

# APT: Improving Specialist LLM Performance with Weakness Case Acquisition and Iterative Preference Training

Jun Rao<sup>1</sup>, Zepeng Lin<sup>1</sup>, Xuebo Liu<sup>1\*</sup>, Xiaopeng Ke<sup>1</sup>, Lian Lian<sup>2</sup>, Dong Jin<sup>2</sup>, Shengjun Cheng<sup>2</sup>, Jun Yu<sup>3</sup>, and Min Zhang<sup>1</sup>

<sup>1</sup>Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China

<sup>2</sup>Huawei Cloud Computing Technologies Co., Ltd.

<sup>3</sup>School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen, China  
{rao7jun,zepenglin1,xiaopk7}@gmail.com, {liuxuebo,zhangmin2021}@hit.edu.cn  
{lianlian3,jindong2,chengshengjun}@huawei.com

## Abstract

Large Language Models (LLMs) often require domain-specific fine-tuning to address targeted tasks, which risks degrading their general capabilities. Maintaining a balance between domain-specific enhancements and general model utility is a key challenge. This paper proposes a novel approach named APT (Weakness Case Acquisition and Iterative Preference Training) to enhance domain-specific performance with self-generated dis-preferred weakness data (bad cases and similar cases). APT uniquely focuses on training the model using only those samples where errors occur, alongside a small, similar set of samples retrieved for this purpose. This targeted training minimizes interference with the model’s existing knowledge base, effectively retaining generic capabilities. Experimental results on the Llama-2 and Mistral-V0.3 models across various benchmarks demonstrate that APT ensures no reduction in generic capacity and achieves superior performance on downstream tasks compared to various existing methods. This validates our method as an effective strategy for enhancing domain-specific capabilities without sacrificing the model’s broader applicability.

## 1 Introduction

Large Language Models (LLMs) can be applied to multiple domains, such as coding (Rasley et al., 2020), mathematics (Yu et al., 2024), general knowledge question answering (Dubois et al., 2023; Rao et al., 2023b). However, substantial differences exist among these domains and tasks, and real-world applications often suffer from limited domain-specific data due to industry barriers. Consequently, enhancing domain models with a small amount of domain-specific data combined with general data (Shi et al., 2023) has emerged as a prominent research focus. Recent work has emphasized self-training to improve task-specific capabilities,

\*Corresponding Author

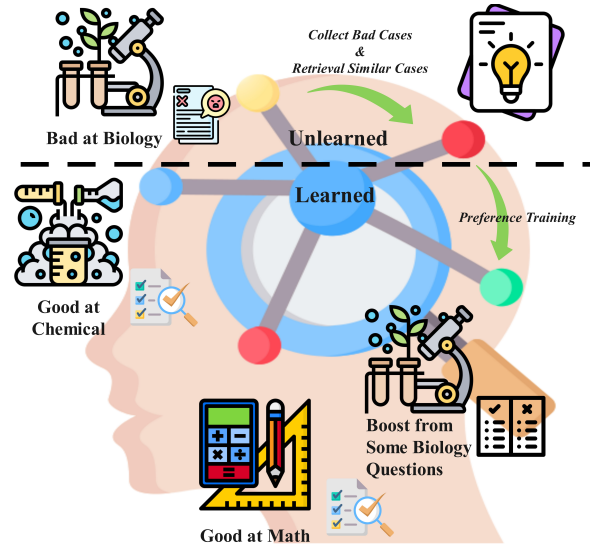


Figure 1: Similar to how people improve in weaker areas by focusing on their mistakes and practicing similar problems, our method identifies model deficiencies and uses targeted training to enhance performance.

particularly for coding and mathematical reasoning tasks (Cobbe et al., 2021; Lai et al., 2024; Xu et al., 2024b; Yu et al., 2024). These studies scale the original datasets through task-specific data generation strategies, resulting in enhanced domain models. Other approaches aim to improve models’ capabilities in specific domains without compromising general task performance, using techniques such as data mixing (Dong et al., 2024) and self-distillation (Yang et al., 2024). Existing methods typically focus on individual tasks or enhance models by constructing more instruction fine-tuned data. The former approach suffers from limited generalization due to its focus on specific tasks, while the latter primarily leverages information from positive samples and lacks feedback from negative samples.

In this paper, we propose the APT method, which utilizes a small number of self-generated negative samples for iterative preference training. The dis-preferred data generation process does not

require complex task-specific data construction. Similar to how humans learn from similar problems shown in Figure 1, our method has the model generate responses to tasks and add incorrect responses (“bad cases”) back into this iteration training data. Additionally, a retrieval module identifies similar samples (“similar cases”) to add more relevant training data. These two types of data are then used for training, which enables the model to learn from its errors and generalize across related tasks, effectively improving its performance on specific tasks. The model training component involves constructing preference data, where model-generated responses are non-preference data and original correct answers are preference data. Through iterative preference learning, the model optimizes its output by recognizing and correcting errors while learning from similar samples. This approach strengthens the model’s ability to handle challenging cases, ultimately improving performance.

Through extensive experimental validation, we show that APT is effective in identifying and optimizing model flaws across several domains, such as mathematical reasoning, coding, and instruction-following. The results indicate improvements of up to 5.9% in Llama2-7b and 6.0% in Mistral-7b, outperforming existing methods such as Dong et al. (2024). Additionally, ablation studies show that acquiring bad case data, retrieving similar cases, and optimizing preference data training objectives and iterations all contribute to enhancing model performance. Final scaling experiments demonstrate the method’s potential for future improvements on stronger models. Our contributions are as follows:

- We present APT, the first preference optimization framework that combines the model’s self-generated data of bad cases and retrieved relevant data to further enhance the model through iterative preference training.
- APT shows the ability to enhance special tasks while ensuring that general-purpose tasks (seven large-scale datasets) are not degraded and has been validated on a number of specific domains, including mathematical reasoning, coding, and instruction following.
- We find that LLMs improve through iterative preference training by learning from retrieval of similar mistakes, much like humans, and even in much stronger domain models.

## 2 Related Work

### 2.1 Alignment Finetuning

Alignment learning trains models to align outputs with human preferences or goals, often using techniques like reward shaping or reinforcement learning to optimize performance and minimize unexpected behaviors. Some methods use instruction fine-tuning (Xu et al., 2023; Peng et al., 2023; Cheng et al., 2024; Rao et al., 2024) to facilitate learning with high-quality examples. Others leverage contrastive data to enhance model consistency with human intentions by increasing positive responses and reducing negative ones. Rafailov et al. (2023) propose Direct Preference Optimization (DPO) to efficiently train large models for knowledge alignment using preference rankings instead of reward models. DPO optimizes classification loss from preference data, making implementing it simpler than reinforcement learning from human feedback (RLHF). Subsequent studies have used DPO for multi-domain alignment. Zhao et al. (2024b) introduces Group Preference Optimization to improve alignment training further.

### 2.2 Self-Training

In the past, most training data sources consisted of existing manually curated supervised corpora. In contrast, as large models evolve, synthetic data demonstrates significant potential for application. Recent efforts leverage the generative power of GPT-4 (OpenAI, 2023) to synthesize various instruction fine-tuning and alignment datasets (Peng et al., 2023; Conover et al., 2023; Xu et al., 2023; Deng et al., 2023, 2024). Earlier work, such as Self-Critiquing (Saunders et al., 2022), used a model’s own output to blend with the original answer distribution, enhancing model performance. Many recent studies (Zhu et al., 2024; Bai et al., 2023), including Nemotron (Parmar et al., 2024), extensively use synthetic data and have established automated synthetic data generation processes. Zeng et al. (2024) employs large models to evolve the instruction fine-tuning dataset, enhancing multi-tasking capabilities. Zhao et al. (2024a) leveraged model output errors to evolve the model and improve its performance. Chen et al. (2024) use LLMs to generate training data from previous iterations, refining policies by distinguishing between self-generated responses and those derived from human annotations. Our work employs a self-play mechanism to address model weaknesses. It identi-

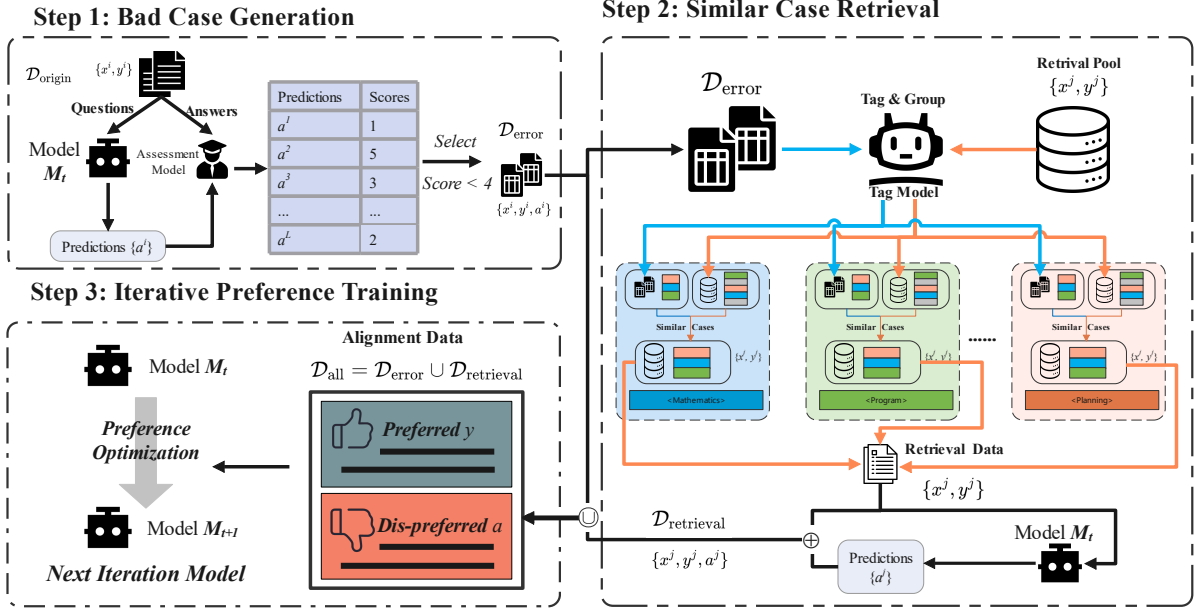


Figure 2: An overview of our approach APT. The main components of our method include: 1) Bad Case Generation: We classify samples with low scores from the assessment model as “Bad Cases”, which the model struggles to learn effectively. The predictions are then combined with the ground truth to create an Error Preference dataset  $D_{\text{error}}$ . Appendix A.8 presents the assessment prompt; 2) Similar Case Retrieval: We use retrieval to identify examples related to  $D_{\text{error}}$  from the retrieval pool, thus forming the dataset  $D_{\text{retrieval}}$ ; and 3) Iterative Preference Optimization:  $D_{\text{error}}$  and  $D_{\text{retrieval}}$  are the final preference datasets. Appendix A.1 presents the algorithm.

ifies poorly processed samples and retrieves similar samples to improve the processing of these cases.

### 3 Method

#### 3.1 Overview

As illustrated in Figure 2, we begin by defining the components involved in this iterative process:

- **Initial Model:** Let the existing model be denoted as  $M_t = f_{\theta}$ , where  $\theta$  represents the model parameters. This model serves as the starting point for subsequent iterations. Additionally, let  $f_{\theta_0}$  be the assessment model (Prometheus 2-7B) (Kim et al., 2024), which is used to evaluate the performance of  $M_t$  throughout the iterative process.
- **Initial Dataset:** We define the task-specific supervised fine-tuning (SFT) dataset as  $D_{\text{origin}} = (x^i, y^i)_{i=1}^L$ , where each pair  $(x^i, y^i)$  consists an input-output example used for model generation and identifying bad cases.
- **Retrieval Pool:** We define the retrieval pool we constructed as  $D_{\text{pool}} = (x^j, y^j)_{j=1}^U$ .
- **Preference Dataset:** The error dataset  $D_{\text{error}} = \{(x^i, y^i, a^i)\}_{i=1}^N$  is constructed by

treating the model’s prediction  $a^i$  as dis-preference and the corresponding ground truth  $y^i$  from  $D_{\text{origin}}$  as preference. The samples in  $D_{\text{error}}$  are specifically selected from those cases where the assessment model  $f_{\theta_0}$  assigns a low score. The dataset  $D_{\text{retrieval}} = \{(x^j, y^j, a^j)\}_{j=1}^M$  is formed by organizing the similar data retrieved during the retrieval phase similar to  $D_{\text{error}}$ . Finally, we obtain the preference dataset  $D_{\text{all}} = D_{\text{error}} \cup D_{\text{retrieval}}$ .

#### 3.2 Bad Case Generation

Similar to the human error correction process, LLMs can improve their weaknesses by actively learning from challenging samples and enhancing their most vulnerable areas. Based on this idea, we use initial model  $M_t$  to self-predict the questions  $\{x^i\}_{i=1}^L$  from the SFT training dataset  $D_{\text{origin}}$  and obtain the corresponding predictions  $\{a^i\}_{i=1}^L$ . These predictions, along with their corresponding ground truth answers  $\{y^i\}_{i=1}^L$  and the original questions  $\{x^i\}_{i=1}^L$ , are evaluated using the Assessment Model  $f_{\theta_0}$ . For each instance, the score  $s^i = f_{\theta_0}(x^i, y^i, a^i)$  quantifies how well the response of the model aligns with human evaluations. By selecting data instance  $(x^i, y^i, a^i)$  with low score  $s^i$ , we identify cases where the model

struggles and fails to handle the data effectively.

### 3.3 Similar Case Retrieval

Bad cases are relatively rare due to the strong capabilities of specialist models. This scarcity may lead to overfitting during model training. Existing approaches to address this include constructing similar question-answer pairs by model-generated rewrites (Lai et al., 2024), regenerating similar questions (Cheng et al., 2024), and employing back-translation (Yu et al., 2024). However, these methods require further design and filtering for specific tasks and lack generalizability. In this work, we adopt a more direct strategy: retrieving data related to the current bad case using the existing high-quality instruction data as the retrieval source.

**Retrieval Pool Construction** The structure of data retrieval pool  $\mathcal{D}_{\text{pool}}$  is as follows: we follow the configuration of Wang et al. (2023a) to collect a comprehensive and representative sample of datasets across various styles. These datasets include: (1) existing NLP resources, such as SuperNI (Wang et al., 2022) and Flan V2 (Chung et al., 2024); (2) datasets created from scratch by humans specifically for instruction tuning, such as Lima (Zhou et al., 2023); (3) datasets generated by proprietary models, including Self-Instruct (Wang et al., 2023b), Unnatural Instructions (Honovich et al., 2023), GPT4-Alpaca (Peng et al., 2023), Baize (Xu et al., 2023), OpenOrca (Lian et al., 2023), WizardLM (Xu et al., 2024a), and MetaMath (Yu et al., 2024); and (4) datasets designed for specialized skills, such as CoT (Wei et al., 2022) for chain-of-thought reasoning.

**Tag-Based Retrieval** We propose the TAG-BASED SIMILARITY method, which leverages a tagging model (Lu et al., 2024) to enhance retrieval precision by categorizing data into finer-grained classes and performing tag-specific similarity searches. Tag statistics are provided in Appendix A.2, and the detailed procedure is as follows: Given an error case preference dataset  $\mathcal{D}_{\text{error}}$  and a retrieval pool  $\mathcal{D}_{\text{pool}}$ , we apply the tagging model to the input  $x$  of each data point. Specifically, the tagging model assigns one or more tags  $\mathcal{T}(x)$  to each input  $x$ , representing its semantic categories such as ‘‘Mathematics’’ or ‘‘Program’’. For each tag  $t \in \mathcal{T}$ , we extract subsets:

$$\mathcal{D}_{\text{error}}^t = \{(x^i, y^i, a^i) \in \mathcal{D}_{\text{error}} : t \in \mathcal{T}(x^i, y^i)\}, \quad (1)$$

$$\mathcal{D}_{\text{pool}}^t = \{(x^j, y^j) \in \mathcal{D}_{\text{pool}} : t \in \mathcal{T}(x^j, y^j)\}, \quad (2)$$

where  $t$  represents a semantic category assigned by the tagging model. Consequently, this process produces multiple subsets,  $(\mathcal{D}_{\text{error}}^t, \mathcal{D}_{\text{pool}}^t)$ , each associated with a specific tag  $t$ . For each pair of subsets  $(\mathcal{D}_{\text{error}}^t, \mathcal{D}_{\text{pool}}^t)$ , we use the Sentence Transformer<sup>1</sup> (Reimers and Gurevych, 2019) to generate embedding vectors for the input  $x$  of each data point, resulting in  $\mathbf{e}(x^i)_{i=1}^N$  and  $\mathbf{e}(x^j)_{j=1}^U$ , respectively. Next, for the subset  $\mathcal{D}_{\text{error}}^t$ , we compute the average embedding vector:

$$\mathbf{e}_{\text{avg}}^t = \frac{1}{|\mathcal{D}_{\text{error}}^t|} \sum_{(x^i, y^i, a^i) \in \mathcal{D}_{\text{error}}^t} \mathbf{e}(x^i). \quad (3)$$

Then, the cosine similarity between  $\mathbf{e}_{\text{avg}}^t$  and the embedding vectors of the samples in  $\mathcal{D}_{\text{pool}}^t$  is computed as follows:

$$S(\mathbf{e}_{\text{avg}}^t, x^j) = \frac{\mathbf{e}_{\text{avg}}^t \cdot \mathbf{e}(x^j)}{\|\mathbf{e}_{\text{avg}}^t\| \|\mathbf{e}(x^j)\|} \quad (4)$$

Based on the computed similarity scores, we select the most similar samples from  $\mathcal{D}_{\text{pool}}^t$  such that their number matches the size of  $\mathcal{D}_{\text{error}}^t$ , forming a retrieved subset  $\mathcal{D}_{\text{retrieval}}^t$ . Finally, we aggregate the retrieved subsets across all tags to construct the complete retrieval set:

$$\mathcal{D}_{\text{retrieval}} = \bigcup_{t \in \mathcal{T}} \mathcal{D}_{\text{retrieval}}^t \quad (5)$$

### 3.4 Iterative Preference Training

To enhance the alignment of LLMs with human preferences, fine-tuning loss is commonly employed the following:

$$L_{\text{SFT}}(\boldsymbol{\theta}) = - \sum_{i=1}^N y_{w_i} \log(\hat{y}_{w_i} | x) \quad (6)$$

Training reward functions is challenging in practice, but DPO (Rafailov et al., 2023) simplifies this process using a predefined preference dataset, which we use  $\mathcal{D}_{\text{all}}$  here. The objective optimization function for this process is as follows:

$$L_{\text{DPO}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ref}}) = \mathbb{E}_{(x, y, a) \sim \mathcal{D}_{\text{all}}} \left[ \ell \left( \lambda \log \frac{p_{\boldsymbol{\theta}}(y | x)}{p_{\boldsymbol{\theta}_{\text{ref}}}(y | x)} - \lambda \log \frac{p_{\boldsymbol{\theta}}(a | x)}{p_{\boldsymbol{\theta}_{\text{ref}}}(a | x)} \right) \right] \quad (7)$$

<sup>1</sup>Here, we use the all-MiniLM-L6-v2 model to generate embedding vectors.



Model	Method	Domain Dataset	Math Reasoning	Coding	Instruction Following	General Capability	AVG
LLama2-7B	Base	-	15.5	19.2	7.5	56.9	24.8
	SFT	GSM	32.1	18.5	34.3	58.5	35.9
		CodeAlpaca	17.7	23.6	29.2	58.7	32.3
		Dolly	19.3	20.9	16.9	59.1	29.1
	Mixed Training	-	34.1	25.0	19.1	59.1	34.3
	+Continued SFT	GSM	32.6	27.4	19.5	59.0	34.6
		CodeAlpaca	31.9	28.1	21.1	58.9	35.0
		Dolly	33.2	28.1	18.0	59.1	34.6
	+DMT (Dong et al., 2024)	GSM	34.0	26.9	18.5	59.0	34.6
		CodeAlpaca	33.4	26.4	18.3	58.5	34.2
		Dolly	33.7	26.0	18.0	58.4	34.0
	+Ours	GSM	<b>39.2</b> (+5.1)	26.1	24.2	<b>59.4</b>	<b>37.2</b> (+2.9)
		CodeAlpaca	34.8	<b>28.4</b> (+3.4)	24.3	<b>59.3</b>	<b>36.7</b> (+2.4)
		Dolly	34.7	27.4	<b>25.0</b> (+5.9)	<b>59.3</b>	<b>36.6</b> (+2.3)
	Mistral-7B-V0.3	Base	-	40.6	33.1	16.4	64.5
SFT		GSM	56.9	41.1	33.8	<b>66.5</b>	49.6
		CodeAlpaca	46.4	43.8	30.3	66.1	46.7
		Dolly	43.8	41.0	26.0	<b>66.8</b>	44.4
Mixed Training		-	58.2	43.7	30.0	66.4	49.6
+Continued SFT		GSM	58.9	43.6	29.4	66.3	49.6
		CodeAlpaca	58.9	44.6	31.2	66.3	50.3
		Dolly	58.7	44.1	27.2	66.3	49.1
+DMT (Dong et al., 2024)		GSM	59.3	44.5	30.1	66.1	50.0
		CodeAlpaca	59.3	42.9	29.0	65.9	49.3
		Dolly	59.1	43.6	29.1	66.0	49.5
+Ours		GSM	<b>61.8</b> (+3.6)	45.9	30.6	<b>66.5</b>	<b>51.2</b> (+1.6)
		CodeAlpaca	60.2	<b>49.7</b> (+6.0)	31.2	<b>66.7</b>	<b>52.0</b> (+2.4)
		Dolly	59.6	47.6	<b>35.0</b> (+5.0)	<b>66.8</b>	<b>52.3</b> (+2.7)

Table 1: Main results on multiple LLMs across domain and general tasks. Continued SFT, DMT, and Ours are built upon the Mixed Training model, which is undergoing further training to improve domain-specific performance. Our approach achieves the best results across all domains while maintaining excellent performance overall.

Here,  $x \sim D$  is sampled from a given distribution  $D$ , and the KL regularization term prevents excessive deviation of the new model  $p_\theta$  from the reference model  $p_{ref}$ , with the regularization parameter  $\lambda > 0$ . However, DPO suffers from a decrease in the probability of choosing, and some work has been done to improve this by using different constraints (Hong et al., 2024; Pal et al., 2024). We directly use SFT loss as a constraint, and the final optimization function used in our experiments is shown below:

$$L(\theta) = L_{DPO}(\theta, \theta_{ref}) + \alpha L_{SFT}(\theta) \quad (8)$$

After training, we obtain an initial model for the next iteration to continue the previous data acquisition and training process.

### 3.5 Efficiency Discussion

Once generated, TAG annotations can be reused indefinitely, eliminating the need for repeated la-

beling. Despite processing a large volume of data, including content from the retrieval pool, the TAG model remains lightweight, with minimal token consumption, ensuring low computational overhead. In fact, the TAG process only requires an additional 5% of the total time, which is negligible in the overall workflow. Furthermore, when a new task arrives, only TAG processing and retrieval are necessary, incurring minimal time costs, which makes our approach both efficient and sustainable.

## 4 Experiment

### 4.1 Experimental Setup

**Training Datasets** Alpaca (Taori et al., 2023; Peng et al., 2023) is generated by the Self-Instruct framework and consists of 52K triplets, which is a common use of general domain data. We use CodeAlpaca, GSM, and Dolly for the special domain data. **CodeAlpaca** (Chaudhary, 2023) is an instruction fine-tuning training data about

the code, with about 20K samples. **GSM8K** dataset (Cobbe et al., 2021), which consists of 9K high-quality arithmetic word problems designed for the grade school level. **Dolly-v2** (Conover et al., 2023) is the open-sourced model trained on a 15K human-annotated instructions dataset. We constructed training datasets by combining three domain-specific datasets (Dolly, GSM8K, and CodeAlpaca) with the Alpaca dataset respectively.

**Baselines** We compare existing base and State-of-the-Art methods for upgrading specialist models, including SFT, Mixed Training, Continued SFT, and DMT. Consistent with current work (Taori et al., 2023; Xu et al., 2023; Zhu et al., 2024) on instruction tuning, we use standard SFT as the most basic comparison method. Mixed Training (Dong et al., 2024) is another baseline that combines multiple domains (GSM, Dolly and CodeAlpaca) directly to train a model. Continued SFT continues training with the Mixed Training model but focuses on the filtered error data, incorporating our tag retrieval. DMT (Dong et al., 2024) learns specialized abilities first and then learns general abilities with a small amount of specialized data. Details are provided in the Appendix A.4, A.5, and A.6.

## 4.2 Main Results

As shown in Table 1, we present the results across multiple tasks and various baseline methods (see Appendix A.7 for full results). Overall, our approach outperforms the baseline methods across different domains. The improvement is particularly significant on Llama-2 on the Alpaca-eval and GSM8K benchmarks, with improvement exceeding 7% and 4%, respectively. This suggests that APT, by selectively enhancing training on data where the model underperforms, can achieve gains similar to human learning from mistakes, significantly improving its performance in that domain. We showcase these improvements through case studies, detailed in the Appendix A.9. Notably, APT does not significantly drop in general capability tasks, indicating stable performance in these areas. This also highlights the importance of retrieving additional general data similar to the erroneous cases. By incorporating more relevant similar general data, the model is able to alleviate the forgetting phenomenon associated with continuous training, thereby preserving its general capabilities. In summary, our method improves performance on domain-specific tasks by targeting areas where

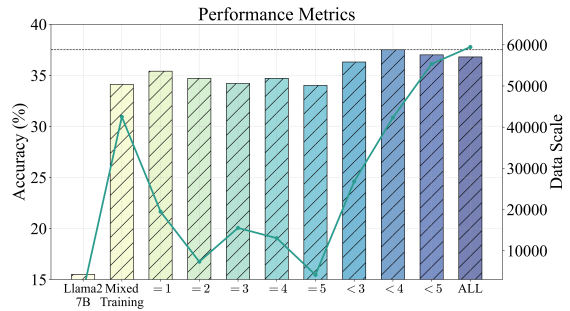


Figure 3: The impact of different response scores and their data volume. Selecting data with larger errors for continuous training will lead to better results and greater efficiency than merging all data (“<4” vs. “ALL”).

the model underperforms, while maintaining stable performance on general tasks.

## 4.3 Analysis on Bad Case Generation

**Effect of the Generation Cases** We test the results using data at all error levels using our training objective, as shown in Figure 3. “=1” represents selecting cases where the Assessment model’s scores are 1, based on the continuation of preference training after Mixed training. Similarly, “<3” represents selecting cases where the Assessment model’s scores are less than 3, and so on. The figure’s line represents the data volume used for training at each setting. The results indicate that data with lower scores yields better performance, suggesting that samples with noticeable or moderate errors play a crucial role in model training.

**Effect of the Assessment Model** To verify whether the assessment model’s scoring effectively helps in filtering out low-quality responses, we replace the original assessment model with the generation model itself to score the generated samples, followed by an iterative training process. We conduct a comparative experiment on LLama-2-7B, as shown in Table 2. We can observe that the model cannot evaluate the quality of its own predictions alone. Therefore, an additional assessment model is required to aid in the evaluation, much like how humans rely on a standard answer or a teacher to identify mistakes during the learning process.

## 4.4 Analysis on Similar Case Retrieval

**Effect of Retrieval Method** We perform experiments on LLama-2 to evaluate the retrieval step, as shown in Table 3. Specifically, we compare the baseline (Mixed training) and further fine-tuning based on mixed training with error samples alone

Domain	Self		Prometheus	
	Domain	AVG	Domain	AVG
GSM	34.8	34.6	39.2	37.2
CodeAlpaca	26.0	34.7	28.4	36.7
Dolly	18.7	34.9	25.0	36.6

Table 2: Comparison of assessment models in which “Self” refers to using the model itself as the assessment model and “Prometheus” refers to the assessment model employed in our method.

(only error) and with three retrieval strategies: mean vector, cluster-based, and our proposed tag-based similarity. Mean Vector Similarity computes the average embedding vector (Zeng et al., 2022; Karpukhin et al., 2020; Zhao et al., 2024c; Rao et al., 2023a) from all error samples and uses it to perform cosine similarity (Rao et al., 2022) comparisons to retrieve relevant items. Cluster Based Similarity groups error samples into clusters using the K-nearest neighbors algorithm (Cover and Hart, 1967; Wu et al., 2023), then calculates cosine similarity between the cluster centers and the candidate samples, including diverse yet similar cases. Our results reveal that retrieving similar samples significantly improves performance compared to direct training on error samples. Among the three retrieval strategies, the performance improvement increases as the retrieval granularity becomes finer (Mean-Vector < Cluster-Based < Tag-Based) (Zhao et al., 2025).

**Effect of Retrieval Data Scale** We compare the results for retrieval sizes of  $1\times$ ,  $2\times$ , and  $3\times$  the original error data within an iteration on the final model’s effectiveness, as shown in Table 4. The results indicate that the model achieves optimal performance with  $1\times$  retrieval size while increasing the data volume to  $2\times$  or  $3\times$  leads to a significant decline (e.g., on Dolly and CodeAlpaca). This precisely indicates that for domain training, it is not the case that the more data, the better. Overly similar or excessive samples may introduce redundancy or noise, negatively impacting the model’s generalization ability.

#### 4.5 Analysis on Iterative Preference Training

**Effect of the Iteration** As shown in Figure 4, we show the results of the assessment of each domain after several rounds of self-iterations. The scores of the corresponding domains on the 3 domains can be further improved after iterative training. This improvement is most obvious the first time, and the

Domain Dataset	Mixed Training	Only Error	Mean Vector	Cluster Based	Tag Based
GSM	34.1	36.3	38.1	38.1	<b>38.1</b>
CodeAlpaca	25.0	26.6	27.7	26.9	<b>28.2</b>
Dolly	19.1	23.5	21.4	21.7	<b>23.7</b>

Table 3: Comparison of different retrieval methods on fine-tuning multiple domains. With our dis-preferred data and training method, the finer-grained retrieved samples can assist in learning about error cases.

Domain Dataset	$1\times$ Scale	$2\times$ Scale	$3\times$ Scale
GSM	<b>38.1</b>	37.7	37.7
CodeAlpaca	<b>28.2</b>	27.2	27.4
Dolly	<b>23.7</b>	21.3	21.2

Table 4: Impact of retrieval size on results. The results show that the best results are obtained by retrieving only an equal amount of data combined with our selected data and training method.

Domain Dataset	SFT	DPO	ORPO	Smaug	Ours
GSM	32.6	36.5	37.2	33.4	<b>38.1</b>
CodeAlpaca	28.1	27.4	26.3	25.4	<b>28.2</b>
Dolly	18.0	22.7	20.4	16.0	<b>23.7</b>

Table 5: Comparison results of optimization goals for multiple domains. SFT means to continue training with the error and retrieve data using SFT loss.

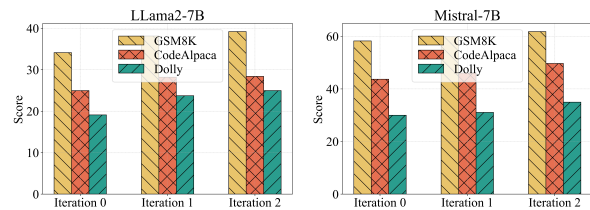


Figure 4: Domain evaluation of iteration. The iteration 0 corresponds to the mixed training model.

improvement in the subsequent iteration process is less relative to the first iteration. This shows the generalization ability of ours in multiple domains and the effectiveness of the iterative operation.

**Effect of the Optimization Objective** We compare here several variant approaches to DPO (Rafailov et al., 2023), namely ORPO (Hong et al., 2024), eliminating the necessity for an additional preference alignment phase and Smaug (Pal et al., 2024), which adds a regular term to mitigate the reduced probability of chosen in the original DPO. From the training results and the average results for each domain of Llama-2 shown in Table 5, introducing the SFT loss constraint can significantly help the model to strengthen the compliance of the instructions and thus obtain a better

Domain	Pred.		APT	
	Domain	AVG	Domain	AVG
GSM	35.9	36.1	39.2	37.2
CodeAlpaca	26.8	35.9	28.4	36.7
Dolly	23.5	36.3	25.0	36.6

Table 6: Exploring the impact of bad case filtering and similar case retrieval. “Pred.” refers to the iterative DPO process using negative generation without selection.

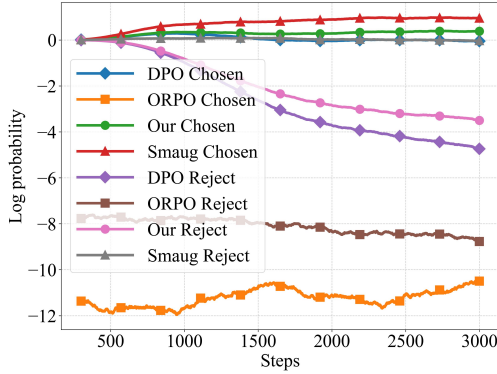


Figure 5: Comparing multiple different preference optimization objectives. The results show that our optimization objective better improves the probability of chosen samples while reducing the probability of rejected ones.

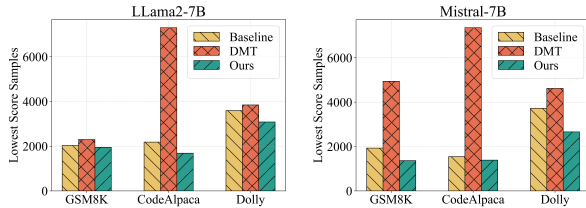


Figure 6: Comparison of the Number of Error Cases across Methods. After our method’s training, the model’s error samples were significantly reduced compared to the baseline and DMT.

domain effect in three domains. We show that models trained with our objective learned to reflect the preference throughout the training process. As shown in Figure 5, we compare the training curves of the logarithmic probabilities chosen and rejected for each optimization objective. As the training progresses, the probability of choosing increases, and the probability of rejected responses decreases, suggesting that our method successfully preserves the contribution of SFT samples to the training while reducing the occurrence of unwanted responses in the model unlike DPO is unchanged.

**Effect of the Bad Case Filtering and Similar Case Retrieval** As shown in Table 6, merely generating negative samples to construct DPO data for

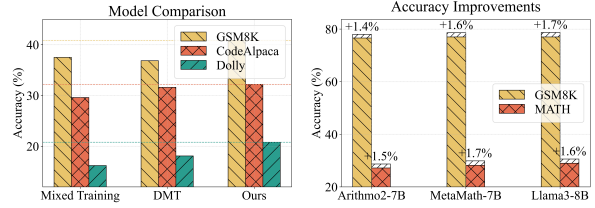


Figure 7: Scalability on Stronger Models. (a) Applying APT to a broader set of domain models enhances the model’s capabilities. (b) Extending APT to advanced mathematical models further improves performance.

iterative preference training fails to significantly enhance model performance. Instead, this training process benefits from the entire method, which includes filtering bad cases and retrieving similar cases. The combination of all the components finally results in the best performance.

#### 4.6 Method Robustness

**Training Robustness** To better illustrate how APT leverages erroneous samples identified by the assessment model to enhance the model’s performance, We show the number of erroneous samples, in which assessment model judgment is less than 4 scores, before (Baseline) and after training (Ours) across three domains in Figure 6. We compare the results of the three methods, mixed training and DMT, and our method with only one iteration. The results show that our method significantly reduces the number of incorrect samples of responses in the training set in several domains while boosting the test set, indicating that our method indeed learns from the training samples.

**Model Scale Robustness** We show scalability with larger model Llama2-13B and several leading models in mathematics, such as Arithmo2 (Jindal, 2023) and MetaMath (Yu et al., 2024) and Llama3-8B (Dubey et al., 2024). For Llama2-13B, we repeated the experiment similar to the main table, and the results show that it still brings an improvement over the existing methods, as shown in Figure 7. Although these mathematical models are sufficiently trained on data from the mathematical domain, additional training on the erroneous data can still enhance their performance. Combining APT still improves the model’s results, demonstrating the method’s scalability potential in a wider range of models with greater scalable potential.



## 5 Conclusion

This work introduces a unified framework APT to enhance domain-specific LLM capabilities further. The framework significantly improves performance across multiple areas, including instruction-following, mathematical reasoning, and coding. We successfully identified model deficiencies and effectively optimized them by finding similar cases for targeted training, thereby improving the overall model performance. Our results demonstrate the potential for further optimization of LLMs on challenging data like humans, providing a promising solution for enhancing model capabilities.

## Limitations

Our current limitations are similar to those typically encountered in reinforcement learning studies (Pal et al., 2024; Ivison et al., 2024). Specifically, the performance of our model is primarily constrained by the quality of the scoring information provided. Since our approach relies heavily on the accuracy of the scoring model, any limitations in its ability to effectively evaluate our outputs may lead to a degradation of the overall iterative reward mechanism. Furthermore, with respect to the retrieval step for error-related data, we acknowledge that our method is currently dependent on the performance of the InsTag model (Lu et al., 2024). While the retrieval process is focused on relevance, it is relatively coarse-grained and may not fully capture the necessary diversity. Additionally, due to resource constraints, we have only conducted experiments on models up to the 13B scale and have not been able to extend our experiments to larger models, such as those with 70B parameters, to further explore the effectiveness of our method at scale.

## Ethics Statement

Our work follows the ACL Ethics Policy. Our findings are based on publicly available datasets for reproducibility purposes. LLMs can contain potential racial and gender bias. The relevant code cannot be linked publicly (github) due to internal company constraints. If you need to do research, please contact the first author by email to get it.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62206076), Guangdong S&T Program

(Grant No. 2024B0101050003), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011491), and Shenzhen Science and Technology Program (Grant Nos. ZDSYS20230626091203008, KJZD20231023094700001, KQTD2024072910215406). We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

## References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. *Qwen technical report*. *arXiv preprint arXiv:2309.16609*.
- Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. <https://github.com/sahil280114/codealpaca>.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. *Evaluating large language models trained on code*. *arXiv preprint arXiv:2107.03374*.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. *Self-play fine-tuning converts weak language models to strong language models*. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Jiale Cheng, Yida Lu, Xiaotao Gu, Pei Ke, Xiao Liu, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. *AutoDetect: Towards a unified framework for automated weakness detection in large language models*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6786–6803, Miami, Florida, USA. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. *Scaling instruction-finetuned language models*. *J. Mach. Learn. Res.*, 25:70:1–70:53.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. *BoolQ: Exploring the surprising difficulty of natural yes/no questions*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind

- Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- 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*.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. [Free dolly: Introducing the world’s first truly open instruction-tuned llm](#).
- T. Cover and P. Hart. 1967. [Nearest neighbor pattern classification](#). *IEEE Transactions on Information Theory*, 13(1):21–27.
- Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. [Improving simultaneous machine translation with monolingual data](#). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12728–12736. AAAI Press.
- Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Min Zhang, and Zhaopeng Tu. 2024. [Newterm: Benchmarking real-time new terms for large language models with annual updates](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 35760–35795. Curran Associates, Inc.
- Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. [How abilities in large language models are affected by supervised fine-tuning data composition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Alpacafarm: A simulation framework for methods that learn from human feedback](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, et al. 2024. [A framework for few-shot language model evaluation](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. [ORPO: Monolithic preference optimization without reference model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. [Unnatural instructions: Tuning language models with \(almost\) no human labor](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. [Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Ashvini Jindal. 2023. [Arithmo-mistral-7b: Mathematical reasoning model](#). Hugging Face.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An open source language model specialized in evaluating other language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. [Step-dpo: Step-wise](#)

- preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval).
- Wing Lian, Bley Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. <https://https://huggingface.co/Open-Orca/OpenOrca>.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.
- Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddhartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.
- Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Dan Su, Chen Zhu, Deepak Narayanan, Aastha Jhunjhunwala, Ayush Dattagupta, et al. 2024. Nemotron-4 15b technical report. *arXiv preprint arXiv:2402.16819*.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Jun Rao, Liang Ding, Shuhan Qi, Meng Fang, Yang Liu, Li Shen, and Dacheng Tao. 2023a. Dynamic contrastive distillation for image-text retrieval. *IEEE Transactions on Multimedia*, pages 1–13.
- Jun Rao, Xuebo Liu, Lian Lian, Shengjun Cheng, Yunjie Liao, and Min Zhang. 2024. CommonIT: Commonality-aware instruction tuning for large language models via data partitions. In *EMNLP*, pages 10064–10083, Miami, Florida, USA. Association for Computational Linguistics.
- Jun Rao, Xv Meng, Liang Ding, Shuhan Qi, Xuebo Liu, Min Zhang, and Dacheng Tao. 2023b. Parameter-efficient and student-friendly knowledge distillation. *IEEE Trans. Multim.*, pages 1–12.
- Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao. 2022. Where does the performance improvement come from - a reproducibility concern about image-text retrieval. In *SIGIR*.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *arXiv preprint arXiv:2305.14705*.
- Chufan Shi, Yixuan Su, Cheng Yang, Yujiu Yang, and Deng Cai. 2023. Specialist or generalist? instruction



- tuning for specific NLP tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15336–15348, Singapore. Association for Computational Linguistics.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. **Challenging BIG-bench tasks and whether chain-of-thought can solve them**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. **Stanford alpaca: An instruction-following llama model**. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, et al. 2023a. **How far can camels go? exploring the state of instruction tuning on open resources**. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. **Self-instruct: Aligning language models with self-generated instructions**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. **Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. **Chain-of-thought prompting elicits reasoning in large language models**. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. **Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. **Wizardlm: Empowering large pre-trained language models to follow complex instructions**. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. **Baize: An open-source chat model with parameter-efficient tuning on self-chat data**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics.
- Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Zhao Wenyi, et al. 2024b. **ChatGLM-math: Improving math problem-solving in large language models with a self-critique pipeline**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9733–9760, Miami, Florida, USA. Association for Computational Linguistics.
- Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. **Self-distillation bridges distribution gap in language model fine-tuning**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. **Meta-math: Bootstrap your own mathematical questions for large language models**. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. **Curriculum learning for dense retrieval distillation**. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. **Automatic instruction evolving for large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6998–7018, Miami, Florida, USA. Association for Computational Linguistics.
- Haokun Zhao, Haixia Han, Jie Shi, Chengyu Du, Jiaqing Liang, and Yanghua Xiao. 2024a. **Large language model can continue evolving from mistakes**. *arXiv preprint arXiv:2404.08707*.



- Siyao Zhao, John Dang, and Aditya Grover. 2024b. [Group preference optimization: Few-shot alignment of large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, and Min Zhang. 2024c. [SEER: Self-aligned evidence extraction for retrieval-augmented generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3041, Miami, Florida, USA. Association for Computational Linguistics.
- Xinping Zhao, Yan Zhong, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Dongfang Li, Baotian Hu, and Min Zhang. 2025. [FunnelRAG: A coarse-to-fine progressive retrieval paradigm for RAG](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3029–3046, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. [LIMA: less is more for alignment](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. 2024. [Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence](#). *arXiv preprint arXiv:2406.11931*.

---

**Algorithm 1** Our Proposed APT Method

---

**Input:** fine-tuned model  $f_\theta$ , assessment model  $f_{\theta_0}$ , labeled dataset  $\mathcal{D}_{\text{origin}} = \{(x^i, y^i)\}_{i=1}^L$ , retrieval datasets pool  $\mathcal{D}_{\text{pool}} = \{(x^j, y^j)\}_{j=1}^U$

**Output:** fine-tuned model  $f_\theta$

- 1: **repeat**
    - # Bad Case Generation
  - 2:   Generate dataset of bad cases  $\mathcal{D}_{\text{error}}$ :
    - Generate the predictions  $\{a^i\}_{i=1}^L$  using fine-tuned model  $f_\theta$  of questions  $\{x^i\}_{i=1}^L$
    - Generate the scores using assessment model  $f_{\theta_0}$  of labeled dataset and predictions  $\{(x^i, y^i, a^i)\}_{i=1}^L$
    - Select low-score cases and generate  $\mathcal{D}_{\text{error}} = \{(x^i, y^i, a^i)\}_{i=1}^N$ , where  $x^i \sim \mathcal{D}_{\text{origin}}$
    - # Similar Case Retrieval
  - 3:   Generate dataset of similar cases  $\mathcal{D}_{\text{retrieval}}$ :
    - Find the similar and the same number of questions  $\{x^j\}_{j=1}^M$  using  $\mathcal{D}_{\text{error}}$  as queries from  $\mathcal{D}_{\text{pool}}$
    - Generate  $\mathcal{D}_{\text{retrieval}} = \{(x^j, y^j, a^j)\}_{i=j}^M$ , where  $x^j \sim \mathcal{D}_{\text{pool}}$  and  $a^j \sim f_\theta(x^j)$
    - # Preference training
  - 4:   Tune  $f_\theta$  with Eq.8 on  $\mathcal{D}_{\text{error}}$  and  $\mathcal{D}_{\text{retrieval}}$  to get new  $f_\theta$
  - 5: **until** convergence or max iteration is reached
- 

## A Appendix

### A.1 APT Algorithm

The workflow is shown in Algorithm 1. APT consists of two major components: data generation and model training. To initiate the process, we must gather comparative data for training. This involves obtaining both positive and negative examples, where the original answers are treated as preferred and the generated answers as disfavored, using model generation. Additionally, we retrieve samples that resemble the negative examples to expand the dataset. The current model generation outputs are also categorized as disfavored, allowing us to construct a new set of preference pairs that include both original and retrieved questions. Model preference training includes our defined formulas and iterative operations. Once the data has been constructed, iterative training can be initiated. In this process, the model is trained with newly generated preference data inputs in each iteration, producing a progressively stronger model.

### A.2 Statistics about Tag-Based Similarity

The Tag-Based Similarity is the core retrieval approach in our framework, and the quality of the tags plays a crucial role in the accuracy of the retrieval process. The most frequently occurring tags in each dataset are as follows: for the Alpaca dataset, the most common tag is “Information Retrieval”; for GSM8k, it is “Mathematics”; for Code\_Alpaca, the primary tag is “Program”; and for Dolly, it is also “Information Retrieval”. In the Retrieval Pool, the tag “Translation” is most common.

### A.3 Case Study about Tag-Based Similarity

For each domain, we will provide two pairs of related data, with one from the retrieval pool and the other from the domain-specific data, and display their respective tag annotations.

### A.4 Evaluation Details

**Evaluations** We use MMLU (Hendrycks et al., 2021), BBH (Suzgun et al., 2023), ARC (Clark et al., 2018), BoolQ (Clark et al., 2019), OpenBookQA (Mihaylov et al., 2018), and Winogrande (Sakaguchi et al., 2020) to evaluate general capabilities, similar to previous works (Shen et al., 2023; Wang et al., 2023a; Longpre et al., 2023). We evaluate the general performance of the model by computing and reporting the average scores across these datasets. We select three domain evaluation benchmarks (GSM (Cobbe et al., 2021), Human-eval (Chen et al., 2021) and Alpaca-eval (Dubois et al., 2023)) to test reasoning, coding, and the following instructions.

**Generic Capacity.** These evaluation benchmarks test the model’s ability to handle a diverse range of question types and domains, comprehensively measuring the model’s generalization ability. MMLU (Hendrycks et al., 2021) consists of a series of questions, ranging from basic to professional levels, across 57 academic subjects. Its multiple-choice format facilitates a relatively straightforward testing process. We use the official MMLU evaluation script and prompts<sup>2</sup>, with modifications to allow for batch processing. We evaluate using five few-shot examples, following the original setup of MMLU. We follow the setup described in the original paper of BBH (Suzgun et al., 2023), and evaluate with chain-of-thought (CoT). The prompts were officially provided with three short

<sup>2</sup><https://github.com/hendrycks/test>

in-context examples. For the CoT setup, we extract the first word after the phrase ‘So the answer is,’ or the entire response if no such substring is present. For ARC, BoolQ, OpenBookQA, and WinGrande tasks, we use EleutherAI LM Harness (Gao et al., 2024) for evaluation and follow its default settings for the assessment. Specifically, the ARC (AI2 Reasoning Challenge) dataset (Clark et al., 2018) provides both easy and challenge-level questions to assess the model’s performance in reasoning tasks of varying difficulty. BoolQ (Clark et al., 2019) tests the model’s ability to answer yes/no questions based on context, while OpenBookQA evaluates its understanding of scientific knowledge. WinoGrande (Sakaguchi et al., 2020) focuses on testing the model’s commonsense reasoning ability. **Coding.** For evaluating the coding capabilities of the models, we employ the HumanEval dataset presented in the Codex paper (Chen et al., 2021). This dataset encompasses 164 programming challenges, wherein models are prompted to finalize a Python function based on its provided docstring. Following the original paper, we calculate the pass@10 to gauge the functional accuracy of the models’ outputs. **Math Reasoning.** The samples of GSM are divided into 7K training and 1K test problems. We report the results of the exact match. **Instruction Following.** We use AlpacaEval (Li et al., 2023) to evaluate the model’s instruction-following capabilities, employing GPT-3.5 as the evaluation model during the assessment.

### A.5 Training Details

For the domain of the SFT model and the baseline mixed training, we utilized the LLama2-7B base and Mistral-7B with a learning rate set to  $5 \times 10^{-6}/1 \times 10^{-6}$  and a batch size of 128. For the DMT and our method, we use the LoRA (Hu et al., 2022) fine-tuning initialization from the baseline model. The entire training was conducted without any weight decay. We applied a linear learning rate schedule, incorporating a warmup phase where the warmup proportion was 3% of the total training steps. In the preference optimization for the LLama2-7B and Mistral-7B using LoRA tuning, we adjusted the learning rate to  $1 \times 10^{-6}/2 \times 10^{-7}$  with a reduced batch size of 32, while still omitting weight decay. The same linear learning rate schedule was employed with a 3% warmup rate. Based on the experimental results, as shown in Figure 8, the regularization parameter in Equation 8 for the

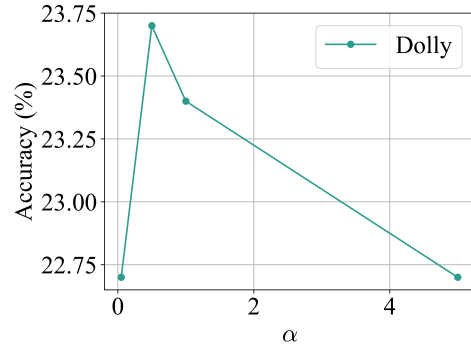


Figure 8: The exploration of hyperparameters.

preference loss is set to 0.5.

### A.6 Retrieval Details

Our retrieval method builds on the error-preference dataset generated during the bad case generation step. Using the tag-based similarity method, we select an equal amount of retrieval data from the retrieval pool to supplement the error-preference dataset, as justified in the ablation study (Figure 3). For example, combining GSM8K (7K) and Alpaca (52K), we identified 42K low-scoring samples (scores < 4) and retrieved an additional 42K samples from the retrieval pool, creating an 84K preference dataset for optimization in the current iteration.

### A.7 General Capability Experiments

In this work, we assess the general capabilities of the model using several widely recognized benchmark tasks: MMLU, BBH, ARC, BoolQ, OpenBookQA, and WinoGrande. To quantify the model’s overall capability and detect any potential forgetting, we report the average performance across these tasks, providing an aggregate measure of the model’s general competence. The results are shown in Table 7. The overall performance of our method demonstrates a slight improvement over the Mixed Training baseline, suggesting that additional domain-specific training does not lead to forgetting in our approach.

### A.8 Prompts

The following prompt is the built-in evaluation prompt of Prometheus 2-7B, used to assess the model’s responses.

### A.9 Case Study

We present detailed case studies. GSM8K rectifies logical errors in reasoning chains to derive

Method	Domain Dataset	MMLU	BBH	ARC-e	ARC-c	Boolq	OpenBookQA	WinoGrande	Avg.
<b>LLama2-7B</b>	-	45.2	40.7	74.6	46.3	77.7	44.2	69.3	56.9
<b>SFT</b>	GSM	46.7	39.2	76.8	49.7	79.9	47.4	69.7	58.5
	CodeAlpaca	47.2	40.6	76.9	50.9	79.3	46.8	69.1	58.7
	Dolly	48.2	39.3	78.3	52.5	79.3	47.0	69.5	59.1
<b>Mixed Training</b>	-	45.6	39.2	79.0	53.1	80.2	47.0	69.6	59.1
<b>+Continued SFT</b>	GSM	45.8	39.4	78.5	51.9	80.2	47.2	69.9	59.0
	CodeAlpaca	45.6	39.8	78.2	51.3	80.4	46.8	70.3	58.9
	Dolly	45.5	40.0	78.8	52.4	80.5	46.8	69.6	59.1
<b>+DMT (Dong et al., 2024)</b>	GSM	45.4	37.8	78.7	53.0	81.4	47.2	69.8	59.0
	CodeAlpaca	45.5	37.8	76.9	51.4	80.7	47.4	69.5	58.5
	Dolly	45.3	38.1	77.0	50.9	80.6	47.2	69.9	58.4
<b>+Ours</b>	GSM	45.4	39.4	78.9	52.6	81.0	48.0	70.9	<b>59.4</b>
	CodeAlpaca	45.2	39.7	78.7	52.7	80.7	48.0	70.2	<b>59.3</b>
	Dolly	45.1	39.0	79.4	53.4	80.5	47.8	69.9	<b>59.3</b>
<b>Mistral-7B-V0.3</b>	-	62.5	58.0	78.4	52.6	82.2	44.0	73.6	64.5
<b>SFT</b>	GSM	61.9	58.2	81.8	56.4	84.5	46.8	75.5	<b>66.5</b>
	CodeAlpaca	62.1	58.2	80.4	56.1	84.5	46.4	75.3	66.1
	Dolly	60.7	58.8	82.1	58.2	85.1	47.0	75.5	<b>66.8</b>
<b>Mixed Training</b>	-	61.0	58.5	81.7	57.3	85.6	46.4	74.4	66.4
<b>+Continued SFT</b>	GSM	60.8	58.5	80.9	56.7	85.4	46.2	75.3	66.3
	CodeAlpaca	61.0	58.5	81.2	56.7	85.5	46.2	74.8	66.3
	Dolly	61.0	58.2	81.3	56.6	85.4	46.2	75.1	66.3
<b>+DMT (Dong et al., 2024)</b>	GSM	61.0	59.5	79.0	55.4	85.9	46.8	74.7	66.1
	CodeAlpaca	60.9	58.5	78.6	55.3	86.2	47.2	74.7	65.9
	Dolly	61.0	59.6	78.9	55.6	86.0	46.6	74.7	66.0
<b>+Ours</b>	GSM	60.8	58.2	80.6	57.9	86.5	46.6	75.0	<b>66.5</b>
	CodeAlpaca	60.9	59.1	81.6	57.5	86.0	46.8	74.7	<b>66.7</b>
	Dolly	60.7	58.8	81.9	58.5	86.2	46.4	74.8	<b>66.8</b>

Table 7: The evaluation results of comprehensive performance across multiple tasks on several LLMs show that our approach does not lead to any degradation in performance in the general domain.

accurate final answers. On CodeAlpaca, the model fulfills task requirements more precisely, such as outputting the first ten Fibonacci numbers instead of just the tenth. On Dolly, it enhances its common-sense knowledge, corrects errors in commonsense reasoning, and provides more detailed and accurate answers. The mixed training answer represents the base model’s response, while our answer represents the response of the model after being trained on error cases. In the mixed training answer, the red text highlights the incorrect parts of the model’s response, while in our answer, the red text shows the corrected responses made by the model.



## Prompt for assessing the model's prediction

###Task Description: An instruction (might include an Input inside it), a response to evaluate, and a score rubric representing a evaluation criteria are given.

1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows: "(write a feedback for criteria) [RESULT] (an integer number between 1 and 5)"
4. Please do not generate any other opening, closing, and explanations.

###The instruction to evaluate:

{instruction}

###Response to evaluate:

{response}

###Reference Answer (Score 5):

{reference\_answer}

###Score Rubrics: You need to score the output of the model in terms of factual correctness, meeting the user's needs, logical coherence, and completeness.

Score 1: Answers provide inaccurate or incorrect information and do not fulfil the purpose and need of the user to ask the question, answers are not consistent overall or there are direct logical inconsistencies in different parts of the answer, answers do not provide enough information and important aspects are omitted.

Score 2: Answers provide inaccurate or incorrect information, fulfil the purpose and need of some users to ask questions, answers are consistent overall but there are logical inconsistencies between different parts, answers do not provide enough information and important aspects are omitted.

Score 3: The response provided inaccurate information, met the purpose and needs of the question posed by some users, fulfilled the formatting requirements of the question, and the overall logical coherence of the question was good, but the response did not provide enough information and omitted important aspects.

Score 4: The information provided in the response is accurate, fulfils the purpose and need of the question posed by some users, fulfils the formatting requirements of the question, the overall logical coherence is excellent, and the response provides sufficient information.

Score 5: The information provided in the responses was accurate and based on credible facts or data, the purpose and need of the questions posed by some users and the format of the questions were fully met, the overall logical coherence was excellent, and the responses provided sufficient information and detail.

###Feedback:

### Retrieval Case on gsm8k

**Retrieval data question 1:** Evie is collecting seashells while at the beach. Each day she collects her favorite 10 shells. At the end of 6 days, she gives 2 shells to her brother. How many shells does she have left? **Retrieval data tag 1:** ['mathematics', 'word problem', 'arithmetic', 'problem solve', 'counting', 'logic', 'mathematical operation', 'quantity', 'mathematical equation', 'mathematical reasoning'] **Domain data question 1:** Weng earns \$12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn? **Domain data tag 1:** ['mathematics', 'arithmetic', 'word problem', 'currency conversion', 'time conversion'] **Retrieval data question 2:** Task: Evaluate the statement: Humans will never colonize Mars. **Retrieval data tag 2:** ['evaluation', 'statement analysis', 'inference'] **Domain data question 2:** In 2021, Wayne is 37 years old. His brother Peter is 3 years older than him and their sister Julia is 2 years older than Peter. What year was Julia born in? **Domain data tag 2:** ['mathematics', 'logic', 'arithmetic', 'problem solve', 'data analysis', 'mathematical operation', 'word problem', 'age calculation', 'comparison', 'inference']

### Retrieval Case on Dolly

**Retrieval data question 1:** Come up with two solutions to this equation:  $3x+2y=7$ . **Retrieval data tag 1:** ['mathematics', 'problem solve', 'algebra', 'equation solve', 'mathematical operation', 'mathematical equation'] **Domain data question 1:** Let's imagine I create a fake currency called Yarns with two types of coins: a super-yarn (worth 10 yarns) and a mini-yarn (worth 0.5 yarn). How much money do I have in total if I have 2 super-yarns and three mini-yarns? **Domain data tag 1:** ['mathematics', 'currency', 'quantity', 'calculation', 'output'] **Retrieval data question 2:** Write a code snippet that uses a for loop to print the numbers 1 to 10. **Retrieval data tag 2:** ['program', 'loop', 'printing', 'number'] **Domain data question 2:** Instruction: How do compilers use IR? Input: An intermediate language is the language of an abstract machine designed to aid in the analysis of computer programs. The term comes from their use in compilers, where the source code of a program is translated into a form more suitable for code-improving transformations before being used to generate object or machine code for a target machine. **Domain data tag 2:** ['computer science', 'program', 'translation', 'object-oriented programming']

### Retrieval Case on CodeAlpaca

**Retrieval data question 1:** Can you tell me how to check if an input is in a list in Python? **Retrieval data tag 1:** ['python programming', 'list manipulation', 'input validation'] **Domain data question 1:** Instruction: Edit the Python code so that it prints out the total number of elements in the list. Input: `list = [1, 2, 3, 4, 5]` **Domain data tag 1:** ['python programming', 'list manipulation', 'output printing'] **Retrieval data question 2:** Write a code snippet that uses a for loop to print the numbers 1 to 10. **Retrieval data tag 2:** ['program', 'loop', 'printing', 'number'] **Domain data question 2:** Write a function to find the number of distinct states in a given matrix. `matrix = [[1, 0, 0], [1, 0, 1], [1, 1, 1]]` **Domain data tag 2:** ['program', 'matrix manipulation', 'algorithm', 'data structure', 'problem solve', 'computational thinking', 'logic', 'mathematics', 'problem analysis', 'data analysis', 'algorithm design', 'program concept']

## Dolly Case 1

**Question:** Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction: What are the primary uses of horses according to the text?

### Input: Horses and humans interact in a wide variety of sport competitions and non-competitive recreational pursuits as well as in working activities such as police work, agriculture, entertainment, and therapy. Horses were historically used in warfare, from which a wide variety of riding and driving techniques developed, using many different styles of equipment and methods of control. Many products are derived from horses, including meat, milk, hide, hair, bone, and pharmaceuticals extracted from the urine of pregnant mares. Humans provide domesticated horses with food, water, and shelter as well as attention from specialists such as veterinarians and farriers.

### Response:

**Mixed training answer:** Horses are used for sport, recreation, work, entertainment, and therapy.

**Our answer:** Horses are used for a variety of activities, including sport competitions, non-competitive recreational pursuits, working activities such as **police work, agriculture**, entertainment, and therapy. Horses were also used in **warfare**, which led to the development of riding and driving techniques. Horses also provide humans with products such as **meat, milk, hide, hair, bone, and pharmaceuticals**.

**Reference answer:** Horses can be used for recreational activities and therapy as well as working activities like police work or agriculture. Historically horses were also used in warfare. Horses can also provide products like meat, milk, hide, hair, bone or pharmaceuticals.

## Dolly Case 2

**Question:** Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction: Classify the following as types of birds or types of snakes: robin, cobra, rattlesnake, eagle, viper, raptor, bluejay, cottonmouth, copperhead, sparrow.

### Input:

### Response:

**Mixed training answer:** Birds: robin, **cobra, rattlesnake, viper**, raptor, bluejay.

Snakes: cottonmouth, copperhead, **sparrow**.

**Our answer:** Birds: robin, eagle, bluejay, raptor, **sparrow**

Snakes: **cobra, rattlesnake, viper**, cottonmouth, copperhead.

**Reference answer:** Robin, eagle, raptor, bluejay, and sparrow are all kinds of birds. Cobra, rattlesnake, viper, cottonmouth, and copperhead are several types of snakes.

### Dolly Case 3

**Question:** Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction: Tell me if these drinks are alcoholic or not: Coca-Cola, Mountain Dew, Budweiser, Rum, Vodka, Tomato Juice, Orange Juice, Sprite, Tequila, Margarita

### Input:

### Response:

**Mixed training answer:** Coca-Cola, Mountain Dew are alcoholic drinks. Budweiser, Rum, Vodka, Tequila are alcoholic drinks. Tomato Juice, Orange Juice, Sprite are not alcoholic drinks.

**Our answer:** Coca-Cola: Not Alcoholic

Mountain Dew: Not Alcoholic

Budweiser: Alcoholic

Rum: Alcoholic

Vodka: Alcoholic

Tomato Juice: Not Alcoholic

Orange Juice: Not Alcoholic

Sprite: Not Alcoholic

Tequila: Alcoholic

Margarita: Alcoholic

**Reference answer:** Budweiser, Rum, Vodka, Tequila and Margarita are alcoholic beverages.

### GSM8k Case 1

**Question:** Four people lost a total of 103 kilograms of weight. The first person lost 27 kilograms. The second person lost 7 kilograms less than the first person. The two remaining people lost the same amount. How many kilograms did each of the last two people lose?

**Mixed training answer:** Let  $x$  be the amount of weight the second person lost. The first person lost 27 kilograms, so the second person lost  $(x + 7)$  kilograms. The first and second person lost a total of  $27 + x + 7 = 34$  kilograms. The remaining two people lost a total of  $103 - 34 = 69$  kilograms. The last two people lost the same amount, so each of them lost  $(69/2 = 34.5)$  kilograms. The answer is 34.5.

**Our answer:** Let  $x$  be the amount of weight the second person lost. The first person lost 27 kilograms, so the second person lost  $(27 - 7 = 20)$  kilograms. The first two people lost a total of  $27 + 20 = 47$  kilograms. The last two people lost a total of  $103 - 47 = 56$  kilograms. So the last two people lost  $(56/2 = 28)$  kilograms each. The answer is 28.

**Reference answer:** Second person =  $(27 - 7 = 20)$  kg.  $103 - 27 - 20 = 56$  kg.  $56/2 = 28$  kg. The last two people each lost 28 kilograms of weight. The answer is 28.



### GSM8k Case 2

**Question:** Tim has 30 less apples than Martha, and Harry has half as many apples as Tim. If Martha has 68 apples, how many apples does Harry have?

**Mixed training answer:** Let  $x$  be the number of apples Harry has. Tim has 30 less apples than Martha, so he has  $x - 30$  apples. Harry has half as many apples as Tim, so he has  $\frac{x-30}{2}$  apples. Martha has 68 apples, so Tim has  $x - 30 = 68$  apples. So Harry has  $\frac{68}{2} = 34$  apples. The answer is 34.

**Our answer:** Let  $T$  be the number of apples Tim has. Martha has 68 apples. So Tim has  $68 - 30 = 38$  apples. Harry has half as many apples as Tim, so Harry has  $38 \div 2 = 19$  apples. The answer is 19.

**Reference answer:** Tim has  $68 - 30 = 38$  apples. Harry has  $38 \div 2 = 19$  apples. The answer is 19.

### GSM8k Case 3

**Question:** Marcy is a makeup artist and has agreed to do some makeup for her friend's wedding. The only makeup she has limited stock of is lip gloss so she counts how many tubes she needs. Each tube of lip gloss will hold enough lip gloss for 3 people's makeup. Marcy decides to bring 6 tubs of lip gloss, each of which holds 2 tubes of lip gloss, and this will be the exact amount she needs for everyone's makeup. How many people is Marcy painting with makeup?

**Mixed training answer:** First find the total number of tubes of lip gloss Marcy brought:  $6 \text{ tubs} * 2 \text{ tubes/tub} = 6 * 2 = 12 \text{ tubes}$ . Then divide the total number of tubes by the number of people each tube can paint:  $12 \div 3 = 4 \text{ people}$ . The answer is 4.

**Our answer:** Marcy has 6 tubs of lip gloss, each with 2 tubes of lip gloss, so she has  $6 * 2 = 12$  tubes of lip gloss. Each tube of lip gloss is enough for 3 people's makeup, so  $12 * 3 = 36$  people's makeup. The answer is 36.

**Reference answer:** Marcy is bringing 6 tubs of lip gloss \* 2 tubes of lip gloss per tub of lip gloss =  $6 * 2 = 12$  tubes of lip gloss. So she must be applying makeup to 12 tubes of lip gloss \* 3 people per tube of lip gloss =  $12 * 3 = 36$  people. The answer is 36.

## CodeAlpaca Case 1

### Question:

```
1 Create a program that prints the first 10 Fibonacci numbers.
```

### Mix training answer:

```
1 def fibonacci(n):
2     a = 0
3     b = 1
4     if n < 0:
5         print("Incorrect input")
6     elif n == 0:
7         return a
8     elif n == 1:
9         return b
10    else:
11        for i in range(2,n):
12            c = a + b
13            a = b
14            b = c
15        return b
16
17 if __name__ == '__main__':
18     print(fibonacci(10))
```

### Our answer:

```
1 def fibonacci(n):
2     a = 0
3     b = 1
4     if n < 0:
5         print("Incorrect input")
6     elif n == 0:
7         return a
8     elif n == 1:
9         return b
10    else:
11        for i in range(2,n):
12            c = a + b
13            a = b
14            b = c
15        return b
16
17 for i in range(10):
18     print(fibonacci(i))
```

### Reference answer:

```
1 """
2 Print the first ten Fibonacci numbers
3 """
4
5 def print_first_ten_fibonacci_numbers():
6     """Print the first ten Fibonacci numbers."""
7
8     n1, n2 = 0, 1
9     num = 0
10    while num < 10:
11        print(n1)
12        nth = n1 + n2
13        n1 = n2
14        n2 = nth
15        num += 1
16
17 if __name__ == '__main__':
18     print_first_ten_fibonacci_numbers()
```

## CodeAlpaca Case 2

### Question:

```
1 Write a function to calculate the arithmetic mean of an array of numbers
2 Input: arr = [5, 8, 2, 4, 7, 10]
3 Output:
```

### Mix training answer:

```
1 5.5
```

### Our answer:

```
1 def mean(arr):
2     sum = 0
3     for num in arr:
4         sum += num
5     return sum / len(arr)
```

### Reference answer:

```
1 def calculate_mean(arr):
2     """
3     Calculate the arithmetic mean of an array of numbers.
4     """
5     if len(arr) == 0:
6         return 0
7
8     return sum(arr) / len(arr)
```

## CodeAlpaca Case 3

### Question:

```
1 Edit the following function so that it counts the even numbers in the given
   array.
2
3 Input: arr = [1, 2, 3, 4]
4
5 Output:
```

### Mix training answer:

```
1 2
```

### Our answer:

```
1 def count_even_numbers(arr):
2     count = 0
3     for num in arr:
4         if num % 2 == 0:
5             count += 1
6     return count
```

### Reference answer:

```
1 def countEvens(arr):
2     count = 0
3     for num in arr:
4         if num % 2 == 0:
5             count += 1
6     return count
```