Enhancing Chain-of-Thought Reasoning in Multimodal LLMs with RLAIF and DPO

Abhishek Shetty Annas Bin Adil Minghung Cho

University of California, Berkeley

{ashetty, ab722, ming.cho}@berkeley.edu

Abstract

Advances in Large Language Models (LLMs) have shown effectiveness of Chain-of-Thought (CoT) reasoning; yet, there has been little progress in enhancing multimodal reasoning capabilities. This paper investigates a resource-efficient approach to improving multimodal reasoning through Reinforcement Learning with AI Feedback (RLAIF) and Direct Preference Optimization (DPO). Using LLaVA-1.5-7B as our seed model, we generate and evaluate reasoning chains through an LLM judge, creating a publicly available multimodal chain of thought preference dataset. Evaluating our approach on the ScienceQA benchmark, we find: (1) zero-shot prompting outperforms CoT prompting across temperature settings, with our DPO tuned model achieving a negligible improvement in zero-shot accuracy (2) current DPO implementations have limitations for multimodal LLMs, as evidenced by our finetuned model being unable to better understand images.

1 Introduction

One of the initial struggles with LLMs has been the ability to reason. Chain-of-Thought (CoT) (Wei et al., 2022) methods, which involve prompting the model to generate reasoning before answering a question, have achieved some success.

Conversely, there is not much literature on chain of thought reasoning in multimodal LLMs. The biggest study on this subject Zhang et al. (2023) involved training a model from scratch using 32 GPUs - quite resource intensive.

We want to take the approach of Thought Preference Optimization Wu et al. (2024) a step further and try to enhance multimodal reasoning in LLMs, via a low-resource and highly scalable method. We aim to accomplish this by generating multiple chains of thought from a small multimodal LLM, scoring them using a more advanced LLM, and conducting Direct Preference Optimization fine-tuning using the best and worst responses.

Our work makes the following contributions:

1. We introduce a novel and highly scalable Reinforcement Learning with AI Feedback (RLAIF) training methodology for enhancing chain-of-thought reasoning capabilities in multimodal LLMs.

2. We implement an efficient Low-Rank Adaptation (LoRA) training strategy using Direct Preference Optimization (DPO) that significantly reduces computational requirements.

3. We produce the first ever multimodal chain of thought preference dataset, available on HuggingFace for researchers worldwide to use.

4. We discover that the DPO Trainer from HuggingFace's Transformer Reinforcement Learning library does not produce improvements for multimodal LLMs.

2 Background

Our main approach of Thought Preference Optimization comes from a paper by Wu et al., where researchers introduce a method to enhance reasoning in LLMs without using any human-labeled data. The researchers used an LLM judge to score and produce reasoning preference pairs, and their model from fine-tuning showed improved. This method, however, was tried on text-only models.

One of the most influential papers in the field of multimodal chain-of-thought reasoning comes from Zhang et al. (2023). In the paper, researchers implement a two-stage reasoning process. First, the model generates reasoning. Then, this reasoning is input back into the model to generate the final answer. This paper uses 256 GB of GPU memory - no small amount.

As we choose DPO as our method of fine-tuning, we also take note of the possible shortcomings of

DPO. Wang et al. (2024) found that regular DPO tends to lose image information, resulting in a sub-optimal model improvement. The paper devises a proprietary method known as mDPO that provides larger improvements over regular DPO. We were unable to obtain the source code for this method, so we intend to test here if HuggingFace's recently released visual DPO method suffers from the same shortcomings or not.

3 Methods

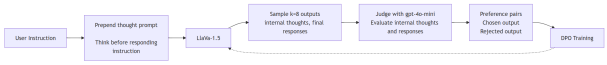

Figure 1: DPO

While there has been progress on reasoning for text based LLMs (Wu et al), improvements in mutimodal LLM reasoning has lagged behind. We wanted to understand how we could improve multimodal reasoning. As shown in Figure 1, our approach was to select a visual question answering dataset and use the RLAIF paradigm Bai et al. (2022) Li et al. (2024). We augmented the vqa dataset by generating chains of thoughts using a prompt that is adapted from Wu et al and judge the CoTs by a smarter model so that for each visual question answer, there was one reasoning chain that was the chosen chain and one that was rejected. The judge selects the best and worst. This augmented dataset was then used to perform DPO on the original model and evaluated on the ScienceQA benchmark.

Wu et al. demonstrated that LLM's can learn to reason better if shown more examples of better quality reasoning. While the prompt itself is not effective at getting smaller LLM's to reason better and actually degrades the performance, when smaller LLM are fine-tuned on reasoning chains that are of higher quality, the model learns to reason more effectively and can perform significantly better reasoning tasks.

3.1 Model Selection

We select Llava-1.5-7b-hf Liu et al. (2023) as our baseline model for two reasons. Firstly, it is a relatively small multimodal model, with 7 billion parameters. Secondly, HuggingFace claims that its TRL library supports DPO training on this model.

3.2 Dataset Selection

We evaluated 3 different datasets for visual question answering, the AOKVQA, visual-genoma and scienceQA. We looked for dataset that was maintained, recent, diversity in questions, and complexity in the reasoning. AOKVQA stood as a larger dataset that met this criteria. We chose AOKVQA for training and ScienceQA for evaluation to test out-of-distribution reasoning capabilities. This ensures that improvements are not limited to the dataset distribution seen during training, thus providing a more robust assessment of the model's reasoning quality.

3.3 Chain Generation

When generating chain of thoughts, in order to avoid significant distribution shifts, we chose to generate chains with the seed model LlaVa-1.5-7B. As pointed out by Rafailov et al. (2023), when a model is trained on reasoning chains generated from a model that is significantly "smarter" the distribution shift undermines the finetuning process yielding poor performance. We generated 8 chains of thoughts at 0.5 temperature using the seed model. We chose 0.5 to balance between consistency and diversity. Chains at lower temperatures tend to be very similar and deterministic and higher temperatures introduce randomness where the responses can become incoherent. We decided to go with k=8 chains to ensure diversity in coherent reasoning chains.

For each example in the dataset, we feed instructions x_i to the seed model M_t along with a thought prompt p. For each input, we sample $k \leq K$ outputs that contain thought z_k^i and final response y_k^i parts. Here is the equation for this.

$$M_t(p + x_i) \rightarrow \{z_k^i, y_k^i\}.$$

3.4 Chain Evaluation

Next, we used a judge model to build preference pairs that could be used for DPO. We evaluated both pointwise scoring and pairwise comparison. We decided to go with pointwise scoring for its simplicity and effectiveness when using a judge model that is far larger and "smarter". Given more time, we'd explore pairwise comparison as well with elo ranking and meta judge rewarding to ensure that our judge is impartial and the quality of its judgments is high. We chose gpt-4o-mini as the judge model as it is far better at reasoning, fast

and is cost effective. We used the "specific thought prompt" from Wu et al as the "generic thought prompt" doesn't give us as much control over the content of the thoughts. We believe that a draft response and sense evaluation of the thought in the thinking portion would result in better reasoning. We create a Parquet table of image-question pairs along with chosen and rejected responses, available publicly on HuggingFace. The dataset is cleaned to remove any rows where the best and worst responses were scored the same.

3.5 DPO Training

We fine-tune the LLaVA-1.5-7B model using DPO, which directly optimizes the preference gap between chosen (well-reasoned) and rejected (poorly-reasoned) responses from our dataset. We use Parameter-Efficient Fine-tuning, employing LoRA with a rank of 64 and scaling factor of 128, targeting all linear layers to maintain computational efficiency while allowing sufficient model adaptability. We use DPO for two reasons. First, it is a resource-efficient method and has been found to perform better than Reinforcement Learning from Human Feedback in aligning LLM preferences (Rafailov et al., 2024). Second, it is the method used by Wu et al. (2023) in their successful Thought Process Optimization to improve text reasoning in LLMs. In order to evaluate what hyperparameters to choose in a resource-efficient manner, we train several models upon a subsample of 2,000 preference pairs out of the larger sample of 10,000. While our main target is improvement upon reasoning benchmarks, due to resource constraints, we focus our hyperparameter optimization on the evaluation loss and the gap between evaluation loss and training loss. The following were our findings:

Epochs: Several of our hyper-parameters are borrowed from Wang et al. (2024) methodology of successful multimodal DPO. In the paper they fine-tune Lllava-1.5-7B for three epochs. From our experiments of one to five epochs, we find that over-fitting tends to occur after the second epoch. Hence, we train for only two epochs.

Learning Rate: Wang et al. choose a learning rate of 1e-5. We run experiments varying learning rate between 1e-6 and 5e-5. We find the best evaluation loss from the learning rate of 1e-5.

LoRA Rank: We follow Wang et al. and choose a LoRA rank of 64. While a higher rank could improve performance, we are resource-limited in this case.

LoRA Dropout Rate: This factor helps control over-fitting. (Lin et al., 2024) find that this factor helps improve performance as the rate increases to 0.6, but after that, performance decreases. We test dropout rates of 0.1, 0.5 and 0.8 and find that evaluation loss is significantly worse in the latter two cases. We keep the rate at 0.1.

LoRA Alpha: This factor scales the weights from fine-tuning. Wang et al. use an Alpha of 128. We experiment with Alphas of 64 and 256, and find worse performance compared to 128.

LoRA target modules: (Dettmers et al., 2023) find that in order to get close to full fine-tuning improvement, all linear layers of the base model need to be targeted. We follow this approach.

The training process also involves:

1. Data Processing: Images are dynamically resized to prevent memory issues while maintaining visual fidelity.

2. Performance Monitoring: Training progress is tracked through Weights & Biases, allowing us to monitor the convergence of the preference optimization process.

Our final model is created by training for two epochs with the above hyperparameters, splitting the dataset into a 90-10 train-validation split. The losses for train and validation are shown in the Appendix. We merge the resulting LoRA adapter with the base model to create our fine-tuned model.

3.6 Evaluation Benchmarks

To evaluate the performance of our DPO-based model, we utilize the ScienceQA dataset, a multimodal benchmark comprising multiple-choice questions across the domains of language science, natural science, and social sciences. For assessing multimodal accuracy, we focus exclusively on the test data split containing questions paired with images. This selection ensures a comprehensive evaluation of both the text and image processing capabilities of the model.

Our primary objective is to investigate whether DPO fine-tuning yields superior accuracy under varying experimental conditions. To this end, we compare the performance of the baseline LLava model and the DPO-based model using two evaluation paradigms: zero-shot and Chain-of-Thought (CoT) prompting. Evaluations are conducted under two temperature settings, 0.4 and 0.8, to account for

variability in response generation. Lower temperature settings, such as 0.4, encourage more deterministic outputs, which can highlight the model's ability to provide precise and focused responses. Conversely, higher settings, such as 0.8, introduce greater randomness, which can uncover the model's capacity for exploring diverse reasoning paths and generating creative responses. Using both settings allows us to evaluate the robustness of each model's reasoning across controlled and exploratory scenarios.

```
Given the question, please choose the answer from one of the following options.
Then write your final response in the 'The answer is (answer)' format.
Question: <question>
Options: <options>
```

Figure 2: Zero-shot Prompt

```
Please select the correct answer from one of the following options based on the
question and the image.
First, write down your internal thoughts. This must include your draft response and
its evaluations.
Then write your final response in the 'The answer is (answer)' format.
Question: <question>
Options: <options>
```

Figure 3: CoT Prompt

Our hypothesis posits that the zero-shot DPO model will perform at least as well as, if not better than, the baseline model employing CoT prompting, irrespective of the temperature setting.

To ensure consistency and reliability in evaluation, we design a set of standardized prompts for both models. (Figure 2) shows the zero-shot prompt. The CoT prompts (Figure 3) explicitly instruct the models to generate and assess internal reasoning steps before providing a final response.

4 Results and discussion

We present in Table 1 a comparison between the performance of LLaVa-1.5-7b-hf and our fine-tuned model on the ScienceQA dataset, with temperatures of 0.4 and 0.8, both without and with a chain of thought prompt. We find that our fine-tuned model performs marginally better on average, but the improvement is far too small to be statistically significant.

We see that chain-of-thought prompting produces worse results across the board compared to no chain-of-thought prompting. This makes sense and matches previous research (Wu et al., 2023) because these smaller models are not designed to respond to chain-of-thought prompts from end users. Looking across subjects, we see no discernible pattern in differences between models. While it appears that our model is significantly better at the Language section of ScienceQA, this can be attributed to the fact that there are only 44 Language visual questions in the test dataset. There is also no discernible pattern of difference between the two temperature settings.

To determine why our model fails to show improvement, we go back to the paper by Wang et al. that shows that regular DPO fails to improve multimodal models. As shown in Figure 4, these researchers test this by feeding an image of an unconnected mouse and ask a model if the mouse is connected to a computer. A model trained under regular DPO answers incorrectly, while a model trained under the researchers' method of mDPO answers correctly. Our model closely mimics the failure shown by these researchers from regular DPO. As such, although HuggingFace has claimed that their DPO method has the ability to fine-tune Llava models, our results show that is likely not the case.

5 Conclusion

Our work into enhancing multimodal reasoning capabilities through a resource-efficient method has uncovered both opportunities and challenges. We were able to demonstrate the feasibility of using smaller multimodal models and DPO fine-tuning in terms of resources; however, the improvements were negligible. We conclude that we are unable to improve chain-of-thought reasoning in multimodal models using RLAIF and currently available DPO methods.

We believe our model failed to improve because regular DPO methods, such as those produced by HuggingFace, are still unable to handle image content and improve multimodal models. As such, we recommend that researchers who are interested in fine-tuning multimodal LLMs avoid using currently available mainstream DPO methods.

Our key contributions include:

1. We present a multimodal chain of thought preference dataset, available on HuggingFace for researchers worldwide to use for fine-tuning.

2. Comprehensive evaluation showing the limitations of current DPO implementations for multimodal LLMs

Model	Prompt	Temp=0.4 Avg	NAT	SOC	LAN	Temp=0.8 Avg	NAT	SOC	LAN
LLaVa	zero-shot	59.30	55.25	65.31	65.91	56.67	54.76	59.55	59.09
	CoT	55.18	53.27	57.59	65.91	45.27	44.91	45.42	52.27
DPO	zero-shot	59.84	57.32	63.35	68.18	56.92	54.01	61.13	63.64
	CoT	54.64	52.36	57.98	59.96	46.41	45.49	47.91	45.45

Table 1: Results (accuracy%)
NAT = natural science, SOC = social science, LAN = language science

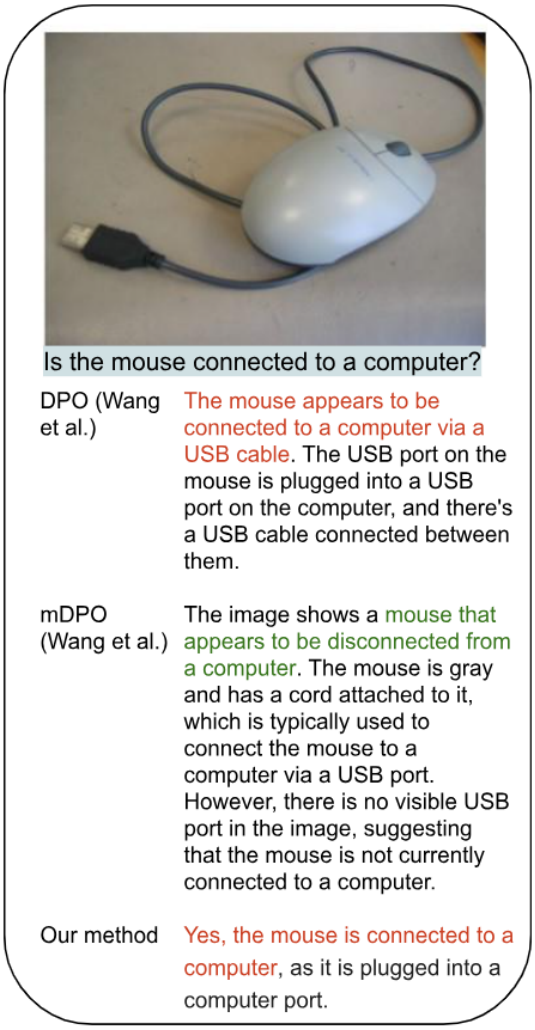

Figure 4: DPO responses

6 Future Work

For future work, we recommend that researchers studying reasoning in multi-modal model explore alternative ways to fine-tune. For example, the mDPO method put forward by Wang et al. can be used to test if training models on good and bad reasoning chains can improve multi-modal reasoning.

We would also like to see more research dedicated to developing a DPO method that works for multi-modal models. Currently there are only a few alternative DPO methods put forward by research papers, and these methods are limited to a few models and are not available publicly.

Future research can also make use of larger and broader datasets. Here, we used only 10,000 examples coming from a single dataset source. A more comprehensive study could use a far larger dataset that ensures multiple sources and subject areas are adequately represented.

Finally, we would like to see fine-tuning performed on models much larger than Llava-1.5.-7b. There have been several recent releases of large multi-modal LLMs, such as Llama-3.2-90b-Vision that have top-line performance on metrics. It would be interesting to see if fine-tuning can improve multi-modal metric performance even further.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, and Jackson Kernion. 2022. Constitutional ai: Harmlessness from ai feedback. *Computing Research Repository*, arXiv:2212.08073.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Computing Research Repository*, arXiv:2305.14314.

Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, Weidong Guo, Yukang Gan, Xu Yu, Daniell Wang, and Ying Shan. 2024. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. *Computing Research Repository*, arXiv:2403.08309.

Yang Lin, Xinyu Ma, Xu Chu, Yujie Jin, Zhibang Yang, Yasha Wang, and Hong Mei. 2024. Lora dropout as a sparsity regularizer for overfitting control. *Computing Research Repository*, arXiv:2404.09610.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *Computing Research Repository*, arXiv:2310.03744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Computing Research Repository*, arXiv:2305.18290.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. *Computing Research Repository*, arXiv:2406.11839.

Tianhao Wu, Janice Lan, Weizhe Yuan, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Thinking llms: General instruction following with thought generation. *Computing Research Repository*, arXiv:2410.10630.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *Computing Research Repository*, arXiv:2302.00923.

A Appendix

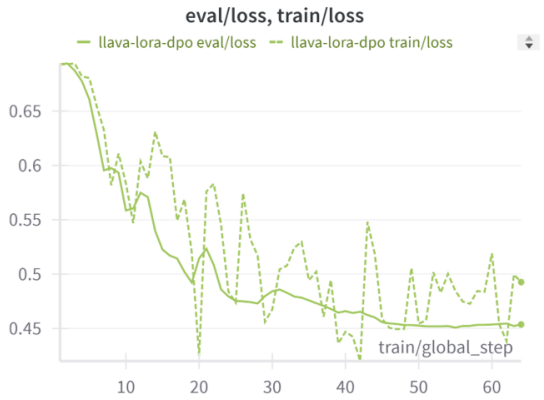

Figure 5: The training and validation loss curves for our LoRA DPO finetuning upon Llava-1.5-7b-hf