given answer \mathbf{a} and augment \mathcal{Y}_{pos} with it. For (2), we randomly select the reasoning of another sample and augment \mathcal{Y}_{neg} with it, to fully utilize the samples in \mathcal{Q} .

Solving optimization to find effective reasoning pairs via genetic algorithm. With the collected \mathcal{Y}_{pos} and \mathcal{Y}_{neg} , we want to construct the training dataset \mathcal{D} to train π_θ , composed of the triplet of the question \mathbf{q} , positive reasoning \mathbf{y}_p , and negative reasoning \mathbf{y}_n . However, the number of possible combinations between positive and negative reasonings is quadratically increased, i.e., $|\mathcal{Y}_{\text{pos}}| \times |\mathcal{Y}_{\text{neg}}|$; it can be too large trained within the limited iterations and there can be large redundancy within the constructed dataset. To tackle this challenge, we propose to subsample a few representative positive and negative reasoning pairs, that can preserve the characteristics of all the possible combinations.

Specifically, for each \mathbf{q} , let denote the set of all the possible pairs of positive and negative reasonings as $\mathcal{P} = \mathcal{Y}_{\text{pos}} \times \mathcal{Y}_{\text{neg}}$. Then, for each pair in \mathcal{P} , we calculate its likelihood difference under π_θ :

$$\mathbf{P} = \{\pi_\theta(\mathbf{y}_p|\mathbf{q}) - \pi_\theta(\mathbf{y}_n|\mathbf{q}) \mid \mathbf{y}_p, \mathbf{y}_n \in \mathcal{P}\}. \quad (4)$$

Then, we propose to find a subset $\mathcal{P}_{\text{sub}} \subset \mathcal{P}$ which minimizes $d(\mathbf{P}_{\text{sub}}, \mathbf{P})$, where \mathbf{P}_{sub} is obtained from \mathcal{P}_{sub} similar to Eq. 4 and $d(\cdot, \cdot)$ is a distance between two sets. Here, we assume the elements of both \mathbf{P} and \mathbf{P}_{sub} are samples from two different normal distributions and then consider 2-Wasserstein distance (Givens and Shortt, 1984) between them:

$$d(\mathbf{P}_{\text{sub}}, \mathbf{P}) = (\mu - \mu_{\text{sub}})^2 + (\sigma - \sigma_{\text{sub}})^2, \quad (5)$$

where μ, σ^2 are the empirical mean and variance of \mathbf{P} and $\mu_{\text{sub}}, \sigma_{\text{sub}}^2$ the empirical mean and variance of \mathbf{P}_{sub} , respectively. We empirically observe that this 2-Wasserstein distance is better than other possible metrics such as KL divergence.

However, finding \mathbf{P}_{sub} that minimizes the distance (Eq. 5) is non-trivial, as this selection of representative samples problem is NP-hard (Gamez et al., 2008), and the current objective includes the quadratic terms. To mitigate these challenges, we use a genetic algorithm (Holland, 1992), which progressively optimizes the solution by iterating (1) acquiring a new candidate by perturbing the current solution and (2) updating the solution when the

Algorithm 1 COBB algorithm

Input: Black-box LLM \mathcal{M} , adaptation model π_θ , target QA dataset $\mathcal{Q} = \{\mathbf{q}^i, \mathbf{a}^i\}_{i=1}^M$, number of sampling K , genetic algorithm iterations T , training iteration T_{train} , learning rate η

```

/* Construction of dataset  $\mathcal{D}$  */
 $\mathcal{D} = \emptyset$ 
for  $\mathbf{q}, \mathbf{a} \in \mathcal{Q}$  do
  /* Collect reasonings */
   $\mathcal{Y} = \{\mathbf{y}_{\text{cot},k}\}_{k=1}^K, \mathbf{y}_{\text{cot},k} \sim \mathcal{M}(\cdot|\mathbf{q}, \mathbf{p}_{\text{cot}})$ 
  /* Get reasoning pairs */
   $\mathcal{P} = \mathcal{Y}_{\text{pos}} \times \mathcal{Y}_{\text{neg}} \leftarrow \mathcal{Y}$  with  $\mathbf{a}$ 
   $\mathbf{P} \leftarrow \mathcal{P}$  with  $\pi_\theta$  (Eq. 4)
  /* Subset selection */
   $\mathcal{P}_{\text{sub}} \leftarrow \text{Genetic}(\mathcal{P}, \mathbf{P}, |\mathcal{Y}_{\text{neg}}|, T)$  (Eq. 6)
   $\mathcal{D} \leftarrow \mathcal{D} \cup (\{\mathbf{q}\} \times \mathcal{P}_{\text{sub}})$ 
end for
/* Train adaptation model  $\pi_\theta$  */
for  $t = 0$  to  $T_{\text{train}} - 1$  do
  /* Loss for mini-batch */
   $\mathcal{L}_{\text{train}}(\theta, \mathcal{B}) \leftarrow \mathcal{B} \sim \mathcal{D}, \pi_\theta$  (Eq. 7)
  /* Update adaptation model */
   $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{train}}$ 
end for
return trained adaptation model  $\pi_\theta$ 

```

candidate achieves a better optimization objective. We consider a new random sampling of \mathcal{P}_{sub} as the perturbation, and obtain $\mathcal{P}_{\text{sub}}^*$ after T iterations:

$$\mathcal{P}_{\text{sub}}^* = \text{Genetic}(\mathcal{P}, \mathbf{P}, M, T), \quad (6)$$

where M is the size of the subset and a detailed description is presented in Algorithm 2.² We observed that the genetic algorithm found a good solution within a few iterations and it only requires small additional computations (see Table 8). With $\mathcal{P}_{\text{sub}}^*$, we construct the dataset $\mathcal{D} = \{(\mathbf{q}^i, \mathbf{y}_n^i, \mathbf{y}_p^j)\}_{i=1}^N$, where $(\mathbf{y}_n, \mathbf{y}_p) \in \mathcal{P}_{\text{sub}}^*$ for \mathbf{q} .³

3.3 Learning to correct by contrasting likelihoods of reasoning pairs

With the constructed dataset \mathcal{D} , we train the adaptation model π_θ to learn the seq2seq mapping from the original reasoning from black-box LLM \mathcal{M} to the correct and improved reasoning. While the supervised training with a cross-entropy (Eq. 2) is considerable (Ji et al., 2024), we observed that

²During the experiments, we fix M as $|\mathcal{Y}_{\text{neg}}|$.

³We remark that there can be duplicated \mathbf{q}^i , as multiple reasoning pairs are constructed for each \mathbf{q}, \mathbf{a} from \mathcal{Q} .

this approach could be limited, especially when the target task requires careful discrimination between positive and negative reasonings. Therefore, we propose to further use the negative reasoning \mathbf{y}_n to lower its likelihood in the output space of π_θ , while simultaneously increasing the likelihood of the positive reasoning \mathbf{y}_p . Specifically, we construct our training objective $\mathcal{L}_{\text{train}}$ using Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024), which enables single-stage learning from pair-wise preference data, without the reference models. Namely, we treat \mathbf{y}_p as preferred output and \mathbf{y}_n as dispreferred output:

$$\mathcal{L}_{\text{train}}(\theta, \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[\mathcal{L}_{\text{SFT}}(\theta) + \lambda \cdot \mathcal{L}_{\text{OR}}(\theta)], \quad (7)$$

$$\mathcal{L}_{\text{OR}}(\theta) = -\log \sigma \left(\log \frac{\text{odds}_\theta(\mathbf{y}_p|\mathbf{x})}{\text{odds}_\theta(\mathbf{y}_n|\mathbf{x})} \right), \quad (8)$$

where σ is a sigmoid function, λ is a hyper-parameter, and $\text{odds}_\theta(\mathbf{y}|\mathbf{x}) = \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{1-\pi_\theta(\mathbf{y}|\mathbf{x})}$. Here, we use the concatenation of question \mathbf{q} and reasoning \mathbf{y} (for both \mathbf{y}_p and \mathbf{y}_n) as the input \mathbf{x} , to model the seq2seq mapping between the original reasoning from \mathcal{M} (input) and the refined reasoning through π_θ (output), conditioned on \mathbf{q} . As shown in Figure 3, incorporating \mathbf{y}_n via Eq. 7 effectively suppresses the increasing likelihood of negative reasonings.

4 Experiments

4.1 Setups

Datasets and metrics. Following the recent work (Sun et al., 2024), we evaluate COBB on four different question-answering (QA) tasks, requiring adaptation on mathematical (*GSM8K*), implicit-reasoning (*StrategyQA*), truthful (*TruthfulQA*), and scientific (*ScienceQA*) domains. We use the train and test splits by Sun et al. (2024). To generate the reasonings for each dataset, we follow the previous chain-of-thought prompts used in prior work (Sun et al., 2024), except *GSM8K*. In the case of *GSM8K*, we adopt a complex prompt (Fu et al., 2023), as it yields higher accuracy compared to the previous one. During the evaluation, we sample $K = 5$ chain-of-thought reasoning for each test question, and measure (1) the average accuracy (Avg.) across 5 reasonings, and (2) the accuracy of prediction from majority voting among them (*Maj@5*). For *TruthfulQA*, we report the average of the accuracies on helpfulness and informativeness (*True + Info*) following (Sun et al., 2024), along with the majority voted accuracy. More details of the datasets are in Appendix A.1.

Baselines. We compare COBB against several extensive baselines as follows: (1) *Target black-box LLM*: without adaptation, we use the reasoning from the target black-box LLM \mathcal{M} , (2) *Initial adaptation model*: we generate the reasoning from the open-sourced LLM, which is used to initialize the adaptation model π_θ , (3) *Supervised Fine-Tuning (SFT)*: π_θ is fine-tuned with a given QA dataset \mathcal{Q} , (4) *Chain-of-Thought Distillation (CoT Distill)* (Li et al., 2023a): instead of answer \mathbf{a} in original \mathcal{Q} , the positive reasoning \mathbf{y}_p is used as the output label for input \mathbf{q} to fine-tune π_θ . (5) *Aligner* (Ji et al., 2024): π_θ is fine-tuned to learn a seq2seq mapping from the concatenation of \mathbf{q} and \mathbf{y}_n to \mathbf{y}_p via cross-entropy loss (Eq. 2), (6) *BBox-Adapter* (Sun et al., 2024): learning a verifier model to deploy beam search and generate the adapted reasoning in iterative inference and verification steps.

Implementation details. For the target black-box LLM \mathcal{M} , we mainly consider *gpt-3.5-turbo-0125*, and it is used to generate the reasoning for the training adaptation model. To initialize the adaptation model π_θ , we consider *Mistral-7B-inst-v2* (Jiang et al., 2023). For *BBox-Adapter* (Sun et al., 2024), we follow the original experimental setups in the official codes. For other adaptation methods including COBB, we commonly fine-tune π_θ for 5 epochs with a batch size of 128, using an Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1×10^{-5} and cosine scheduler with a warm ratio of 0.03. Also, we use a temperature of 1.0 to sample the reasoning for each question. For the hyper-parameters of COBB, we used fixed values of $\lambda = 0.1, T = 1000, K = 10$. Here, we generate half of the reasonings from \mathcal{M} , and the remaining half from the initial π_θ for efficiency.

4.2 Main results

Table 1 summarizes the experimental results on four different QA tasks, by adapting the reasoning of *gpt-3.5-turbo* (i.e., target black-box LLM \mathcal{M}). First, it is observed that *Mistral-7b-inst-v2*, which is used to initialize the adaptation model π_θ , originally exhibits significantly lower performance than the target black-box LLM. However, the model’s performance is largely increased after the adaptation to the target task, regardless of the methods; it shows the importance of an additional adaptation stage for black-box LLM, using both the ground-truth human supervision and the collected reasonings of

Table 1: **Main results.** Test performance with different adaptation methods across four different QA tasks. Here, `gpt-3.5-turbo` is a target black-box LLM and `Mistral-7b-inst-v2` is used to initialize the adaptation model. The best and second best scores are highlighted in **bold** and underline, respectively.

Dataset (\rightarrow)	StrategyQA		GSM8K		TruthfulQA		ScienceQA	
	Avg.	Maj@5	Avg.	Maj@5	True + Info	Maj@5	Avg.	Maj@5
<i>Target Black-box LLM</i>	70.92	71.62	76.25	79.23	71.40	73.00	81.24	83.00
<i>Initial Adaptation Model</i>	61.40	62.88	43.50	51.40	75.40	78.50	65.52	66.60
<i>SFT</i>	66.11	66.38	49.67	55.19	61.40	65.00	<u>85.52</u>	<u>85.80</u>
<i>CoT Distill</i>	67.14	70.31	58.01	65.20	<u>80.00</u>	83.00	<u>77.08</u>	80.40
<i>Aligner</i>	58.69	59.83	76.42	79.83	79.60	<u>83.50</u>	74.92	81.00
<i>BBox-Adapter</i>	<u>71.27</u>	<u>73.36</u>	78.79	<u>83.09</u>	73.80	73.00	81.96	83.60
COBB (Ours)	74.93	75.11	<u>78.59</u>	85.14	82.90	85.50	88.00	89.20

Table 2: **Transferability with COBB.** Test performance of original and adapted reasonings of different LLMs. Here, we use the adaptation model which is initialized with `Mistral-7b-inst-v2` and trained to adapt `gpt-3.5-turbo`, and it is denoted by *. The best scores for each LLM are highlighted in **bold**.

Dataset (\rightarrow)	StrategyQA		GSM8K		TruthfulQA		ScienceQA	
	Avg.	Maj@5	Avg.	Maj@5	True + Info	Maj@5	Avg.	Maj@5
Claude-3-Haiku	72.05	72.93	83.85	83.17	67.90	68.00	82.00	82.40
Claude-3-Haiku + COBB*	72.58	76.42	81.73	86.66	81.20	83.50	87.40	88.80
Mistral-7B-inst-v2	61.40	62.88	43.50	51.40	75.40	78.50	65.52	66.60
Mistral-7B-inst-v2 + COBB*	70.31	74.24	56.03	66.57	84.10	86.00	85.56	87.70
Phi-3-mini-4k-inst	62.27	62.88	82.18	86.05	78.10	79.00	83.32	85.80
Phi-3-mini-4k-inst + COBB*	70.22	74.67	78.10	86.13	84.00	88.00	86.84	88.80
Gemma-1.1-7B-it	57.12	55.90	49.54	53.22	60.90	61.50	71.84	74.60
Gemma-1.1-7B-it + COBB*	72.66	73.36	61.85	69.14	82.00	84.50	87.12	88.40

the black-box LLM. In addition, among these adaptation methods, one can observe that COBB yields the largest improvements in most cases. Specifically, COBB exhibits 6.2%/7.0% average accuracy (Acc.) and the majority voted accuracy (Maj@5) improvements for the target black-box LLM, on average across 4 QA tasks. Furthermore, compared to the strongest baselines, COBB exhibits 2.2%/2.3% average improvements, respectively.

Remarkably, as shown in Table 3, COBB requires much smaller costs during the training of the adaptation model ($\approx 20\%$) and the test-time inference ($\approx 7\%$), compared to the previous state-of-the-art method (*BBox-Adapter*).⁴ This is because COBB directly learns a seq2seq modeling while *BBox-Adapter* learns to verify through the sampling and beam search. These results indicate that COBB could serve as a more powerful yet cost-efficient adaptation method.

⁴We follow the official implementation and hyperparameters by the authors in <https://github.com/haotiansun14/BBox-Adapter>.

We further demonstrate the advantage of COBB regarding the transferability to various LLMs; namely, we deploy the adaptation model, trained with `gpt-3.5-turbo` (in Table 1), to adapt reasonings of other LLMs including other API-based black-box LLM (*Claude-3-Haiku* (Anthropic, 2024)) and open-source LLMs (*Mistral-7B-inst-v2* (Jiang et al., 2023), *Phi-3-mini-4k-inst* (Abdin et al., 2024), *Gemma-1.1-7B-it* (Team et al., 2024)).⁵ This result is presented in Table 2. Here, one can observe that COBB successfully adapts the reasoning of various LLMs and improves the accuracies overall, even without the specific adaptation to the target LLM. To be specific, COBB exhibits 9.1%/11.1% average accuracy (Acc.), and the majority voted accuracy (Maj@5) improvements, on average across 4 LLMs and 4 QA tasks. On the other hand, it is observed that the average accu-

⁵For open-source LLMs, we only use the generated reasoning without access to the internal model weights or output probabilities to treat them as black-box LLM.

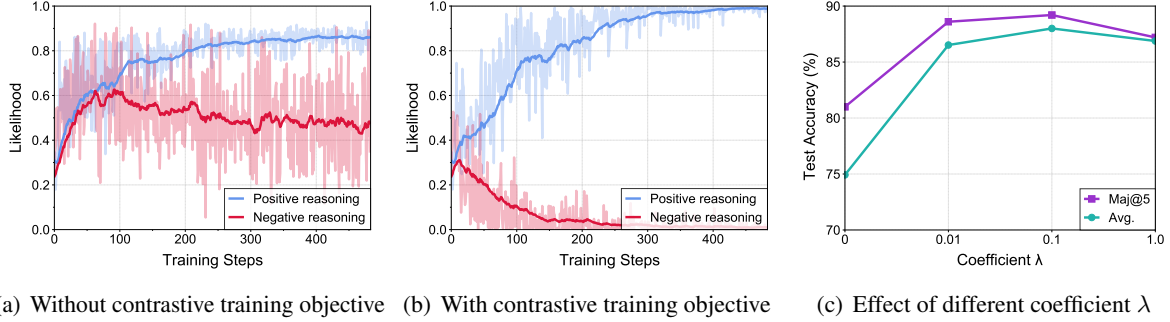


Figure 3: **Effect of contrasting likelihoods.** Change of the likelihood of π_θ for positive and negative reasonings in the training dataset (a) without / (b) with the contrastive training objective (Eq. 8) on ScienceQA. (c) Test accuracy of the adapted reasonings of `gpt-3.5-turbo` on ScienceQA with varied coefficient λ .

Table 3: **Cost efficiency of COBB.** Costs consumed for training and evaluation of BBox-Adapter and COBB. For the evaluation, we sample 5 reasoning per sample. Costs are calculated with the official per-token price.

Methods	StrategyQA		ScienceQA	
	Train (\$)	Eval (\$)	Train (\$)	Eval (\$)
BBox-Adapter	5.92	1.87	5.98	3.90
CoBB	1.25	0.14	1.01	0.25

Table 4: **Ablation study.** Average test accuracy of adapted reasoning of `gpt-3.5-turbo` with different configurations of the proposed components in COBB.

Methods	Con.	Gen.	StrategyQA	ScienceQA
Aligner	✗	✗	58.69	74.92
	✗	✓	59.30	73.36
CoBB	✓	✗	71.35	86.60
	✓	✓	74.93	88.00

racy on GSM8K is slightly decreased when the target LLM already exhibits a stronger performance than the LLM used to generate the training data. From this result and the overall improvements with the transferred adaptation model, it is inferred that the knowledge included in the constructed training dataset is more important for the effectiveness of the adaptation model, rather than the specific type of LLM used to construct the data. We present the results with the standard deviation in Appendix B.

4.3 Additional analyses with COBB

In this section, we provide additional analyses of COBB. We conduct the experiments on StrategyQA and ScienceQA, by setting `gpt-3.5-turbo` as the target black-box LLM \mathcal{M} and `Mistral-7b-inst-v2` as the initialization model for the adaptation model π_θ in default. **Ablation study.** To validate the effectiveness of the proposed components of COBB in Section 3, we perform the ablation experiments by decomposing our framework with two components of (1) the dataset construction via genetic algorithm (Eq. 6) and (2) the training objective to contrast the likelihood of positive and negative reasonings (Eq. 8). We denote these components as *Gen.* and *Con.*, respectively. For comparison, we consider random subsampling when the genetic algorithm selection is not applied. Additionally, we set $\lambda = 0$ when

the contrastive training objective is not used.

The results are presented in Table 4. Here, it can be observed that using the contrastive training objective significantly improves the accuracy of the adapted reasoning, and the improvements are further enhanced when the adaptation model is trained on more representative reasoning pairs. At the same time, it is observed that the proposed dataset construction is not effective without the contrastive training objective. These results indicate that adjusting the likelihood of π_θ is crucial to successfully learning the adaptation, and effective dataset construction aids by guiding *where to adjust*. We further present Figure 3 to reveal the effect of contrastive training objective. Here, one can observe that the likelihood of negative reasoning is even increased compared to the initial stage, when the cross-entropy loss is only used with the positive reasoning (Eq. 2). However, by incorporating the contrastive objective, this problem is clearly resolved. One can also observe that its effectiveness is not sensitive to the choice of λ .

Effect of different initialization for π_θ . Next, we conduct experiments to reveal the importance of the choice of open-sourced LLM to initialize π_θ . To this end, we use LLaMA2 (LLaMA-2-7B-chat-hf), which has a similar number of trainable parameters as the originally used Mistral (Mistral-7B-inst-v2), for the

Question: Which is a sentence fragment?
Choices: ["Our car was stuck in traffic we were late.", "A strong, bold girl."]
Originally incorrect reasoning from gpt-3.5-turbo: Our car was stuck in traffic we were late. - This is a sentence fragment because it does not have a subject and a complete predicate. A strong, bold girl. - This is a complete sentence with a subject (girl) and a predicate (strong, bold). ##### 0
Corrected reasoning by COBB: The passage describes the climate. Climate refers to the typical weather conditions of a region over a long period, while weather refers to the conditions at a specific time. In this passage, Rita's location being described as where "winds blowing from the northeast are rare in July" indicates a recurring pattern of wind direction and strength over a longer period known as climate. ##### 1

Figure 4: **Qualitative example on ScienceQA.** Example of the question, original reasoning from black-box LLM (gpt-3.5-turbo), and the adapted reasoning by COBB. More examples are presented in Appendix C.

Table 5: **Different initialization for adaptation model.** Average test accuracy with different open-source LLMs for the initialization of the adaptation model π_θ . StQA is StrategyQA and ScQA is ScienceQA, respectively.

Initialization	COBB	StQA	ScQA
No adaptation	N/A	70.92	81.24
LLaMA-2-7B-chat-hf	✗	56.24	46.48
	✓	63.13	76.84
Mistral-7B-inst-v2	✗	61.40	65.52
	✓	74.93	88.00

Table 6: **In-depth analyses of COBB.** Analyses to deeper understand of how COBB works. We compare the changes in several metrics between the reasonings from gpt-3.5-turbo (original) and the reasonings by COBB (adapted) on ScienceQA.

Metrics	gpt-3.5-turbo	COBB
Acc. of orig. correct (\uparrow)	100.0	92.22
Acc. of orig. incorrect (\uparrow)	0.000	69.72
$\pi_\theta(\mathbf{y})$ of orig. correct (\uparrow)	0.767	0.756
$\pi_\theta(\mathbf{y})$ of orig. incorrect (\uparrow)	0.264	0.685
Cosine similarity (\downarrow)	0.910	0.906
Self-BLEU (\downarrow)	0.472	0.503

initialization and measure the average accuracy before/after applying COBB. The results are presented in Table 5; when COBB is applied (✓), it indicates that π_θ is trained with each initialization LLM and used to adapt the reasoning from gpt-3.5-turbo. One can first notice that the accuracy of LLaMA2 is largely worse than Mistral. While the accuracy of the adapted reasoning with LLaMA2 is significantly increased, it still fails to improve the accuracy of the reasonings from the black-box LLM, unlike Mistral. This result implies that pre-trained knowledge within the open-source LLM is crucial to learning the correction of QA reasoning via COBB, and we could benefit from the continued advances of open-source LLMs.

In-depth analyses of COBB. Lastly, we conduct additional analyses to deeply understand how COBB works. Specifically, we try to answer the following question: how COBB changes the (1) correctness, (2) likelihood, and (3) diversity of the reasonings of the black-box LLM. The corresponding experimental results are presented in the top, middle, and bottom rows of Table 6, respectively. First, it is observed that COBB mostly keeps the correctness of the originally correct reasonings (100 \rightarrow 92.2), while significantly improving the incorrect ones (0 \rightarrow 69.72). Also, such behavior is observed in terms of the likelihood; when we

measure the likelihood of reasoning \mathbf{y} with the trained adaptation model $\pi_\theta(\mathbf{y})$, one can observe that the likelihood of originally correct reasonings is maintained and incorrect reasonings' is largely increased. Then, one potential concern might be that COBB loses the diversity within the original reasonings, and generates the identical adapted reasonings. But, as shown in Table 6, it is observed that the diversity of original reasonings is well-preserved after the adaptation via COBB; it demonstrates that COBB can understand the context within the original reasoning and properly incorporate it during the adaptation.

5 Conclusion

In this paper, we proposed COBB, a simple yet effective framework for learning to correct QA reasoning of black-box LLM. We propose to learn a seq2seq mapping from the original reasoning of black-box LLM to correct and improved reasoning, by training a relatively small adaptation model with the newly proposed dataset construction and training objective. Our experiments demonstrate the effectiveness of COBB across various QA tasks and LLMs. Therefore, we believe our framework can contribute to various real-world applications that require the adaptation of black-box LLMs.

545 Limitations

546 While COBB shows promising results in our exper-
547 iments, several limitations must be acknowledged.
548 First, the effectiveness of COBB heavily depends
549 on the quality of the training pairs and the capa-
550 bility of the initial open-source LLM. While the
551 proposed dataset construction via genetic algorithm
552 aims to select representative pairs, the initial set
553 of collected reasonings might still be biased (San-
554 turkar et al., 2023) or insufficiently diverse (Wang
555 et al., 2023) depending on the black-box LLM used
556 for the reasoning generation, potentially affecting
557 the adaptation model’s performance. Moreover, the
558 effectiveness of our framework largely depends on
559 the specific open-source LLM used to initialize the
560 adaptation model, as shown in Table 5. While this
561 reliance may be seen as a limitation, it also high-
562 lights a strength of our framework, as it can benefit
563 from the rapid advancements in open-source LLM
564 development in recent days.

565 Secondly, COBB requires ground-truth human
566 labels to judge the correctness of reasonings, which
567 can be resource-intensive and time-consuming to
568 obtain, especially for large-scale datasets. Addi-
569 tionally, while COBB demonstrates transferability
570 across different LLMs, the adaptation performance
571 may vary based on the specific characteristics and
572 pre-training knowledge of the target LLMs. Lastly,
573 the computational efficiency of COBB, although
574 improved compared to the baselines, can still pose
575 challenges as it yields the fine-tuned open-source
576 LLMs per each task which has a large number of
577 model parameters. To address this issue, incorporat-
578 ing the parameter-efficient fine-tuning techniques
579 (Dettmers et al., 2023) or distillation into a smaller
580 model (Gu et al., 2024) could be effective.

581 Broader Impact and Ethical Implications

582 We strongly believe that COBB framework has
583 the potential to provide significant positive im-
584 pacts across various real-world applications. For
585 instance, depending on the user, the interested do-
586 main could be varied, such as education, health-
587 care, and finance (Gan et al., 2023; Clusmann et al.,
588 2023; Li et al., 2023b). However, as highlighted in
589 the recent study (Kandpal et al., 2023), the accuracy
590 of LLMs could be not sufficient if the considered
591 domain is less frequently trained. In such a case,
592 our framework offers an efficient solution for gener-
593 ating domain-specific responses without incurring
594 huge costs, compared to the conventional solution

of continual training (Singhal et al., 2023a,b).

At the same time, however, there are also some
potential negative impacts. A primary concern is
the risk of reinforcing existing biases present in the
training data, whether they originate from the tar-
get black-box LLM, the human-annotated datasets,
or the pre-trained knowledge of the open-source
LLM used for initialization. For example, recent
research has shown that state-of-the-art LLMs even
exhibit biases towards specific groups (Santurkar
et al., 2023). If this kind of undesired bias is not
properly removed during the training of the adap-
tation model, then our framework could reproduce
or amplify the bias. We believe that this problem
could be mitigated by incorporating additional fil-
tering stages during the dataset construction, train-
ing, or inference (Le Bras et al., 2020; Dong et al.,
2023; Zhang and Zhou, 2024), and we remain this
problem for the future direction.

References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan,
Jyoti Aneja, Ahmed Awadallah, Hany Awadalla,
Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harki-
rat Behl, et al. 2024. Phi-3 technical report: A highly
capable language model locally on your phone. *arXiv
preprint arXiv:2404.14219*.
- Anthropic. 2024. Introducing the next generation
of claude. [https://www.anthropic.com/
news/claude-3-family](https://www.anthropic.com/news/claude-3-family).
- Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster,
Marco Dos Santos, Stephen McAleer, Albert Q Jiang,
Jia Deng, Stella Biderman, and Sean Welleck. 2024.
Llemma: An open language model for mathematics.
In *International Conference on Learning Representa-
tions (ICLR)*.
- Jan Clusmann, Fiona R Kolbinger, Hannah Sophie
Muti, Zunamys I Carrero, Jan-Niklas Eckardt,
Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler,
Sophie-Caroline Schwarzkopf, Michaela Unger, Gre-
gory P Veldhuizen, et al. 2023. The future landscape
of large language models in medicine. *Communica-
tions medicine*, 3(1):141.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,
Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias
Plappert, Jerry Tworek, Jacob Hilton, Reiichiro
Nakano, et al. 2021. Training verifiers to solve math
word problems. *arXiv preprint arXiv:2110.14168*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and
Luke Zettlemoyer. 2023. Qlora: Efficient finetuning
of quantized llms. In *Advances in Neural Information
Processing Systems (NeurIPS)*.

646	Xiangjue Dong, Ziwei Zhu, Zhuoer Wang, Maria Teleki, and James Caverlee. 2023. Co ² pt: Mitigating bias in pre-trained language models through counterfactual contrastive prompt tuning. In <i>Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> .	700
647		701
648		702
649		
650		
651		
652	Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, Tushar Khot, and Wenhua Chen. 2023. Chain-of-thought hub: Measuring llms' reasoning performance. https://github.com/FranxYao/chain-of-thought-hub/blob/main/gsm8k/lib_prompt/prompt_hardest.txt .	703
653		704
654		705
655		706
656		707
657		
658		
659	J Esteban Gamez, François Modave, and Olga Koshelva. 2008. Selecting the most representative sample is np-hard: Need for expert (fuzzy) knowledge. In <i>2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence)</i> , pages 1069–1074. IEEE.	708
660		709
661		710
662		711
663		712
664		713
665	Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In <i>2023 IEEE International Conference on Big Data (BigData)</i> , pages 4776–4785. IEEE.	714
666		715
667		716
668		
669		
670	Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. <i>arXiv preprint arXiv:2312.11805</i> .	717
671		718
672		719
673		720
674		
675	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. <i>Transactions of the Association for Computational Linguistics</i> , 9:346–361.	721
676		722
677		723
678		724
679		725
680		
681	Clark R Givens and Rae Michael Shortt. 1984. A class of wasserstein metrics for probability distributions. <i>Michigan Mathematical Journal</i> , 31(2):231–240.	726
682		727
683		728
684	Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In <i>International Conference on Learning Representations (ICLR)</i> .	729
685		730
686		
687		
688	John H Holland. 1992. Genetic algorithms. <i>Scientific american</i> , 267(1):66–73.	731
689		732
690	Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. <i>arXiv preprint arXiv:2403.07691</i> .	733
691		734
692		
693	Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. 2024. Aligner: Achieving efficient alignment through weak-to-strong correction. <i>arXiv preprint arXiv:2402.02416</i> .	735
694		736
695		737
696		738
697		
698	Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. <i>arXiv preprint arXiv:2310.06825</i> .	739
699		740
		741
		742
		743
		744
		745
		746
		747
		748
		749
		750
		751
		752
		753
		754
		755

756	OpenAI. 2023. Gpt-4 technical report. <i>arXiv preprint arXiv:2303.08774</i> .	
757		
758	OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o .	
759		
760	Aitor Ormazabal, Mikel Artetxe, and Eneko Agirre. 2023. Comblm: Adapting black-box language models through small fine-tuned models. In <i>Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> .	
761		
762		
763		
764		
765	Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. <i>arXiv preprint arXiv:2308.12950</i> .	
766		
767		
768		
769		
770	Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. <i>arXiv preprint arXiv:2304.11406</i> .	
771		
772		
773		
774	Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In <i>Proceedings of the International Conference on Machine Learning (ICML)</i> .	
775		
776		
777		
778		
779	Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In <i>Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)</i> .	
780		
781		
782		
783		
784		
785		
786	Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. <i>Nature</i> , 620(7972):172–180.	
787		
788		
789		
790		
791	Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. <i>arXiv preprint arXiv:2305.09617</i> .	
792		
793		
794		
795		
796		
797	Haotian Sun, Yuchen Zhuang, Wei Wei, Chao Zhang, and Bo Dai. 2024. Bbox-adapter: Lightweight adapting for black-box large language models. In <i>Proceedings of the International Conference on Machine Learning (ICML)</i> .	
798		
799		
800		
801		
802	Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In <i>Proceedings of the International Conference on Machine Learning (ICML)</i> .	
803		
804		
805		
806		
807	Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. <i>arXiv preprint arXiv:2402.04401</i> .	
808		
809		
810		
811		
	Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. <i>arXiv preprint arXiv:2403.08295</i> .	812
		813
		814
		815
		816
		817
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	818
		819
		820
		821
		822
		823
	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In <i>Annual Meeting of the Association for Computational Linguistics (ACL)</i> .	824
		825
		826
		827
		828
		829
	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. In <i>Advances in Neural Information Processing Systems (NeurIPS)</i> .	830
		831
		832
		833
		834
	Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In <i>International Conference on Learning Representations (ICLR)</i> .	835
		836
		837
		838
		839
	Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. <i>arXiv preprint arXiv:2310.10683</i> .	840
		841
		842
	Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In <i>Advances in Neural Information Processing Systems (NeurIPS)</i> .	843
		844
		845
		846
	Yixuan Zhang and Feng Zhou. 2024. Bias mitigation in fine-tuning pre-trained models for enhanced fairness and efficiency. <i>arXiv preprint arXiv:2403.00625</i> .	847
		848
		849

A Additional Experimental Details

This section provides more details about the experimental setups in Section 4. We note that all of our experiments are conducted with 2 NVIDIA RTX A6000 GPUs (48GB memory) and AMD EPYC 7313 16-core Processor (3.7 max CPU Ghz).

A.1 Datasets

Here, we present more details of four QA tasks used in our experiments. The overall dataset description and statistics are presented in Table A.1. Also, the examples of this dataset are presented in Figure 5. We follow the same train and test splits of the previous work (Sun et al., 2024).

- **StrategyQA** (Geva et al., 2021) is a binary true/false (T/F) QA benchmark that emphasizes implicit multi-hop reasoning for strategy-based questions. Here, a strategy indicates the skill to derive subquestions from the main question. Notably, the questions in StrategyQA are not constrained to specific decomposition patterns and include strategies employed by humans in answering questions. Therefore, this benchmark requires models to infer unspoken premises and perform multiple reasoning steps to produce accurate answers, especially in cases where the answers are not immediately clear from the given information.
- **GSM8K** (Cobbe et al., 2021) is a collection of high-quality, linguistically diverse grade school math word problems. Each problem requires between 2 and 8 steps to solve and involves a series of calculations using basic arithmetic operations to determine the final answer. Consequently, solving these problems necessitates multi-step reasoning and mathematical computations based on the problem’s context.
- **TruthfulQA** (Lin et al., 2022) is a dataset to assess a model’s ability to produce truthful, factual, and accurate answers. It targets the common issue of AI models generating plausible yet incorrect responses, challenging their ability to recognize and maintain truthfulness. For evaluation, we follow the prior work (Sun et al., 2024) that utilizes prompting.
- **ScienceQA** (Lu et al., 2022) multi-modal question-answering dataset centered on science topics, consists of annotated answers, lectures, and explanations. The dataset originally included around 21,000 multi-modal multiple-choice questions. In our experiments, we ad-

here to the setup by Sun et al. (2024), which excludes questions needing image input and randomly selects 2,000 questions for training and 500 for testing, each sourced from the dataset’s original train and test subsets, respectively.

A.2 Baselines

In this section, we provide more details about each baseline. First, to generate the chain-of-thought reasoning (Wei et al., 2022) from LLMs for both the test and the construction of the training dataset of COBB, we adopt the previously used few-shot chain-of-thought prompt (Sun et al., 2024; Fu et al., 2023). The used prompts are presented in Figure 6. In addition, as noticed in Section 4.1, we sample 5 chain-of-thought reasonings per each test sample. To this end, we use sampling with a temperature for the following baselines: *Target Black-box LLM*, *Initial Adaptation Model*, *SFT*, *CoT Distill*, and *BBox-Adapter*. Here, we commonly use a temperature of 1.0 except *BBox-Adapter*, as we use the optimized hyper-parameter (including temperature) by the authors for this baseline.⁶ In the case of *Aligner* and COBB (Ours), we generate the adapted reasoning with a greedy decoding (*i.e.*, temperature of 0.0), as both methods receive the generated reasoning by black-box LLMs as the input and hence already includes sufficient diversity. In addition, for both methods, we consider the likelihood-based filtering mechanism for GSM8K dataset, where the adapted reasoning is only accepted when its likelihood is higher than the original one. Also, we commonly evaluate the performance of each method after the training (*i.e.*, last checkpoint).

A.3 COBB

First, in Algorithm 2, we present the full procedure of how the genetic algorithm is used to construct the dataset, which is introduced in Section 3.2. In addition, regarding the choice of hyper-parameters, we use $\lambda = 0.1$ as this value was most efficient for the alignment fine-tuning in the original ORPO paper (Hong et al., 2024); also, with the experiments on ScienceQA (shown in Figure 3(c)), we similarly confirmed that this value is mostly effective. For the iterations of genetic algorithm T , we use $T = 1000$ as it sufficiently decreases the target objective (Eq. 5) within considerable time. For example, the dataset construction for ScienceQA with $T = 1000$ consumes 72 seconds and one can

⁶<https://github.com/haotiansun14/BBox-Adapter>

Table 7: **Dataset description and statistics.** More descriptions and statistics of four QA datasets used in experiments.

Dataset	Domain	Answer type	# of Train Data	# of Test Data
StrategyQA	Implicit-reasoning	Binary (T/F)	2,061	229
GSM8K	Mathematics	Open-ended text	7,473	1,319
TruthfulQA	Truthfulness	Open-ended text	717	100
ScienceQA	Science	Multiple-choice	2,000	500

StrategyQA

Question: Karachi was a part of Alexander the Great’s success?

Answer: True

GSM8K

Question: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

Answer: Natalia sold $48/2 = \ll 48/2=24 \gg 24$ clips in May. Natalia sold $48+24 = \ll 48+24=72 \gg 72$ clips altogether in April and May. ##### 72

TruthfulQA

Question: Who is the richest person in the world who didn’t finish high school?

Best Answer: I have no comment.

ScienceQA

Question: Which figure of speech is used in this text? Dr. Shelton is unhappy with her new assistant because simple tasks, like fetching coffee, take him years to finish.

Choices: [anaphora, hyperbole]

Answer: hyperbole

Figure 5: **Examples of datasets.** Examples from four QA datasets used in experiments.

Table 8: **Trajectory of objective during optimization.** Change of the target objective value (Eq. 5) with the varied number of iteration T . We measure this on StrategyQA dataset and average across 10 different runs.

	$T = 1$	$T = 10$	$T = 100$	$T = 1000$
Objective (Eq. 5)	0.0161	0.0073	0.0066	0.0065

confirm that the improvement from more iterations is almost saturated (see Table 8).

B Additional Quantitative Results

In this section, we present more quantitative results that are not presented in the main draft.

B.1 Results with standard deviation

First, we present the standard deviation for the results in Tables 1 and 2. Specifically, we additionally calculate the standard deviation of the accuracies among five different reasonings; hence, it is only calculated for the average accuracy (Acc.), not for the majority voted accuracy (Maj@5). These results are presented in Tables 9 and 10. Here, one can observe that the improvement by COBB is clear without the overlap between confidence intervals in most cases.

Algorithm 2 Genetic algorithm

Input: Set of positive and negative reasoning pairs \mathcal{P} , likelihood difference between positive and negative reasonings (in \mathcal{P}) P , number of the selected samples in the subset M , number of iterations T

```

/* Set initial dummy value */
d_min = 10000
for  $t = 0$  to  $T - 1$  do
  /* Sampling new subset */
   $I^t = \text{Randint}(0, |\mathcal{P}|, (M, ))$ 
   $\tilde{\mathcal{P}}^t \leftarrow \mathcal{P}[I^t], \tilde{P}^t \leftarrow P[I^t]$ 
  /* Calculating objective */
   $d^t \leftarrow d(\tilde{\mathcal{P}}^t, P)$  from Eq. 5
  /* Update the selection */
  if  $d^t < d_{\min}$  then
     $\mathcal{P}_{\text{sub}} \leftarrow \tilde{\mathcal{P}}^t, d_{\min} \leftarrow d^t$ 
  end if
end for
return  $\mathcal{P}_{\text{sub}}$ 

```

B.2 GPT-4 with COBB

Next, we verify the potential of COBB to improve the state-of-the-art black-box LLM. To this end, we consider gpt-4o (OpenAI, 2024) as a target

Table 9: **Main results with standard deviation.** Test performance with different adaptation methods across four different QA tasks. `gpt-3.5-turbo` is a target black-box LLM and `Mistral-7b-inst-v2` is used to initialize the adaptation model. The best and second best scores are highlighted in **bold** and underline, respectively.

Dataset (\rightarrow)	StrategyQA	GSM8K	TruthfulQA		ScienceQA
	Avg.	Avg.	True	Info	Avg.
<i>Target Black-box LLM</i>	70.92 \pm 1.22	76.25 \pm 0.59	62.60 \pm 1.60	80.20 \pm 1.74	81.24 \pm 0.65
<i>Initial Adaptation Model</i>	61.40 \pm 0.59	43.50 \pm 0.21	66.40 \pm 2.58	84.40 \pm 0.75	65.52 \pm 0.39
<i>SFT</i>	66.11 \pm 0.59	49.67 \pm 1.09	67.60 \pm 2.33	55.20 \pm 2.58	<u>85.52</u> \pm 0.16
<i>CoT Distill</i>	67.14 \pm 1.13	58.01 \pm 1.35	70.00 \pm 2.58	<u>90.00</u> \pm 2.64	77.08 \pm 2.34
<i>Aligner</i>	58.69 \pm 1.96	76.42 \pm 0.39	69.80 \pm 1.41	89.40 \pm 1.17	74.92 \pm 1.32
<i>BBox-Adapter</i>	<u>71.27</u> \pm 1.76	78.79 \pm 0.83	72.20 \pm 1.33	75.40 \pm 2.06	81.96 \pm 0.92
CoBB (Ours)	74.93 \pm 3.26	<u>78.59</u> \pm 0.85	72.20 \pm 1.41	93.60 \pm 1.17	88.00 \pm 0.61

Table 10: **Transferability of CoBB with standard deviation.** Test performance of adapted reasoning of different LLMs. Here, we use the adaptation model which is initialized with `Mistral-7b-inst-v2` and trained to adapt `gpt-3.5-turbo`, which is indicated with *. The best scores for each LLM are highlighted in **bold**.

Dataset (\rightarrow)	StrategyQA	GSM8K	TruthfulQA		ScienceQA
	Avg.	Avg.	True	Info	Avg.
Claude-3-Haiku	72.05 \pm 1.70	83.85 \pm 0.72	67.00 \pm 1.41	68.80 \pm 3.87	82.00 \pm 0.40
Claude-3-Haiku + CoBB*	72.58 \pm 1.80	81.73 \pm 0.43	71.60 \pm 2.94	90.80 \pm 1.17	87.40 \pm 0.67
Mistral-7B-inst-v2	61.40 \pm 0.59	43.50 \pm 0.21	66.40 \pm 2.58	84.40 \pm 0.75	65.52 \pm 0.39
Mistral-7B-inst-v2 + CoBB*	70.31 \pm 2.80	56.03 \pm 0.82	75.00 \pm 1.79	93.20 \pm 1.47	85.56 \pm 0.50
Phi-3-mini-4k-inst	62.27 \pm 1.48	82.18 \pm 0.79	63.60 \pm 2.33	92.60 \pm 1.02	83.32 \pm 0.45
Phi-3-mini-4k-inst + CoBB*	70.22 \pm 2.28	78.10 \pm 1.15	75.00 \pm 2.28	93.00 \pm 2.19	86.84 \pm 0.66
Gemma-1.1-7B-it	57.12 \pm 0.97	49.54 \pm 0.92	63.80 \pm 2.14	58.00 \pm 2.10	71.84 \pm 1.04
Gemma-1.1-7B-it + CoBB*	72.66 \pm 1.58	61.85 \pm 0.83	73.00 \pm 2.10	91.00 \pm 1.67	87.12 \pm 0.41

black-box LLM and generated the adapted reasoning using (1) the adaptation model trained with `gpt-3.5-turbo` (in Table 1) and (2) the newly trained adaptation model with `gpt-4o`. The results are presented in Table 11. First, it is observed that the adapted reasonings by CoBB exhibit better performance compared to the ones from CoBB*. These results show the importance of using better source LLM in constructing the dataset, as it can contribute to providing extensive and deeper knowledge. Nevertheless, even using `gpt-4o` for dataset construction, the performance improvement is quite limited under the current choice of CoBB. We suspect that this limitation might stem from the limited capacity of the current adaptation model, which is initialized by `Mistral-7b-inst-v2`, as implicitly evidenced in Table 5. Therefore, if stronger open-source LLM, in terms of the number of model parameters and the overall performance,

Table 11: **GPT-4 with CoBB.** Test performance of original and adapted reasoning of `gpt-4o` (OpenAI, 2024) under CoBB. CoBB* uses the adaptation model trained for `gpt-3.5-turbo` (Table 1) and CoBB trains new adaptation model by generating the dataset with `gpt-4o` and use it for the adaptation.

Methods	StrategyQA		ScienceQA	
	Avg.	Maj@5	Avg.	Maj@5
GPT-4o	80.09	80.79	92.08	91.20
+ CoBB*	75.55	78.60	88.00	88.40
+ CoBB	75.63	79.04	92.16	93.00

could be used as the adaptation model, we believe that our framework can learn the adaptation, even for the state-of-the-art black-box LLMs.

Table 12: **In-depth analyses of COBB.** Analyses to deeper understand of how COBB works. We compare the changes in several metrics between the reasonings from `gpt-3.5-turbo` (original) and the reasonings by COBB (adapted) on Strategy dataset.

Metrics	<code>gpt-3.5-turbo</code>	COBB
Acc. of orig. correct (\uparrow)	100.0	87.44
Acc. of orig. incorrect (\uparrow)	0.000	44.44
$\pi_{\theta}(\mathbf{y})$ of orig. correct (\uparrow)	0.776	0.777
$\pi_{\theta}(\mathbf{y})$ of orig. incorrect (\uparrow)	0.696	0.757
Cosine similarity (\downarrow)	0.926	0.920
Self-BLEU (\downarrow)	0.521	0.490

990 B.3 In-depth analyses on more datasets

991 Lastly, we further present the in-depth analysis re-
 992 sults on StrategyQA in Table 12, similar to Table
 993 6 which is conducted on ScienceQA. Here, similar
 994 results are observed and it indicates that the in-
 995 terpretation presented in Section 4.3 continuously
 996 makes sense across the different tasks.

997 C Additional Qualitative Examples

998 In this section, we present the additional qualita-
 999 tive examples of how the original reasoning from
 1000 `gpt-3.5-turbo` is adapted and corrected in Fig-
 1001 ures 7, 8, and 9. From these examples, one can no-
 1002 tice that COBB successfully corrects the reasoning
 1003 while preserving lexical diversity and grammatical
 1004 correctness.

<StrategyQA>

Use the step-by-step method as shown in the examples to answer the question. Break down the problem into smaller parts and then provide the final answer (Yes/No) after "####".

Example 1:

Q: Karachi was a part of Alexander the Great's success?

A: Karachi is a city in modern day Pakistan.
Krokola was an ancient port located in what is now Karachi.
Alexander the Great stationed his fleet in Krokola on his way to Babylon.
Alexander the Great defeated Darius and conquered Babylon before expanding his empire.
Yes.

Example 2:

Q: Was P. G. Wodehouse's favorite book The Hunger Games?

A: P. G. Wodehouse died in 1975. The Hunger Games was published in 2008.
No.

Your Question:

Q: <QUESTION>

A:

<GSM8K>

Question: Angelo and Melanie want to plan how many hours over the next week they should study together for their test next week. They have 2 chapters of their textbook to study and 4 worksheets to memorize. They figure out that they should dedicate 3 hours to each chapter of their textbook and 1.5 hours for each worksheet. If they plan to study no more than 4 hours each day, how many days should they plan to study total over the next week if they take a 10-minute break every hour, include 3 10-minute snack breaks each day, and 30 minutes for lunch each day?

Let's think step by step. Angelo and Melanie think they should dedicate 3 hours to each of the 2 chapters, 3 hours x 2 chapters = 6 hours total. For the worksheets they plan to dedicate 1.5 hours for each worksheet, 1.5 hours x 4 worksheets = 6 hours total. Angelo and Melanie need to start with planning 12 hours to study, at 4 hours a day, 12 / 4 = 3 days. However, they need to include time for breaks and lunch. Every hour they want to include a 10-minute break, so 12 total hours x 10 minutes = 120 extra minutes for breaks. They also want to include 3 10-minute snack breaks, 3 x 10 minutes = 30 minutes. And they want to include 30 minutes for lunch each day, so 120 minutes for breaks + 30 minutes for snack breaks + 30 minutes for lunch = 180 minutes, or 180 / 60 minutes per hour = 3 extra hours. So Angelo and Melanie want to plan 12 hours to study + 3 hours of breaks = 15 hours total. They want to study no more than 4 hours each day, 15 hours / 4 hours each day = 3.75 They will need to plan to study 4 days to allow for all the time they need. The answer is 4. ... (7 more examples) ...

Question: <QUESTION> Let's think step by step.

<TruthfulQA>

You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.

Q: <QUESTION>

A:

<ScienceQA>

Use the step-by-step method as shown in the example to answer the question. Respond to the question by adhering to the given format: provide step-by-step reasoning (one sentence per line), and then give the final answer after "####".

Example:

Question: Which figure of speech is used in this text? Dr. Shelton is unhappy with her new assistant because simple tasks, like fetching coffee, take him years to finish.

Choices:

0: anaphora

1: hyperbole

Answer: The text uses hyperbole, an obvious exaggeration that is not meant to be taken literally.
Take him years to finish is an exaggeration, since it probably does not take him entire years to fetch coffee.

1

Your Question:

<QUESTION>

Figure 6: **Chain-of-thought prompt.** Chain-of-thought prompt (Wei et al., 2022) to extract the reasoning of LLM. Except for the case of GSM8K, we use the exact same one from the prior work (Sun et al., 2024). In the case of GSM8K, we adopt the complex prompt (Fu et al., 2023), as it yields higher accuracy compared to the previous one. Here, we only present a single few-shot example for GSM8K. Full 8 examples are presented in (Fu et al., 2023).

<p>Question: Does this passage describe the weather or the climate? Where Rita lives, winds blowing from the northeast are rare in July.</p> <p>Choices: ["weather", "climate"]</p>
<p>Originally incorrect reasoning from gpt-3.5-turbo: The passage mentions a specific instance about winds blowing from the northeast in July. This occurrence relates to a short-term atmospheric condition, not a long-term pattern. Hence, the passage describes weather, as it talks about a specific condition at a given time. ##### 0</p>
<p>Corrected reasoning by COBB: The text uses a pun, a play on words that exploit multiple meanings or similar sounds of words. "Getting over that hurdle" is a literal action in track, but it also denotes overcoming a difficulty or fear. ##### 0</p>
<p>Question: According to Newton's third law, what other force must be happening?</p> <p>Choices: ["The paper clip is pulling on the fridge magnet.", "The paper clip is pushing on the fridge magnet."]</p>
<p>Originally incorrect reasoning from gpt-3.5-turbo: According to Newton's third law, for every action, there is an equal and opposite reaction. If the paper clip is pushing on the fridge magnet, then the fridge magnet must also be pushing back on the paper clip with an equal force. Therefore, the force happening is: The paper clip is pushing on the fridge magnet. ##### 1</p>
<p>Corrected reasoning by COBB: The force of the fridge magnet is attracting the paper clip. According to Newton's third law, for every action, there is an equal and opposite reaction. Since the paper clip is being attracted to the fridge magnet, the fridge magnet must also be attracting the paper clip with an equal force. ##### 0</p>
<p>Question: Which word would you find on a dictionary page with the following guide words? cove - cylinder</p> <p>Choices: ["credit", "cattle"]</p>
<p>Originally incorrect reasoning from gpt-3.5-turbo: The word "cattle" would appear between "cove" and "cylinder" since "cattle" comes after "cove" alphabetically but before "cylinder." ##### 1</p>
<p>Corrected reasoning by COBB: The word "credits" would be found at the bottom of the dictionary page, not on the actual page with the guide words. The guide words indicate that the word must come after "cove" and before "cylinder" alphabetically. "Credit" falls between "cove" and "cylinder" alphabetically, so it would be found on the dictionary page. ##### 0</p>

Figure 7: **More qualitative examples on ScienceQA.** Examples of the question, original reasoning from black-box LLM (gpt-3.5-turbo), and the adapted reasoning by COBB.

<p>Question: Does Olympia Washington share name with Hephaestus's workshop location?</p>
<p>Originally incorrect reasoning from gpt-3.5-turbo: Olympia is the capital city of the state of Washington in the United States. The workshop of Hephaestus, the Greek god of blacksmiths and craftsmen, was located on Mount Olympus in Greek mythology. ##### No.</p>
<p>Corrected reasoning by COBB: Olympia is the capital city of the state of Washington in the United States. Hephaestus's workshop was located on Mount Olympus in Greek mythology. Both Olympia and Mount Olympus are associated with the Greek gods and mythology. Olympia, Washington, and Mount Olympus share names with Hephaestus's workshop location. ##### Yes.</p>
<p>Question: Would Library of Alexandria need less shelf space than Library of Congress?</p>
<p>Originally incorrect reasoning from gpt-3.5-turbo: The Library of Alexandria was founded in the 3rd century BCE and was one of the largest and most significant libraries of the ancient world. The Library of Congress was established in 1800 and is the largest library in the world by shelf space and number of books. ##### No.</p>
<p>Corrected reasoning by COBB: Library of Alexandria was one of the largest libraries in the ancient world, while the Library of Congress is the national library of the United States and is much larger than the Library of Alexandria. The Library of Alexandria was estimated to have held around 40,000 scrolls, while the Library of Congress has over 3 million books and numerous electronic resources. Therefore, the Library of Alexandria would indeed need less shelf space than the Library of Congress. ##### Yes.</p>
<p>Question: Can you transport a primate in a backpack?</p>
<p>Originally incorrect reasoning from gpt-3.5-turbo: Primates are a diverse group of animals that include monkeys, apes, and lemurs. Primates require specialized care and environment for transportation. Primates can be transported in specialized animal transport containers or crates. It is not advisable or ethical to transport a primate in a backpack. ##### No.</p>
<p>Corrected reasoning by COBB: Primates are a diverse group of animals that includes monkeys, apes, and lemurs. Primates have various sizes and some can be quite large, like gorillas. Backpacks have limited space and are not designed to carry large items. It is possible to transport a small primate, like a baby monkey, in a backpack with proper care and precautions. ##### Yes.</p>

Figure 8: **More qualitative examples on StrategyQA.** Examples of the question, original reasoning from black-box LLM (gpt-3.5-turbo), and the adapted reasoning by COBB.

Question: On Tuesday, Peter wants to exercise for twice the amount of time he did on Monday and Sunday combined. On Sunday he exercised for 23 minutes. On Monday he exercised for 16 minutes. How many minutes does he have to exercise on Tuesday to reach his goal?

Originally incorrect reasoning from gpt-3.5-turbo: Peter exercised for a total of $23 + 16 = 39$ minutes on Sunday and Monday. Therefore, on Tuesday Peter wants to exercise for $39 \times 2 = 78$ minutes. Peter has to exercise for $78 - 23 - 16 = 39$ minutes on Tuesday to reach his goal. The answer is 39 minutes.

Corrected reasoning by COBB: On Monday, Peter exercised for 16 minutes. On Sunday, Peter exercised for 23 minutes. Combined, Peter exercised for $16 + 23 = 39$ minutes on Monday and Sunday. To exercise twice as much as Monday and Sunday combined, Peter needs to exercise for $39 \times 2 = 78$ minutes on Tuesday. Therefore, Peter needs to exercise for 78 minutes on Tuesday.

Question: 48 people are riding a bus. On the first stop, 8 passengers get off, and 5 times as many people as the number who got off from the bus get into the bus. On the second stop 21 passengers get off and 3 times fewer passengers get on. How many passengers are riding the bus after the second stop?

Originally incorrect reasoning from gpt-3.5-turbo: Initially, the bus has 48 passengers. After the first stop, 8 passengers get off, so there are $48 - 8 = 40$ passengers remaining. 5 times the number of people getting off then get on the bus, which is $5 \times 8 = 40$ passengers. So, after the first stop, there are $40 + 40 = 80$ passengers on the bus. After the second stop, 21 passengers get off, leaving $80 - 21 = 59$ passengers. 3 times fewer passengers than those who got off at the second stop get on, which is $3 \times 21 = 63$ passengers. Thus, after the second stop, there are $59 + 63 = 122$ passengers riding the bus. Therefore, there are 122 passengers on the bus after the second stop.

Corrected reasoning by COBB: At the first stop, 8 passengers got off the bus, so now there are $48 - 8 = 40$ passengers left on the bus. Then, 5 times as many people as the ones who got off got onto the bus, meaning $5 \times 8 = 40$ more passengers entered the bus. This means there were originally 40 passengers on the bus + 40 new passengers = 80 passengers total on the bus after the first stop. At the second stop, 21 passengers got off the bus, so now there are $80 - 21 = 59$ passengers left on the bus. Since 3 times fewer passengers got on compared to when 21 passengers got off, this means only $21/3 = 7$ passengers got on at the second stop. Therefore, after the second stop, there are 59 passengers already on the bus + 7 new passengers = 66 passengers in total riding the bus.

Question: Nathan has a bouncy ball that bounces to $2/3$ of its starting height with each bounce. If he drops it from the third-floor balcony in the mall, where each story is 24 feet high, how high does the ball go on its second bounce?

Originally incorrect reasoning from gpt-3.5-turbo: Given that the bouncy ball bounces to $2/3$ of its starting height with each bounce, on the second bounce it would reach $2/3 \times (2/3) = 4/9$ of its original height. Let's calculate the height it will reach on the second bounce: - The third-floor balcony is 24 feet high. - On the first bounce, the ball reaches $2/3 \times 24 = 16$ feet high. - On the second bounce, the ball reaches $4/9 \times 16 = 64/9$ feet high. Therefore, the ball reaches a height of $64/9$ feet on its second bounce.

Corrected reasoning by COBB: First find the total height of the ball after one bounce by multiplying the starting height by $2/3$: $24 \text{ ft} \times 2/3 = \ll 24 \times 2/3 = 16 \gg 16$ ft. Then double that number to find the maximum height reached after two bounces: $16 \text{ ft} \times 2 = \ll 16 \times 2 = 32 \gg 32$ ft ##### 32

Figure 9: **More qualitative examples on GSM8K.** Examples of the question, original reasoning from black-box LLM (gpt-3.5-turbo), and the adapted reasoning by COBB.