

# ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback

Ju-Seung Byun\*, Jiyun Chun\*, Jihyung Kil, Andrew Perrault

Department of Computer Science and Engineering

The Ohio State University

{byun.83, chun.203, kil.15, perrault.17}@osu.edu

## Abstract

Large Multimodal Models (LMMs) excel at comprehending human instructions and demonstrate remarkable results across a broad spectrum of tasks. Reinforcement Learning from Human Feedback (RLHF) and AI Feedback (RLAIF) further refine LLMs by aligning them with specific preferences. These methods primarily use ranking-based feedback for entire generations. With advanced AI models (Teacher), such as GPT-4 and Claude 3 Opus, we can request various types of detailed feedback that are expensive for humans to provide. We propose a two-stage algorithm **ARES** that Alternates REinforcement Learning (RL) and SUpervised Fine-Tuning (SFT). First, we ask the Teacher to score how much each sentence contributes to solving the problem in a Chain-of-Thought (CoT). This sentence-level feedback allows us to consider individual valuable segments, providing more granular rewards for the RL procedure. Second, we ask the Teacher to correct wrong reasoning after the RL stage. The RL procedure requires substantial hyperparameter tuning and often generates errors such as repetitive words and incomplete sentences. With correction feedback, we stabilize the RL fine-tuned model through SFT. We conduct experiments on the multi-modal datasets ScienceQA and A-OKVQA to demonstrate the effectiveness of our proposal. The ARES rationale achieves around 70% win rate compared to baseline models judged by GPT-4o. Additionally, we observe that the improved rationale reasoning leads to a 2.5% increase in inference answer accuracy on average for the multi-modal datasets. <sup>1</sup>

## 1 Introduction

Large Language Models (LLMs) and Large Multimodal Models demonstrate remarkable performance across diverse language and multi-modal

\*Equal contribution

<sup>1</sup>Code: <https://github.com/Amyyyyeah/ARES>

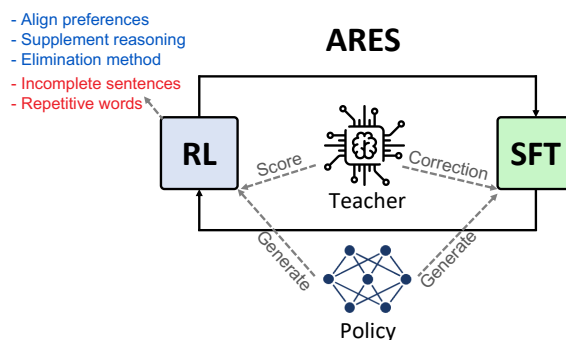


Figure 1: **Overview of ARES:** ARES alternates RL and SFT. The Teacher provides scores (rewards) for each sentence for RL. **Blue** indicates the advantages of RL, and **red** indicates potential degeneration. ARES corrects the issues through the Teacher’s correction feedback.

tasks (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; Zhang et al., 2022a; Liu et al., 2023a; Goyal et al., 2023). However, these Large Models (LMs) often generate toxic and biased content (Gehman et al., 2020; Tamkin et al., 2021) because LMs are primarily trained to predict the next token based on extensive corpus datasets. To align LM behavior more closely with user preferences, previous works (Glaese et al., 2022; Ouyang et al., 2022) fine-tune their models using Reinforcement Learning from Human Feedback (RLHF). Furthermore, with advancements in LMs, advanced LM feedback can replace costly human feedback, yielding Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023; Bai et al., 2022; Yuan et al., 2024).

However, RLHF and RLAIF encounter two significant challenges. First, both methods often utilize ranking-based feedback (Ouyang et al., 2022), which orders the generated samples by preference. For instance, if sample *A* is preferred over sample *B*, the model is fine-tuned to generate more outputs like *A* and fewer like *B*. However, if *B* contains certain valuable parts, they may be discarded. To alleviate these issues, Lightman et al. (2023);

Luo et al. (2024) propose sentence-level feedback, applying it solely to the reward model without Reinforcement Learning (RL), called the Process-supervised Reward Model (PRM). It demonstrates its potential through search algorithms on the PRM, such as best-of- $N$  or Monte Carlo Tree Search (MCTS). Furthermore, Wang et al. (2024) demonstrate the effectiveness of RL with sentence-level feedback by heuristically scoring each sentence in math problems, where evaluating the predicted answer is straightforward. Thus, sentence-level feedback exhibits significant promise compared to existing ranking-based feedback. Nonetheless, acquiring sentence-level feedback is more costly than ranking-based feedback.

Second, the RL process is inherently unstable and requires extensive hyperparameter tuning (Eimer et al., 2023). This instability often results in the generation of repetitive words and truncated sentences. Hyperparameter tuning becomes an enormous burden as the model size increases, making exhaustive tuning seemingly impossible, especially for individuals. The existing RLHF method (Ouyang et al., 2022) recycles the dataset used in pretraining within the loss function with Proximal Policy Optimization (PPO) (Schulman et al., 2017) to mitigate this degeneration problem. However, this approach prevents the model from fully maximizing the sum of rewards through RL and may limit the opportunity for diverse improvements, which differ from the pretraining distribution.

In this work, we aim to address the two challenges mentioned above through various types of feedback using an advanced AI model as a Teacher. Many advanced AI models, including GPT-4 and Claude 3 Opus, are already used as evaluators for many tasks and generate reliable human-level answers (Liu et al., 2023b; Sottana et al., 2023). 1) We request a score from the Teacher for each sentence, ranging from 0.0 to 1.0. Each score indicates how much a sentence contributes to solving the problem. This provides detailed reward feedback to the training model and can be applied to both mathematical and more general multi-modal Chain-of-Thought (CoT) (or rationale reasoning) problems. 2) We ask the Teacher to identify and correct minor errors in the RL results, such as incorrect or cut-off parts. With this corrected dataset, we fine-tune the model using Supervised Fine-Tuning (SFT). This stage allows the model to maximize the rewards while properly deviating from the pretraining distribution. In summary, we propose a hybrid

algorithm **ARES** that Alternates **RE**inforcement Learning and **S**upervised Fine-Tuning.

To evaluate how much rationale reasoning can be improved through the ARES framework, we use the ScienceQA and A-OKVQA datasets, which are large-scale, multi-modal datasets that include rationale reasoning data. We use Multimodal-CoT (MM-CoT) (Zhang et al., 2023b) as the baseline. MM-CoT employs two separate models: one model is responsible for generating rationale, and the other model, an inference model, processes the concatenated input (problem and generated rationale). This distinct framework enhances performance, even with relatively smaller models like Flan-Alpaca<sub>Base</sub> (Chia et al., 2023) (251M) and Flan-Alpaca<sub>Large</sub> (790M) with ViT feature (Dosovitskiy et al., 2021). We perform ARES on the rationale reasoning model of MM-CoT. We compare ARES rationale reasoning with that of MM-CoT through GPT-4o, determining which rationale is better and computing the win rate. Additionally, we check whether the improved rationale reasoning leads to better answer accuracy. Our results show that our rationale reasoning outperforms the baselines with around 70% win rate and demonstrates 2.5% increase in inference answer accuracy on average for the different model sizes and the multi-modal tasks.

In summary, our key contributions are:

- We propose ARES, leveraging diverse types of feedback from advanced AI (Teacher) to enhance CoT reasoning for multi-modal tasks.
- ARES properly reflects the direction in which the model is fine-tuned through RL and stabilizes the RL fine-tuning procedure with SFT.
- ARES generates better rationale chains compared to baselines judged by GPT-4o and improves inference answer accuracy for multi-modal tasks.

## 2 Methodology

This section briefly introduces the preliminaries in Section 2.1 and presents our two-stage hybrid algorithm **ARES** that Alternates **RE**inforcement Learning and **S**upervised Fine-Tuning. 1) We request a score for each sentence in the Chain-of-Thought (CoT) from the advanced AI model (Teacher) to determine how much it contributes to solving the problem (Section 2.2). We perform Reinforcement

Learning (RL) with the score feedback on our training model. 2) The Teacher corrects minor errors such as truncated or slightly incorrect sentences, thereby performing Supervised Fine-Tuning (SFT) (Section 2.2).

## 2.1 Preliminaries

For an input  $x \in X$ , a Transformer-based (Vaswani et al., 2023) model  $\pi_\theta(\cdot|x)$  parameterized by  $\theta$  generates the output  $y$  composed of sentences  $\{s_0, s_1, \dots, s_k\}$ .

$$\pi_\theta(y | x) = \prod_{t=0}^k \pi_\theta(y_t | x, y_{<t}), \quad (1)$$

where  $y_{<t}$  indicates previous tokens. To proceed with RL finetuning, Ouyang et al. (2022) train an outcome-supervised reward model (ORM) using ranking-based feedback. With more fine-grained feedback like sentence-level, Lightman et al. (2023); Wang et al. (2024) train a process-supervised reward model (PRM). Instead of training a reward model, we request score feedback  $r(x \cup s_{<t}, s_t)$  for each sentence from an advanced AI such as GPT-4 where  $s_{<t}$  represents previous sentences.

## 2.2 Reinforcement Learning

Reinforcement Learning (RL) fine-tunes our model  $\pi_\theta$  to maximize sum of sentence rewards from an advanced AI model such as GPT-4 and Claude 3 Opus. The RL objective is as follows:

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\substack{x \sim X \\ s_t \sim \pi_\theta(\cdot|x, s_{<t})}} \left[ \sum_{i=0}^k \gamma^i r(x \cup s_{<t}, s_t) \right]. \quad (2)$$

where  $\gamma \leq 1.0$  is a discount factor. We use Proximal Policy Optimization (PPO) (Schulman et al., 2017) to achieve this RL objective, treating sentences as actions (Equation 3).

$$\begin{aligned} & \operatorname{maximize}_{\theta} \mathbb{E}_t \left[ \frac{\pi_\theta(s_t|x \cup s_{<t})}{\pi_{\theta_{\text{old}}}(s_t|x \cup s_{<t})} \hat{A}_t \right] \\ & \text{s.t. } \mathbb{E}_t \left[ \text{KL}(\pi_\theta(\cdot|x \cup s_{<t}), \pi_{\theta_{\text{old}}}(\cdot|x \cup s_{<t})) \right] \leq \delta \end{aligned} \quad (3)$$

where  $\pi_{\text{old}}$  is the original policy (baseline model) and  $\hat{A}_t$  is an advantage estimator at timestep  $t$ . PPO is commonly leveraged in Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and AI Feedback (RLAIF) (Bai et al., 2022). PPO’s conservative update prevents the

training model from deviating too far from the original model, thus avoiding degeneration.

**Sentence-Level Nuanced Feedback:** We request a score between 0.0 and 1.0 for each sentence in CoT through the advanced AI for RL. The closer the score is to 1.0, the more relevant and helpful it is to solving the problem. Table 5 presents the prompt format. We additionally shift the reward distribution by  $-0.5$  to center it at 0 (Zheng et al., 2023). Therefore, the actual range is from  $-0.5$  to 0.5. Using these nuanced scores, the RL fine-tuned model exhibits emergent behaviors (please refer to Section 4). This allows us to understand the direction in which the model is intended to change through RL.

## Advantages of Using Advanced AI for Score Feedback:

Although calling the API has disadvantages, such as incurring costs or facing usage limits, there exist several advantages to using the advanced AI for feedback. First, there is no need to train a reward model. Second, as the RL fine-tuned model begins to generate out-of-distribution outputs that differ from the data used to train the reward model, it becomes challenging for the trained reward model to provide accurate rewards. However, this out-of-distribution problem is effectively addressed with the advanced AI.

**RL Challenge:** One of the challenging factors for RL is hyperparameter tuning (Eimer et al., 2023). This often results in generating repetitive words and truncated sentences (Ouyang et al., 2022). Additionally, as the model size increases, finding working hyperparameters becomes infeasible for individuals. To alleviate this issue, we utilize correction feedback from the advanced AI as the second stage (Section 2.3), and proceed with the supervised fine-tuning to stabilize the RL fine-tuned model.

## 2.3 Correction: Supervised Fine-Tuning

The RL fine-tuning procedure makes model changes to maximize the reward sum, such as correcting mistakes or explaining why other options cannot be the answer. However, without highly tuned hyperparameters (Eimer et al., 2023), the model after the RL phase may result in errors such as repeated sentences, truncated sentences, or incorrect content for some data points. (See examples in Appendix D.)

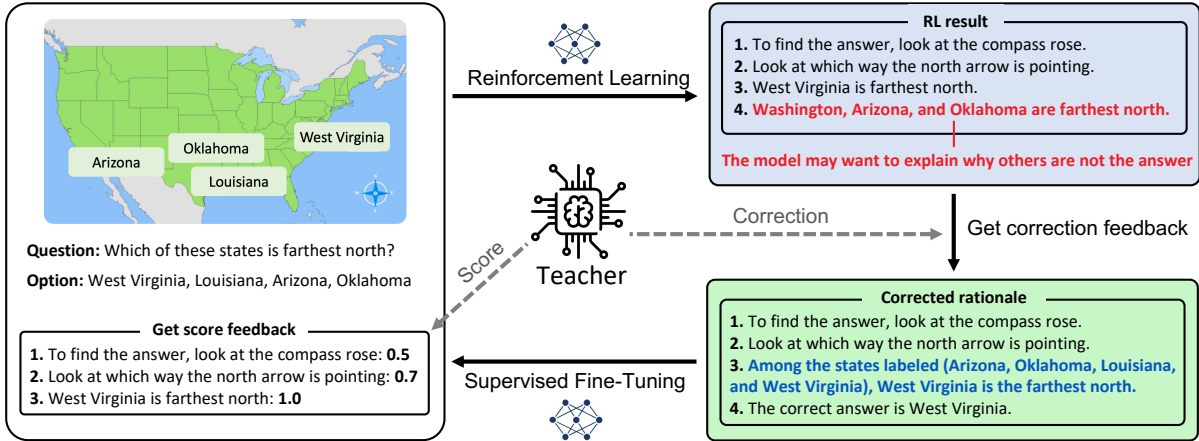


Figure 2: ARES pipeline: For a given model generating rationale reasoning, we request an AI model’s sentence-level scores (ranging from 0.0 to 1.0). The closer the score is to 1.0, the more it helps solve the problem. We proceed with the RL stage using these sentence-level scores. After RL, the training model may produce incorrect parts (colored in red), so we enhance the training model by requesting correction feedback from the AI model (colored in blue) and then proceed with the supervised fine-tuning stage.

**Correction Feedback:** Given the success of LLMs and LMMs in a wide range of areas (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022a), we are not restricted to requesting feedback in the form of scores. We request correction feedback from advanced AI (Teacher) for sentences containing errors after the RL process, and obtain a corrected dataset  $X_{\text{corrected}}$ . Since the supervised fine-tuning is more stable and finding appropriate hyperparameters is easier than RL, we proceed with supervised fine-tuning using  $X_{\text{corrected}}$  exactly as in common autoregressive model (Vaswani et al., 2023) training to stabilize the RL fine-tuned model. This reduces the burden of RL’s exhaustive hyperparameter tuning and properly guides the direction in which the training model wants to change.

**How Correction Feedback Helps RL:** RL increases the probability of positively rewarded actions (or sentences) and decreases the probability for negative rewards. The direction of learning is determined by the reward (scalar) value. However, the opposite direction of the reward is sometimes required. For example, suppose there is a truncated sentence  $s_{\text{truncated}}$  in CoT.  $s_{\text{truncated}}$  gets a negative score because it is an incomplete sentence (Table 13). If there is no correction stage, the probability of  $s_{\text{truncated}}$  is simply reduced. *What if  $s_{\text{truncated}}$  contains some valuable part?* This valuable part is ignored, and its probability decreases. To alleviate this issue, we instead receive the corrected sentence as feedback and encourage the training model to

generate complete sentences, which is very challenging to achieve with only RL. Table 16 shows more examples of how the correction stage helps the RL stage by maintaining the reasoning context while changing the erroneous parts.

Additionally, RL is primarily fine-tuned through PPO (Schulman et al., 2017) to prevent the model from deviating too much from the original model. The KL divergence penalty further prevents deviation. However, this penalty often causes the model’s degeneration. As a solution, InstructGPT (Ouyang et al., 2022) proposes PPO-ptx, where the supervised fine-tuning term with the pretraining dataset is included in the loss function. While this aims to align the training model with specific preferences, it tends to anchor the model to the pretraining dataset. Instead, we perform supervised fine-tuning through the Teacher’s correction feedback to allow the training model to more freely adapt and meet specific preferences without degeneration.

## 2.4 Algorithm Details

We propose a hybrid algorithm Alternating between REinforcement learning and Supervised fine-tuning (ARES). Figure 2 illustrates the ARES pipeline. First, we prepare a model with a given training dataset and generate rationale reasoning composed of several sentences for input. For the RL procedure to align the training model with a preference, we request scores for each sentence. The RL result may include some incorrect parts (colored in red as the 4th sentence in the RL result

box), but it aims to maximize the rewards provided. Next, we request correction feedback and create a corrected dataset (colored in blue as the 3rd sentence in the Corrected Rationale box) for the supervised fine-tuning stage. We then repeat the process from the RL stage until convergence.

### 3 Experimental Setup

**Data:** We first evaluate our proposed method on the ScienceQA (Lu et al., 2022a) dataset, a large-scale, multi-modal science dataset designed to assess multi-hop reasoning abilities. We choose ScienceQA because it contains reasoning chains to derive the answer. Each problem consists of a question, multiple options, multi-modal contexts, a correct answer, and an annotated lecture or solution chain (note that around 9.5% lack the solution chain). In addition, we conduct experiments on A-OKVQA (Schwenk et al., 2022), a knowledge-based multi-modal benchmark with a diverse set of challenging questions paired with rationales, demanding non-trivial commonsense knowledge (see Appendix B).

**Baselines:** We mainly compare our method with Multimodal-CoT (MM-CoT) (Zhang et al., 2023b) as the baseline because it utilizes reasoning chains to solve multi-modal tasks. MM-CoT leverages two distinct models: the first generates a rationale for a given problem, and the second, an inference model, takes the concatenated input (problem and generated rationale). This separated framework shows improved performance, even for relatively small models such as Flan-Alpaca<sub>Base</sub> (Chia et al., 2023) (251M) and Flan-Alpaca<sub>Large</sub> (790M). We use the rationale model provided by MM-CoT for ScienceQA and retrain the rationale model ourselves for A-OKVQA because there is no provided model.

**Prompts for Feedback:** Since our proposed ARES requests different types of feedback for each stage, a corresponding prompt exists separately. We use Claude 3 Haiku for all training to get feedback because it is approximately 20 times cheaper than the top competing models, yet still demonstrates decent performance. We first request scores ranging from 0.0 to 1.0 for each sentence in CoT to proceed with the RL stage. To obtain reasonable scores, we let Haiku consider the starting point of thought, the process of elimination, or true statements. (See

Table 5.)

In order to collect the corrected dataset for the SFT stage, we let Haiku refer to the given problem and correct the answer as the prompt. We ask Haiku to maintain the format of the existing rationale chains as much as possible and correct only the parts that require correction. The RL stage often makes the training model generate repetitive sentences. This repetition is not easily removed even by GPT-4 when the repetitive sentence exists in the middle of rationale reasoning. To reduce the burden of feedback, we simply hard-code the removal of repetitive sentences before adding the generated rationale to the prompt. (See Appendix C.2.)

**Training Details:** For the ARES<sub>Base</sub> RL stage, we use a learning rate of  $2e-5$  and 10 epochs for PPO with a batch size of 8 for both ScienceQA and A-OKVQA. The learning rate for ARES<sub>Large</sub> is  $2e-5$  with 5 epochs for PPO and a batch size of 2 for both tasks. We proceed with 2 rounds of our pipeline for ARES<sub>Base</sub> and 2 rounds for ARES<sub>Large</sub> for ScienceQA. For A-OKVQA, we proceed with 1 round for both model sizes. For the SFT stage for correction, we follow the hyperparameters used in MM-CoT for both model sizes. Additionally, we replace MM-CoT’s inference model, which is the same size as the rationale model, with the Low-Rank Adaptation (LoRA) (Hu et al., 2021) added to the rationale model (Figure 4). The LoRA adapter effectively utilizes the rationale model’s features with a small number of weights, enabling 2x–14x faster inference compared to MM-CoT, which introduces a separate inference model (See the time comparison in Table 7 and Table 8). For more detailed settings, please refer to Appendix C.

**Evaluation Metrics:** We use two main metrics to test how our pipeline (ARES) improves rationale reasoning quality. First, we evaluate ARES’s rationale reasoning quality against baseline models since we enhance our model based on them. For two different model sizes (Flan-Alpaca<sub>Base</sub> and Flan-Alpaca<sub>Large</sub>) and two tasks (ScienceQA and A-OKVQA), rationale reasoning quality is evaluated by GPT-4o-2024-05-13 and the win rate is calculated (Section 4.3). The GPT-4 series is actively used as an evaluation metric, replacing human judgment for various domains (Liu et al., 2023b; Sottana et al., 2023). Second, we assess

how the improved rationale reasoning impacts answer accuracy (Section 4.4). This evaluation is also performed on both model sizes and tasks. Additionally, we analyze how the RL stage fine-tunes the training model and maximizes the sum of rewards in Section 4.1.

## 4 Experimental Results

In this section, we evaluate our pipeline ARES that Alternates REinforcement Learning and Supervised Fine-Tuning by requesting diverse types of feedback for an advanced AI model (Teacher) (Claude 3 Haiku). The goal of ARES is to improve the rationale reasoning quality. We demonstrate how ARES enhances rationale reasoning in the following sections.

### 4.1 Emergent Behavior Through RL

Through RL, a training model is aligned to a specific preference. Essentially, the model increases the probability of helpful sentences receiving good rewards and reduces the probability of incorrect or meaningless sentences. However, this process produces some interesting additional results.

First, it supplements rationale reasoning for some problems where rationale reasoning is insufficient. In particular, 9.5% of problems in ScienceQA have empty rationale reasoning (solution) data. The model generates nothing before the RL stage for these problems but starts generating reasoning chains afterward (See Table 14). We observe this especially when utilizing PPO’s advantage normalization or when the learning rate is large.

Second, the training model begins to explain why other options are not the answer (See Table 15). The process of elimination is a useful method for deriving answers when options are given.

### 4.2 Guide RL with Correction

Despite the benefits of RL, hyperparameter tuning often requires massive effort. Without meticulous tuning, the RL fine-tuned model may produce errors such as repetitive or incomplete sentences. To address these issues, we add a supervised fine-tuning (SFT) stage after RL to correct these errors. SFT is more stable than RL. We evaluate how well the SFT stage corrects errors caused by the RL stage for various RL hyperparameters. We test various RL hyperparameters such as learning rate = {5e-6, 1e-5, 2e-5, 5e-5}, batch size = {2, 4, 8, 16,

ScienceQA	Win Rate
<b>ARES</b> <sub>Base</sub> vs MM-CoT <sub>Base</sub>	69.76%
<b>ARES</b> <sub>Large</sub> vs MM-CoT <sub>Large</sub>	73.76%
A-OKVQA	Win Rate
<b>ARES</b> <sub>Base</sub> vs MM-CoT <sub>Base</sub>	69.11%
<b>ARES</b> <sub>Large</sub> vs MM-CoT <sub>Large</sub>	66.96%

Table 1: We train baseline models, MM-CoT, with the ARES pipeline and ask GPT-4o to evaluate which rationale reasoning is better. We compare each baseline for two model sizes (**ARES**<sub>Base</sub> and **ARES**<sub>Large</sub>) and two tasks (ScienceQA and A-OKVQA).

32}, and PPO epoch = {5, 10, 15}. As a result of RL, we observe that some of the sentences in rationale chains are repetitive or truncated (see Table 13 and 12). The SFT stage, with correction feedback, reflects the direction in which the model is fine-tuned through RL and appropriately guides it (Table 13 and 16). However, excessive RL learning rates or epochs cause serious degeneration of the model, such as producing no output or generating strange words, and the results of correction feedback are also unreasonable.

### 4.3 Rationale Reasoning Comparison

We check whether ARES improves the quality of rationale reasoning compared to the baseline model. GPT-4o evaluates which rationale chain is better between the rationale generated by ARES and the rationale generated by the baseline model. We randomly shuffle the rationale chains and provide them as Option A and Option B (see Appendix A.3) for a fair evaluation (Yu et al., 2023). We conduct our experiments with two different model sizes, Flan-Base and Flan-Large with ViT feature, on ScienceQA and A-OKVQA. Table 1 shows that ARES achieves around 70% win rate against each corresponding baseline model for both datasets.

### 4.4 Inference Accuracy

We investigate whether the improved rationale also contributes to answer inference accuracy. Table 2 shows the main results of answer inference on the ScienceQA. We evaluate our base model against the MM-CoT baseline. **ARES**<sub>Base</sub> achieves a 2.79% improvement compared to the corresponding baseline (MM-CoT<sub>Base</sub>). The large model (**ARES**<sub>Large</sub>) shows some minimal improvement compared to the corresponding baseline. However, it’s worth noting

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN (Yu et al., 2019)	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down (Anderson et al., 2018)	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN (Kim et al., 2018)	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF (Peng et al., 2019)	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT (Kim et al., 2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al., 2022b)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al., 2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA <sub>Base</sub> (Khashabi et al., 2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA <sub>Base</sub> w/ CoT (Lu et al., 2022a)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
LLaMA-Adapter (Zhang et al., 2023a)	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA (Liu et al., 2023a)	13B	90.36	95.95*	88.00	89.49	88.00	90.66	90.93	90.90*	90.92
InstructBLIP (Dai et al., 2023)	11B	-	-	-	-	90.70*	-	-	-	-
MM-CoT <sub>Base</sub> (Zhang et al., 2023b)	251M+251M	84.59	92.46	83.45	83.87	83.29	85.64	86.34	85.23	85.95
<b>ARES<sub>Base</sub></b> (Ours)	251M+30M	<b>87.92</b>	<b>92.58</b>	<b>85.91</b>	<b>86.61</b>	<b>85.82</b>	<b>88.36</b>	<b>88.88</b>	<b>87.48</b>	<b>88.38</b>
MM-CoT <sub>Large</sub> (Zhang et al., 2023b)	790M+790M	90.76	93.59	86.55	89.69	87.85	89.55	90.90	89.12	90.26
<b>ARES<sub>Large</sub></b> (Ours)	790M+76M	<b>91.21*</b>	92.80	<b>89.45*</b>	<b>90.27*</b>	<b>88.35</b>	<b>91.22*</b>	<b>91.48*</b>	<b>90.38</b>	<b>91.09*</b>

Table 2: Main results on the ScienceQA test set (%). Size = backbone size. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Other results are sourced from Lu et al. (2022a) and Zhang et al. (2023b). Results in bold represent the better performance corresponding baseline. (\*) indicates the best performance.

that despite this seemingly small gain, **ARES<sub>Large</sub>** beats 13B LLaVA (Liu et al., 2023a). This minimal improvement may be due to the 9.5% of ScienceQA problems needing more rationale reasoning (around 9.5% problems have empty rationale reasoning). The RL stages can only eliminate some empty rationale reasoning, which requires numerous ARES pipeline rounds. Above all, our main goal is to assess how the RL stage works and how the SFT stage aids RL.

Table 3 shows the results of answer inference on the A-OKVQA. We retrain MM-CoT<sub>Base</sub> and MM-CoT<sub>Large</sub> and evaluate these on the validation set as in (Zhang et al., 2023b) because the test set is hidden. In our experiments, MM-CoT models perform around 10% better than the reported accuracy in (Zhang et al., 2023b). ARES achieves 4.45% gains against MM-CoT<sub>Base</sub> and 2.35% for MM-CoT<sub>Large</sub>.

In addition, we demonstrate that two stages, RL and SFT, are essential through an ablation study. Figure 3 shows the rationale reasoning for 4 cases. The baseline model (MM-CoT) produces the same rationale reasoning as the dataset. However, the corrected reasoning for MM-CoT without the RL stage has insufficient information compared to the reasoning of ARES that performs RL (refer to Table 17 for more examples). Table 4 also shows that inference accuracy gradually improves as each part of ARES is executed. 1st RL indicates a single RL run on MM-CoT, and 1st ARES means one

round of the ARES pipeline. 1st ARES & 2nd RL represents the second RL on 1st ARES, and finally, 2nd ARES refers to two rounds of ARES.

Model	Accuracy
IPVR (OPT-66B)	48.6
ViLBERT	49.1
MM-CoT <sub>Base</sub>	60.96
<b>ARES<sub>Base</sub></b> (Ours)	<b>65.41</b>
MM-CoT <sub>Large</sub>	65.68
<b>ARES<sub>Large</sub></b> (Ours)	<b>68.03</b>

Table 3: Results of ARES on A-OKVQA. We mainly compare different-sized MM-CoT baselines (Zhang et al., 2023b). We retrain the MM-CoTs and run the ARES pipeline on these models. We evaluate these models on the validation set because the test set is hidden.

## 5 Related Work

Chain-of-Thought (CoT) is a multi-step reasoning method for problem-solving that encourages LLMs to consider the intermediate reasoning steps. Zero-Shot-CoT (Kojima et al., 2023) promotes CoT by using prompts such as "Let's think step by step" for LLMs. For Few-Shot-CoT (Zhang et al., 2022b; Wei et al., 2023), a few examples with reasoning processes are provided, allowing the model to refer to these examples and understand how to perform CoT. Wei et al. (2023) reveal that this CoT

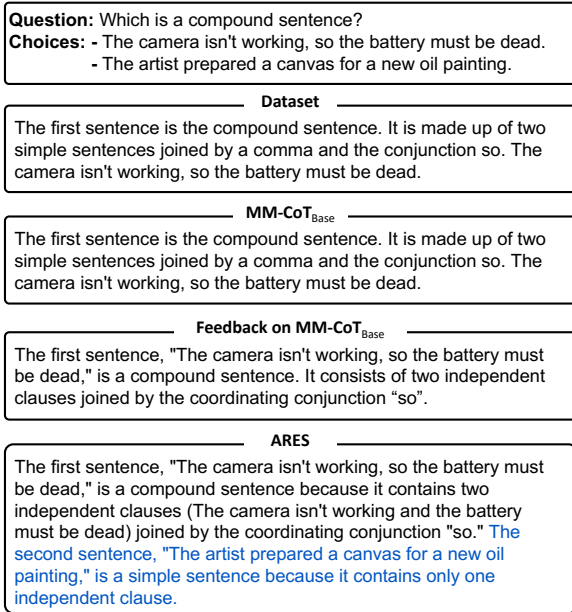


Figure 3: Comparison of rationales: dataset, baseline, correction feedback for baseline, and ARES.

technique positively impacts the performance of large models ( $> 100B$ ), but has minimal effect on smaller models. MM-CoT (Zhang et al., 2023b) suggest that CoT is beneficial even for relatively small models, such as 200M, if the model that generates intermediate reasoning and the model that infers the answer are separated. We find that simply adding a LoRA adapter (Hu et al., 2021) to the reasoning model results in comparable performance. This framework enables the LoRA adapter to effectively utilize all features, from raw text to latent features, and generates answers 2x–14x faster than MM-CoT, which uses a separate inference model (See Table 7 and Table 8). This speed advantage arises from the fact that our framework does not require a rationale as input, whereas the separate inference model framework must first generate the rationale before using it as input.

Reinforcement Learning from Human Feedback (RLHF) (Glaese et al., 2022; Ouyang et al., 2022) and AI Feedback (RLAIF) (Bai et al., 2022) align LLMs with user preferences. Ouyang et al. (2022) collects ranked feedback from human labelers and uses this feedback to perform Reinforcement Learning (RL). Constitutional AI (CAI) (Bai et al., 2022) collects ranked AI feedback rather than costly human feedback and handles harmfulness with RL. Both approaches learn outcome-supervised reward models (ORM) using ranking-based feedback. Lightman et al. (2023), instead, propose a

Model	Accuracy
MM-CoT <sub>Base</sub>	85.95
1st RL	86.70
1st ARES	87.81
1st ARES & 2nd RL	87.88
2nd ARES	88.38

Table 4: Ablation study: The accuracy gradually improves as each stage of ARES is added.

process-supervised reward model (PRM) that leverages sentence-level feedback for CoT. Lightman et al. (2023); Luo et al. (2024) evaluate each trained ORM and PRM with searching algorithms such as best-of- $N$  or Monte Carlo Tree Search (MCTS) by selecting the highest-scored solution, demonstrating that the PRM-selected solution outperforms the ORM-selected one. Wang et al. (2024) perform RL using PRM, providing heuristic sentence-level scores for math problems that are simple to grade. As an LLM is trained with RL and starts generating outputs different from the original distribution, these reward models would not correctly provide rewards (Pitis, 2023; Byun and Perrault, 2024). Instead of training a reward model for a more general task, we perform RL by requesting sentence-level rewards from advanced AI models such as GPT-4.

## 6 Conclusion

We propose a hybrid algorithm, ARES, which Alternates REinforcement Learning (RL) and Supervised Fine-Tuning (SFT) to enhance multi-modal rationale reasoning for ScienceQA and AOKVQA. ARES leverages two types of feedback: 1) ARES requests a score from a Teacher (we used Claude 3 Haiku) for sentence-level nuanced feedback and proceeds with RL. 2) ARES requests the Teacher to correct rationale chains after RL, stabilizing the RL fine-tuned model with SFT. ARES is designed to aid the RL procedure without massive hyperparameter tuning while properly reflecting the desired changes from RL. We evaluate the improvement in rationale reasoning produced by ARES compared to baselines using GPT-4o, and assess how much the improved rationale chains enhance inference accuracy for the two multi-modal tasks. We hope our work inspires further research on utilizing various types of AI feedback.



## Limitations

Although we address general multi-modal rationale models beyond mathematical problems, receiving feedback from AI models still needs to be more reliable for more complex tasks such as graduate-level math or expert-level knowledge. For instance, some A-OKVQA problems even contain challenging questions requiring external knowledge beyond the image alone. This challenge highlights the necessity for future research to develop methods that can effectively incorporate external knowledge sources into the model. Additionally, if the model is not publicly available for free, using the API incurs costs, and there are daily usage limits.

## Acknowledgments

We sincerely appreciate the generous support of computational resources provided by the Ohio Supercomputer Center (Center, 1987).

## References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). *Preprint*, arXiv:1707.07998.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. [Constitutional ai: Harmlessness from ai feedback](#). *Preprint*, arXiv:2212.08073.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Ju-Seung Byun and Andrew Perrault. 2024. [Symmetric reinforcement learning loss for robust learning on diverse tasks and model scales](#). *Preprint*, arXiv:2405.17618.
- Ohio Supercomputer Center. 1987. [Ohio supercomputer center](#).
- Yew Ken Chia, Pengfei Hong, and Soujanya Poria. 2023. [Flan-alpaca: Instruction tuning from humans and machines](#). <https://github.com/declare-lab/flan-alpaca>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#). *Preprint*, arXiv:2204.02311.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). *Preprint*, arXiv:2305.06500.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). *Preprint*, arXiv:2010.11929.
- Theresa Eimer, Marius Lindauer, and Roberta Raileanu. 2023. [Hyperparameters in reinforcement learning and how to tune them](#). *Preprint*, arXiv:2306.01324.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [Realtocixityprompts: Evaluating neural toxic degeneration in language models](#). *Preprint*, arXiv:2009.11462.
- Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh,

- Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. [Improving alignment of dialogue agents via targeted human judgements](#). *Preprint*, arXiv:2209.14375.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. [News summarization and evaluation in the era of gpt-3](#). *Preprint*, arXiv:2209.12356.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Daniel Khoshabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. [Unifiedqa: Crossing format boundaries with a single qa system](#). *Preprint*, arXiv:2005.00700.
- Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. [Bilinear attention networks](#). *Preprint*, arXiv:1805.07932.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. [Vilt: Vision-and-language transformer without convolution or region supervision](#). *Preprint*, arXiv:2102.03334.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Large language models are zero-shot reasoners](#). *Preprint*, arXiv:2205.11916.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. [Rlaif: Scaling reinforcement learning from human feedback with ai feedback](#). *Preprint*, arXiv:2309.00267.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. [Visualbert: A simple and performant baseline for vision and language](#). *Preprint*, arXiv:1908.03557.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. [Let’s verify step by step](#). *Preprint*, arXiv:2305.20050.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-eval: Nlg evaluation using gpt-4 with better human alignment](#). *Preprint*, arXiv:2303.16634.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. [Learn to explain: Multimodal reasoning via thought chains for science question answering](#). In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2022b. [Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning](#). *Preprint*, arXiv:2110.13214.
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. [Improve mathematical reasoning in language models by automated process supervision](#). *Preprint*, arXiv:2406.06592.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. 2019. [Dynamic fusion with intra- and inter-modality attention flow for visual question answering](#). *Preprint*, arXiv:1812.05252.
- Silviu Pitis. 2023. [Failure modes of learning reward models for LLMs and other sequence models](#). In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Preprint*, arXiv:1910.10683.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *CoRR*, abs/1707.06347.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. [A-okvqa: A benchmark for visual question answering using world knowledge](#). *Preprint*, arXiv:2206.01718.
- Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. [Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. [Understanding the capabilities, limitations, and societal impact of large language models](#). *Preprint*, arXiv:2102.02503.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#). *Preprint*, arXiv:1706.03762.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024. [Math-shepherd: Verify and reinforce llms step-by-step without human annotations](#). *Preprint*, arXiv:2312.08935.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. [Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning](#). *Preprint*, arXiv:2305.13660.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. [Deep modular co-attention networks for visual question answering](#). *Preprint*, arXiv:1906.10770.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. [Self-rewarding language models](#). *Preprint*, arXiv:2401.10020.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023a. [Llama-adapter: Efficient fine-tuning of language models with zero-init attention](#). *Preprint*, arXiv:2303.16199.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. [Opt: Open pre-trained transformer language models](#). *Preprint*, arXiv:2205.01068.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. [Automatic chain of thought prompting in large language models](#). *Preprint*, arXiv:2210.03493.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. [Multi-modal chain-of-thought reasoning in language models](#). *Preprint*, arXiv:2302.00923.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. [Secrets of rlhf in large language models part i: Ppo](#). *Preprint*, arXiv:2307.04964.

## A Prompts

### A.1 Prompt for Sentence-Level Nuanced Feedback

The prompt for obtaining sentence-level nuanced feedback by Claude 3 Haiku is illustrated in Table 5. Each reasoning sentence is assigned a value between 0.0 and 1.0.

- Values close to 0.0 indicate completely incorrect rationales.
- A value of 0.5 represents a neutral rationale, such as an initial thought process or true statements that aid in guiding guesses towards the correct answer.
- Values close to 1.0 denote a correct or highly relevant rationale.

These scores enable our model to discern the direction of changes through Reinforcement Learning (RL), reflecting the extent to which a sentence aids in resolving the problem.

### A.2 Prompt for Correction Feedback

Due to the challenges mentioned in Section 2.2, we adopt the correction feedback approach. The following are the specific instructions for obtaining correction feedback using Claude 3 Haiku. We have established the following seven rules for obtaining correction feedback using Claude 3 Haiku: The prompt is presented as a Table 9.

### A.3 Prompt for Win Rate Evaluation

We prompt GPT-4o (2024-05-13) to choose which generated rationale is better for solving the question because we don't have gold rationales. Given two generated rationales (e.g., MM-CoT<sub>Base</sub> and ARES<sub>Base</sub>), we ask GPT-4o: "You are given two rationale options (A or B). Your job is to select the better rationale between A and B for solving the given problem with the given image, choices, hint, and answer. Please output only 'A' or 'B'." (see Table 6). Yu et al. (2023) find that ChatGPT

Feedback	Prompt Structure
	<p><b>[Prompt when Image is provided]</b></p> <p>There exists a set comprising Image, Options, Hint, and Answer for a Question. The reasoning process used to deduce the answer is provided in JSON format. Fill in "xxx" with values ranging from 0.0 to 1.0, in increments of 0.1. The reasoning may include the starting point of thought, the process of elimination, or true statements, although these may not appear to be directly related to the answer at first glance. A value closer to 0.0 indicates a completely incorrect rationale, 0.5 indicates a neutral rationale such as the initial thought process or true statements that guide later guesses towards the answer, and a value closer to 1.0 denotes a correct or relevant rationale for the question. Please just fill the "xxx" parts and only return the JSON format. If a sentence is repetitive (appeared before), then give 0.0.</p>
Sentence-Level Nuanced Feedback	<p>Question: &lt;Question&gt;  Options: &lt;Choices&gt;  Hint: &lt;Hint&gt;  Answer: &lt;Answer&gt;</p> <pre>{   "&lt;Rationale 1&gt;": xxx,   "&lt;Rationale 2&gt;": xxx,   "&lt;Rationale 3&gt;": xxx }</pre> <p><b>[Prompt when no Image is provided]</b></p> <p>There exists a set comprising Options, Hint, and Answer for a Question. The reasoning process ... <i>&lt;same as the prompt when the Image is provided&gt;</i></p>

Table 5: Prompt structure for sentence-level nuanced feedback in a question-answering system. The table outlines the format for prompts when an image is provided and when no image is provided, detailing how to score the rationale for each sentence in terms of correctness and relevance.

is skewed towards choosing option A, so we randomly swap options A and B for each evaluation to avoid bias.

## B Difficulties with External Knowledge in A-OKVQA

The A-OKVQA dataset includes challenging questions paired with rationales that demand knowledge beyond the information available in the image. These questions cannot be answered simply by querying a knowledge base, as they require a deeper understanding and integration of external knowledge.

Our model faces difficulties with problems that

cannot be resolved using only the information from the image. While our approach is designed to improve rationales by addressing grammatical errors and incomplete or incorrect statements, it struggles with questions that necessitate external knowledge.

Table 10 illustrates an example where the question requires knowledge about the typical PSI range for bicycle tires. This information is not visually apparent from the image of the bicycle alone. To answer this question correctly, one needs external knowledge about standard bicycle maintenance practices and the recommended PSI ranges for different types of bicycle tires. This highlights a challenge for our model, as it must provide correct rationales and answers without access to such ex-

Evaluation	Prompt Structure
Win Rate	<p>You are given two rationales options (A or B). Your job is to select the better rationale between A and B for solving the given problem with the given Image, Choices, Hint, and Answer. Please output only A or B.</p> <p>Question: &lt;Question&gt;            Choices: &lt;Choices&gt;            Hint: &lt;Hint&gt;            Answer: &lt;Answer&gt;</p> <p>OPTION A: &lt;Rationales&gt;            OPTION B: &lt;Rationales&gt;</p>

Table 6: Prompt structure used for evaluating the win rate of generated rationales using GPT-4o.

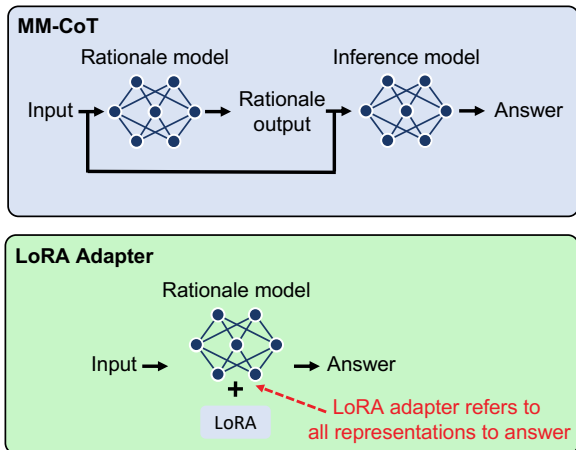


Figure 4: MM-CoT (Zhang et al., 2023b) uses two same size separate models for reasoning and inference. We replace the inference model with a LoRA adapter (only 1/10 the weights) added to the rationale model.

ternal knowledge, relying solely on the image and its internal knowledge base.

In the second example (Table 11), understanding what the first two numbers of an identification tag denote requires specific external knowledge about livestock tagging systems. Such tagging systems often use coded information, where the numbers might represent the birth month, year, or other identification details. This information is not visually apparent from the image and requires familiarity with agricultural practices or livestock management, illustrating the challenge for our model in providing correct rationales and answers without external knowledge.

## C Training Details

ScienceQA has 21K multi-modal problems, with 12K for training, 4K for validation, and 4K for testing. It also includes various difficulty levels from elementary to high school, covering domains like natural science, language science, and social science. In addition, we conduct experiments on A-OKVQA (Schwenk et al., 2022), a knowledge-based multi-modal benchmark with a diverse set of 25K questions A-OKVQA includes 25K questions (17K for training, 1K for validation, and 6K for testing).

We adapt the same T5 encoder-decoder architecture (Raffel et al., 2023) under Base and Large settings, following (Zhang et al., 2023b), and initialized with Flan-Alpaca (Chia et al., 2023).

### C.1 Reinforcement Learning

For the  $\text{ARES}_{\text{Base}}$  and  $\text{ARES}_{\text{Large}}$  training on the ScienceQA and A-OKVQA dataset, we employ the following settings:

**Common Settings:** We use top- $k$  sampling with  $k = 50$  and sample 4 actions. The initial coefficient for the Kullback-Leibler (KL) divergence is set to 0.0001. The range for clipping the probability ratios in PPO is 0.2. The discount factor is set to 1.0. Token length is constrained to 512. We train the model using 4 NVIDIA A100 80GB GPUs.

**$\text{ARES}_{\text{Base}}$  Specific Settings:** We use a learning rate of  $2e-5$  and 10 epochs for PPO with a batch

size of 8. Advantage normalization is applied for  $\text{ARES}_{\text{Base}}$ , and gradient accumulation steps are set to 8.

**$\text{ARES}_{\text{Large}}$  Specific Settings:** The learning rate for  $\text{Flan-Alpaca}_{\text{Large}}$  is  $2e-5$  with 5 epochs for PPO and a batch size of 2 for both tasks. Advantage normalization is not used and gradient accumulation steps are set to 16.

## C.2 Supervised Fine-Tuning

We use a batch size of 8 and train for 20 epochs with a learning rate of  $8e-5$  for  $\text{ARES}_{\text{Base}}$ , following (Zhang et al., 2023b). For  $\text{ARES}_{\text{Large}}$ , we use a batch size of 2 and train for 50 epochs with a learning rate of  $5e-5$ . The output length is set to 64 tokens. Training for  $\text{ARES}_{\text{Base}}$  utilizes 1 A100 GPU, while training for  $\text{ARES}_{\text{Large}}$  utilizes 4 A100 GPUs. In the MM-CoT paper (Zhang et al., 2023b), because the `final_eval` setting was not consistent, we retrained the base model with `final_eval=true` and the large model with `final_eval=false` for consistency.

**Token Cleanup:** In order to collect the corrected dataset, we need to identify tokens representing the end of each sentence, such as periods, question marks, and exclamation marks. In the ScienceQA dataset, a newline character often follows the ‘n’ being added after it. To reduce the burden of feedback, we simply hard-code the removal of repetitive sentences before adding the generated rationale to the prompt. We remove this ‘n’ and also ignore the backslash (\) character. For overlapping sentences, we placed each rationale of a problem into a list. If a rationale sentence was already in the list, we did not include it again during this preprocessing step.

## C.3 LoRA Adapter Training

MM-CoT utilizes two identically sized models for reasoning and inference tasks. In our approach, we replace the inference model with a LoRA adapter (Figure 4), which is added to the rationale model and consists of only one-tenth of the weights.

For LoRA adapter training for ScienceQA and A-OKVQA, we use a LoRA rank of  $r = 64$ , a LoRA  $\alpha = 128$ , and a LoRA dropout rate of 0.05. The learning rate is set to  $8e-5$  for both  $\text{ARES}_{\text{Base}}$  and  $\text{ARES}_{\text{Large}}$ . The batch size is 16 for  $\text{ARES}_{\text{Base}}$  and 4 for  $\text{ARES}_{\text{Large}}$  on the ScienceQA dataset. For

Model	Rationale	Inference	Total
$\text{MM-CoT}_{\text{Base}}$	1h 32m	7m	1h 39m
$\text{ARES}_{\text{Base}}$	-	8m	8m
$\text{MM-CoT}_{\text{Large}}$	5h 23m	12m	5h 35m
$\text{ARES}_{\text{Large}}$	-	24m	24m

Table 7: Time Comparison between MM-CoT and ARES models for ScienceQA test set

Model	Rationale	Inference	Total
$\text{MM-CoT}_{\text{Base}}$	6m	2m	8m
$\text{ARES}_{\text{Base}}$	-	3m	3m
$\text{MM-CoT}_{\text{Large}}$	16m	3m	19m
$\text{ARES}_{\text{Large}}$	-	6m	6m

Table 8: Time Comparison between MM-CoT and ARES models for A-OKVQA test set

the A-OKVQA dataset, the batch size is 4 for both  $\text{ARES}_{\text{Base}}$  and  $\text{ARES}_{\text{Large}}$ .

## C.4 Time Comparison between LoRA Adapter and Inference Model

Rather than introducing a separate model for inference, we achieve comparable performance by adding the LoRA adapter to the rationale model, while simultaneously obtaining a 2x–14x speedup in inference (ARES) compared to MM-CoT, which introduces a separate inference model. We provide the time comparison in Table 7 and Table 8.

This speed gap is mainly due to the fact that the separate inference model requires rationale generation before the inference procedure. However, the LoRA adapter directly refers to the rationale model’s latent features to derive answers, eliminating the need to generate the rationale first.

## D Comparison of Generated Rationales

As mentioned in Section 2.3 and Section 4.1, because RL increases the probability of sentences receiving positive rewards and reduces the probability of sentences receiving negative rewards, the trained model often exhibits specific phenomena. It tends to generate repetitive and incomplete sentences (Table 12 and Table 13). Before the RL steps, the model couldn’t produce rationales, but after RL steps, it starts generating meaningful rationale reasoning (Table 14). Furthermore, it begins to generate reasons why other options are not the answer (Table 15).

As illustrated in Table 17, we compare the solutions from the ScienceQA original dataset,

the rationales generated by the baseline model (**MM-CoT**<sub>Base</sub>), the rationales from the baseline model with correction feedback applied, and the rationales generated by our model (**ARES**<sub>Base</sub>). The first example, "Which property do these three objects have in common?" illustrates that the baseline model generates incorrect rationales such as "The lemon is not (yellow)" and "All three objects are rough. The property that all three objects have in common is rough." However, when we apply correction feedback to the rationales generated by the baseline model and compare it to our proposed method, we see that our approach generates correct rationales that include the correct answer and provide explanations on why other options are not the answer. The second example also shows that our method improves rationale reasoning.

Feedback	Prompt Structure
Correction Feedback	<p data-bbox="544 495 967 528"><b>[Prompt when Image is provided]</b></p> <p data-bbox="544 533 1385 600">Your task involves reviewing a set that includes an Image, Options, Hint, Answer, and Rationales for a Question. Please follow below 7 rules.</p> <ol data-bbox="544 604 1385 1137" style="list-style-type: none"> <li data-bbox="544 604 1385 705">1. Preserve any correct original rationales based on the given answer by incorporating them into the final rationale without making any alterations.</li> <li data-bbox="544 710 1385 777">2. Preserve any original rationales that represent the starting point of thought.</li> <li data-bbox="544 781 1385 848">3. Correct any grammatical errors or incomplete rationales based on the given information without your knowledge.</li> <li data-bbox="544 853 1385 920">4. If there are incorrect rationales based on the given answer, please correct them without removing them based on the given information.</li> <li data-bbox="544 925 1385 992">5. Please take into account the content of the options, hint, and answer when doing this task.</li> <li data-bbox="544 996 1385 1064">6. Fill the corrected rationales inside the {} in the final_rationale according to the given format below, without any additional explanation.</li> <li data-bbox="544 1068 1385 1135">7. Return only the entire set of Rationales within curly braces ({} ) below with the filled one in the step 6.</li> </ol> <p data-bbox="544 1178 807 1317">Question: &lt;Question&gt; Options: &lt;Choices&gt; Hint: &lt;Hint&gt; Answer: &lt;Answer&gt;</p> <p data-bbox="544 1359 946 1534">Rationales: { original_rationale:{&lt;Rationales&gt; final_rationale:{} }</p> <p data-bbox="544 1576 1005 1610"><b>[Prompt when no Image is provided]</b></p> <p data-bbox="544 1615 1385 1715">Your task involves reviewing a set that includes Options, Hint, Answer, and Rationales for a Question. ... <i>&lt;same as the prompt when the image is provided&gt;</i></p>

Table 9: Prompt structure for correction feedback. The table details the rules and format for reviewing and correcting rationales when an image is provided and when no image is provided. Each set includes a question, options, hint, answer, and rationales, with specific instructions on preserving, correcting, and formatting the rationales.



---

**A Challenging Example from A-OKVQA**

---

**Image**



---

**Question**      What is the correct PSI for a bicycle tire?

---

**Choices**      100-120 psi, 80-130 psi, 40-90 psi, 50-80 psi

---

Table 10: A challenging example question from the AOKVQA dataset. The question asks for the correct PSI for a bicycle tire, which requires external knowledge beyond what is depicted in the image. The choices provided represent typical PSI ranges for different types of bicycle tires.

---

**A Challenging Example from A-OKVQA**

---

**Image**



---

**Question**      The first two numbers of the identification tag denote what?

---

**Choices**      birth month, own number, birth year, birth date

---

Table 11: A challenging example question from the AOKVQA dataset. The question asks for the meaning of the first two numbers on an identification tag, which requires external knowledge about livestock tagging systems. The choices provided represent typical information encoded in such tags.

---

**Example of the Repetitive Sentences After the RL Stage**

---

**Problem:** Which tense does the sentence use? "Mona will print her name with care."

**Choices:** [present tense, future tense, past tense]

**Answer:** future tense

---

Before RL

**[Rationale reasoning]**

The sentence is in future tense. You can tell because it uses will before the main verb, print. The verb tells you about something that is going to happen.

---

After RL

**[Rationale reasoning]**

The sentence is in future tense. You can tell because it uses will before the main verb, print. The verb tells you about something that is going to happen. **The verb tells you about something that is going to happen. The verb tells you about something that is going to happen. The verb ends in -s and tells you about something that is going to happen...**

---

Table 12: Example of the repetitive sentences after the RL stage.

---

**Example of the Incomplete Sentence After the RL Stage**

---

**Problem:** What is the volume of a large soup pot?

**Choices:** [7 liters, 7 milliliters]

**Answer:** 7 liters

---

Before RL

**[Rationale reasoning]**

The better estimate for the volume of a large soup pot is 7 liters. 7 milliliters is too little.

---

After RL

**[Rationale reasoning]**

The better estimate for the volume of a large soup pot **is-**

---

Table 13: Example of the incomplete sentence after the RL stage.

<b>Changed Empty Rationale Reasoning Through RL</b>	
<b>Problem:</b> Select the solid.	
<b>Choices:</b> [rain, water in a fishbowl, hammer]	
<b>Answer:</b> hammer	
Before RL	<b>[Rationale reasoning]</b> <empty>
After RL	<b>[Rationale reasoning]</b> Rain is a liquid. A liquid takes the shape of any container it is in. If you put rainwater into a bucket, the rainwater will take the shape of the bucket. But the rainwater will still take up the same amount of space. The water in a fishbowl is a liquid. A liquid takes the shape of any container it is in. If you pour water from a fishbowl into a different container, the water will take the shape of that container. But the water will still take up the same amount of space. A hammer is a solid. A solid has a size and shape of its own. Many hammers are made of iron and wood. The water in a fishbowl is a liquid. A liquid takes the shape of any container it is in. If you put water from a fishbowl into a different container, the water will take the shape of that container. But the water will still take up the same amount of space. A rain is a liquid. A liquid takes the shape of any container it is in. If you put rainwater into a bucket, the rainwater will take the shape of the bucket. But the rainwater will still take up the same amount of space. A hammer is a solid. A solid has a size and shape of its own. Many hammers are made of iron and wood. Both iron and wood are solids.

Table 14: Example of an empty rationale reasoning that starts generating rationale reasoning after the RL stage.

<b>Example of the Process of Elimination After the RL Stage</b>	
<b>Problem:</b> Which is a compound sentence?	
<b>Choices:</b> ["Last night we heard an owl outside in the oak tree.", "I ate all of my dinner, so I can have some pie."]	
<b>Answer:</b> "I ate all of my dinner, so I can have some pie."	
Before RL	<b>[Rationale reasoning]</b> The first sentence is the compound sentence. It is made up of two simple sentences joined by a comma and the conjunction so. I ate all of my dinner, so I can have some pie.
After RL	<b>[Rationale reasoning]</b> The first sentence is the compound sentence. It is made up of two simple sentences joined by a comma and the conjunction so. I ate all of my dinner, so I can have some pie. The second sentence does not have a comma, so it is not a compound sentence. Last night we heard an owl outside in the oak tree.

Table 15: Example of the process of elimination rationale reasoning after the RL stage.


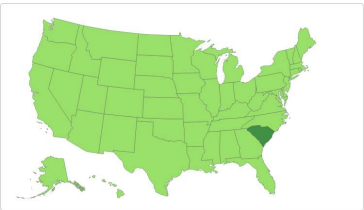
<b>Comparison of Corrected Rationales After RL</b>	
<b>Image</b>	
<b>Question</b>	Will these magnets attract or repel each other?
<b>Choices</b>	repel, <b>attract</b>
<b>After RL</b>	Will these magnets attract or repel? To find out, look at which poles are closest to each other. The north pole of one magnet is closest to the south pole of the other magnet. Poles that are the same repel. So, these magnets will repel each other. The picture shows two magnets that are the same shape and the same size. Both magnets are the same size. So, these magnets will repel each other.
<b>Corrected Rationale</b>	The magnets will attract each other because the north pole of one magnet is closest to the south pole of the other magnet. Opposite poles of magnets attract each other in this diagram. Therefore, the correct answer is that the magnets will attract.
<b>Question</b>	How long is a pen?
<b>Choices</b>	20 kilometers, <b>20 centimeters</b>
<b>After RL</b>	The better estimate for the length of a pen is 20 centimeters. 20 kilometers is too long. 20 centimeters is too short. 20 kilometers is too long.
<b>Corrected Rationale</b>	The better estimate for the length of a pen is 20 centimeters. 20 kilometers is an unreasonable estimate, as it is far too long for a typical pen. 20 centimeters is a more accurate and reasonable estimate for the length of a standard pen.
<b>Image</b>	
<b>Question</b>	What is the capital of South Carolina?
<b>Choices</b>	<b>Columbia</b> , Montgomery, Charleston, Harrisburg
<b>After RL</b>	Columbia is the capital of South Carolina.nThe city of Columbia is Charleston, South Carolina. Charleston is the smallest city in South Carolina.nHayeburg is the smallest city in South Carolina. Montgomery is the smallest city in South Carolina. But the city of Columbia is home to many cultural and historical landmarks.
<b>Corrected Rationale</b>	Columbia is the correct answer as it is the capital city of South Carolina. It is the largest city in the state and serves as the seat of government. The other options, Montgomery (the capital of Alabama), Charleston (a city in South Carolina), and Harrisburg (the capital of Pennsylvania), are not the capital of South Carolina.

Table 16: Examples of how the correction stage corrects mistakes after the RL stage.

### Comparison of Generated Rationales





<b>Image</b>	 <div style="display: flex; justify-content: space-around; font-size: small;"> <div style="text-align: center;"> rain boots</div> <div style="text-align: center;"> rubber duck</div> <div style="text-align: center;"> lemon</div> </div>
<b>Question</b>	Which property do these three objects have in common?
<b>Choices</b>	rough, <b>yellow</b> , sticky
<b>Original Solution</b>	Look at each object. For each object, decide if it has that property. A rough object feels scratchy when you touch it. The rain boots and the rubber duck are not rough. A sticky object can attach or stick to other things. The rain boots and the rubber duck are not sticky. Yellow is a color. This color is yellow. All three objects are yellow. The property that all three objects have in common is yellow.
<b>MM-CoT<sub>Base</sub></b>	Look at each object. For each object, decide if it has that property. A sticky object can attach or stick to other things. The rubber duck is not sticky. Yellow is a color. This color is yellow. The rubber duck and the rain boots are yellow, but the lemon is not. A rough object feels scratchy when you touch it. All three objects are rough. The property that all three objects have in common is rough.
<b>Correction feedback on MM-CoT<sub>Base</sub></b>	The property that all three objects have in common is yellow. The rubber duck and the rain boots are both yellow, and the lemon is also yellow. Although the objects may have other properties like roughness, the common property among them is the color yellow.
<b>ARES<sub>Base</sub> (Ours)</b>	The three objects shown - rubber duck, rain boots, and lemon - all share the common property of being yellow in color. The rubber duck is made of a smooth, non-sticky material, the rain boots are also a yellow-colored object, and the lemon is also a yellow-colored liquid. Therefore, the correct answer is that the common property among these three objects is that they are yellow. The other options, rough and sticky, do not apply to all three objects, as the rubber duck is not rough, and the rain boots are not sticky.
<b>Question</b>	Which word would you find on a dictionary page with the following guide words? lever - loan
<b>Choices</b>	lamb, <b>liquid</b>
<b>Original Solution</b>	Put the words in alphabetical order. Since liquid is between the guide words lever - loan, it would be found on that page.
<b>MM-CoT<sub>Base</sub></b>	Put the words in alphabetical order. Since lamb is between the guide words lever - loan, it would be found on that page.
<b>Correction feedback on MM-CoT<sub>Base</sub></b>	The word "liquid" would be found on the dictionary page with the guide words "lever - loan" because it falls alphabetically between those two words.
<b>ARES<sub>Base</sub> (Ours)</b>	The word "liquid" would be found on the dictionary page with the guide words "lever - loan" because it falls alphabetically between those two words. Therefore, the word "liquid" would be the word found on that dictionary page.

Table 17: Comparison of generated rationales by baseline, baseline with correction feedback, and our model (ARES).