

On the Suitability of Reinforcement Fine-Tuning to Visual Tasks

Xiaxu Chen^{1,2*} Wei Li^{1†} Chunxu Liu^{1,3} Chi Xie^{1,4}

Xiaoyan Hu¹ Chengqian Ma¹ Feng Zhu¹ Rui Zhao^{1‡}

¹SenseTime Research ²Beijing Institute of Technology ³Nanjing University ⁴Tongji University

Abstract

Reinforcement Fine-Tuning (RFT) is proved to be greatly valuable for enhancing the reasoning ability of LLMs. Researchers have been starting to apply RFT to MLLMs, hoping it will also enhance the capabilities of visual understanding. However, these works are at a very early stage and have not examined how suitable RFT actually is for visual tasks. In this work, we endeavor to understand the suitabilities and limitations of RFT for visual tasks, through experimental analysis and observations. We start by quantitative comparisons on various tasks, which shows RFT is generally better than SFT on visual tasks. To check whether such advantages are brought up by the reasoning process, we design a new reward that encourages the model to “think” more, whose results show more thinking can be beneficial for complicated tasks but harmful for simple tasks. We hope this study can provide more insight for the rapid advancements on this topic.

1. Introduction

Recently, Reinforcement Fine-Tuning (RFT) has demonstrated remarkable effectiveness on Large Language Models (LLMs) such as DeepSeek-R1 [4]. By incentivizing the model to engage in more extensive “thinking” during training and inference, RFT significantly enhances its reasoning capabilities for addressing complex language tasks. Relevant techniques include Reinforcement Learning with Human Feedbacks (RLHF) and Reinforcement Learning with Verifiable Rewards (RLVR), which utilizes human preferences or objectively verifiable outcomes as rewards for reinforcement learning.

A natural question emerges: can RFT similarly augment Multimodal Large Language Models (MLLMs), particularly in the realm of visual reasoning? Recent studies [5, 12, 14, 24] have investigated the application of RFT to MLLMs, achieving superior performance on tasks that

explicitly demand robust reasoning skills. These efforts have underscored RFT’s strengths in Few-Shot Classification, Object Detection, and Reasoning Grounding, surpassing the capabilities of Supervised Fine-Tuning (SFT). Nevertheless, the extent of RFT’s applicability to visual tasks remains largely unexplored.

In this study, we examine the impact of RFT on MLLMs, contrasting it with prior approaches such as MLLMs trained with SFT. We begin by implementing RFT on MLLMs and evaluating their performance against SFT across various computer vision tasks from perception classification tasks to those need visual reasoning. Notably, RFT consistently delivers substantial improvements on specific tasks, often outperforming SFT by a wide margin.

We then explore whether the performance advantage of RFT over SFT stems from improved reasoning. To investigate this, we introduce a **Normalized Length Reward** in the RFT framework, encouraging the model to produce lengthier intermediate outputs and engage in prolonged “thinking”. This adjustment enhance the performance on complicated tasks requiring explicit reasoning but decrease it on perception classification tasks, suggesting that the gains are partially attributable to enhancing model’s structured reasoning capabilities from RFT. Besides, disabling the thinking process during inference consistently impairs MLLM performance. We therefore conclude that *current computer vision tasks demands different degrees of reasoning according to their task nature*, and insights gained from RFT on LLMs cannot be directly applied to visual domains.

Our contributions are summarized as follows:

1. We demonstrate that Reinforcement Fine-Tuning (RFT) outperforms SFT across a range of computer vision tasks, from basic classification to those requiring visual reasoning.
2. By encouraging MLLMs to think longer using the Normalized Length Reward in RFT, MLLMs obtain reasonable thinking process and stronger performance on some complicated tasks that require explicit reasoning.
3. We find that encouraging longer thinking process can be harmful on some simple visual tasks, which shows RFT from LLMs require more improvements before applied

*Email: 3220230717@bit.edu.cn

†Email: liwei1@sensetime.com

‡Corresponding author. Email: zhaorui@sensetime.com

to visual tasks.

2. Related Works

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) [1, 2, 6, 9] integrate visual encoders with Large Language Models (LLMs) to enable visual perception and achieve remarkable performance on multimodal tasks [3, 8, 11]. A typical MLLM consists of a vision encoder, an LLM, and a visual projector that maps visual tokens into the semantic space of LLM. Leveraging this architecture, MLLMs have been applied to a wide range of vision and language tasks, including image classification [15, 16], object segmentation [7], object detection [18], information retrieval [10], and visual question answering [1, 2, 6, 9]. In this paper, we further explore the versatility of Reinforcement Fine-Tuning (RFT) in enhancing the test-time scaling ability of MLLMs on various computer vision tasks.

Reasoning in MLLMs. Since the recent surge of reasoning in Large Language Models like Openai-o1 and DeepSeek-R1 [4], the community has been trying to achieve a similar reasoning process for MLLMs. They utilize Reinforcement Learning with Verifiable Rewards, a training approach to enhance language models in tasks with objectively verifiable outcomes. Exploration in this area is still at a very early stage and remains highly immature. Among them, R1-V explores how to transplant R1 directly to MLLMs. VisualThinker-R1-Zero [24] claims to be the first to produce “aha moment” and increased response length for MLLMs, by performing RFT on base models without instruction tuning. Visual-RFT [12] finds that reinforcement fine-tuning is more powerful than supervised fine-tuning on a wide range of tasks like few-shot classification and detection. Such concurrent works all try to transplant R1 from LLMs to MLLMs and prove how powerful R1 is. Differently, this work tries to examine where RFT is suitable and where not on traditional computer vision tasks.

3. Method

Reinforcement Learning in Large Models. Reinforcement Learning with Verifiable Rewards [4, 21, 22] is a training paradigm aimed at improving language models in domains where correctness can be objectively verified, such as mathematics and programming. In contrast to Reinforcement Learning from Human Feedback (RLHF) [17], which depends on a learned reward model, RLVR evaluates outputs using a direct verification function. This eliminates the need for an intermediary reward model while ensuring that the optimization process remains closely tied to the intrinsic correctness measures of tasks. Given an input question q , the policy model π_θ generates a response o and receives a verifiable reward accordingly. The training objective op-

timized in RLVR is expressed as follows:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_\theta(q)} [R_{\text{RLVR}}(q, o)] \quad (1)$$

$$= [R(q, o) - \beta \text{KL}[\pi_\theta(o|q) \parallel \pi_{\text{ref}}(o|q)]], \quad (2)$$

where π_{ref} represents the pre-optimization reference model, R denotes the reward function used for verification, and β is a hyperparameter that regulates the KL-divergence. The reward function R evaluates a given question-output pair (q, o) by determining whether the predicted response o aligns with the ground-truth answer.

Group Relative Policy Optimization. DeepSeek R1-Zero [4] removes the need for supervised fine-tuning by adopting reinforcement learning for post-training. Specifically, it utilizes Group Relative Policy Optimization (GRPO) [21], which differs from conventional reinforcement learning methods like PPO [20] that rely on a critic model for policy evaluation. Instead of using a separate critic, GRPO directly assesses and ranks multiple candidate responses. Given an input query q , the model generates G potential answers o_1, o_2, \dots, o_G based on its current policy $\pi_{\theta_{\text{old}}}$. The rewards associated with these responses, denoted as r_1, r_2, \dots, r_G , are then computed. GRPO normalizes these rewards by calculating their mean and standard deviation:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\})}, \quad (3)$$

where A_i indicates the relative quality of the i -th response. This approach guides the model toward generating higher-quality responses by prioritizing those with relatively superior rewards.

Verifiable Rewards for Vision Tasks. The reward model is a key step in reinforcement learning (RL) that aligns models with preference alignment algorithms, which can be as straightforward as a verification function that checks for exact matches between predictions and ground-truth answers. The RL training process in the recent DeepSeek-R1 [4] model achieves a significant improvement in the model’s reasoning ability through the verifiable reward design. To transfer this strategy to the visual domain, we design different rule-based verifiable reward functions for various visual perception tasks.

Normalized Length Reward. In VisualThinker [24], researchers observed that using a fixed-length reward of +0.001 per additional generated token led to an increase in response length without improving accuracy. The model exploited the reward by generating repetitive content rather than enhancing its reasoning process. To address this issue, we apply sigmoid normalization to the length reward, stabilizing its value between 0 and 1, as follows:

$$R = \frac{1}{1 + \lambda \cdot \exp(-(L - L_0))}, \quad (4)$$

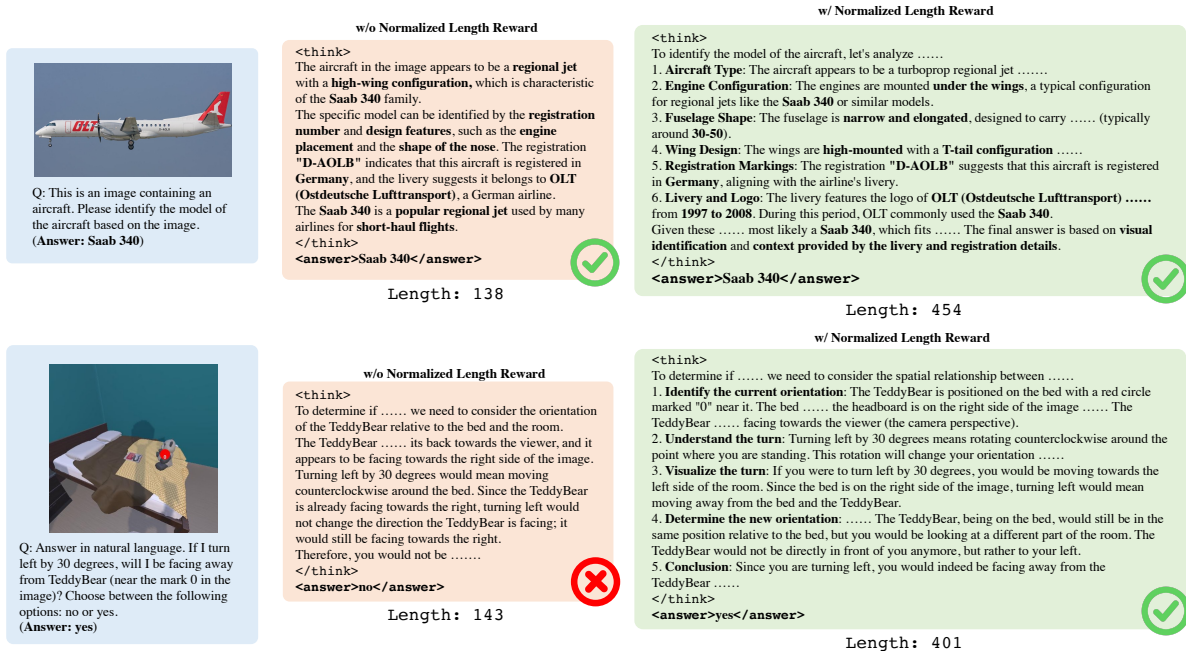


Figure 1. Qualitative Comparison Between w/o and w/ Normalized Length Reward. Best view in zoom.

where λ is a scaling factor set to 1 in our experiments, and L_0 is a predefined length fixed at 100 across all experiments. Therefore, the length reward is kept within a reasonable range, as the diminishing marginal benefits help stabilize response length effectively.

4. Experiments and Findings

4.1. Experimental Setups

Based on Qwen2.5VL-Instruct-7B [1], we conducted all RFT experiments on eight NVIDIA A100 80GB GPUs. The batch size is set to 1 per device, with gradient accumulation in 4 steps. For model generation, the temperature is set to 1, and the KL coefficient is fixed at 0.04. To investigate the impact of response length on model performance, we set the maximum response length at 1,024 tokens. We retain the built-in system prompt and incorporate format-related questions in the user prompt for GRPO training, where each sample generates 8 rollouts. The model is trained for 300 steps with a learning rate of $1e-6$. For supervised fine-tuning, we adopt Qwen2.5VL-Instruct-7B and train it for 300 steps to ensure a fair comparison with RFT, using Llama-Factory framework [23].

4.2. Quantitative Comparison over CV Tasks

We evaluate our model on three benchmarks: Banner Analysis, FGVC Aircraft [13], and SAT [19]. Banner Analysis is a dataset that we collected for sentiment classification. It serves as a benchmark for detecting negative content in ban-

ners and slogans. The dataset comprises real-world images of various banners and contains 331 images for training, specifically curated for this task. As a binary classification benchmark, it takes a banner image as input and outputs either "positive" or "negative", indicating whether the banner contains positive or negative content. FGVC Aircraft [13] is a fine-grained classification task that involves identifying airplanes across 100 distinct classes, totally 6k training images. SAT [19] is a spatial aptitude training dataset designed to challenge users with complex, dynamic spatial tasks that go beyond the static relationships found in traditional datasets. 15k data samples are used for training. These three benchmarks increase in difficulty, with a progressively higher demand for reasoning ability.

In Table 1, we compare the effects of Supervised Fine-Tuning (SFT) and Reinforcement Fine-Tuning (RFT) on Qwen2.5VL-Instruct-7B [1]. Both SFT and RFT models are trained on the corresponding datasets for 300 steps for fair comparison. We investigate the impact of explicitly guiding MLLMs to engage in reasoning during test by requiring three paradigms to generate reasoning process within `<think>` and `</think>` tokens.

The results demonstrate that incorporating `<think>` tokens in prompts leads to underperformance in both training-free and SFT models, revealing that extended intermediate reasoning steps may induce overthinking. In contrast, RFT models, while direct generating answer reduces accuracy, explicit reasoning steps substantially enhance performance. This paradigm shift implies that RFT training could

Table 1. **Performance Comparison Between SFT and RFT.** The three benchmarks listed has an increasing demand on reasoning ability. Banner Analysis is our constructed real-world benchmark for recognizing malicious content.

Benchmarks	Qwen2.5VL-Instruct-7B					
	Training Free		SFT		RFT	
	w/o <think>	w/ <think>	w/o <think>	w/ <think>	w/o <think>	w/ <think>
Banner-Analysis	80.72	80.11	90.74	90.21	97.22	97.83
FGVC-Aircraft	55.48	47.55	75.85	75.84	63.57	83.32
SAT	56.85	49.84	58.24	59.62	61.01	62.23

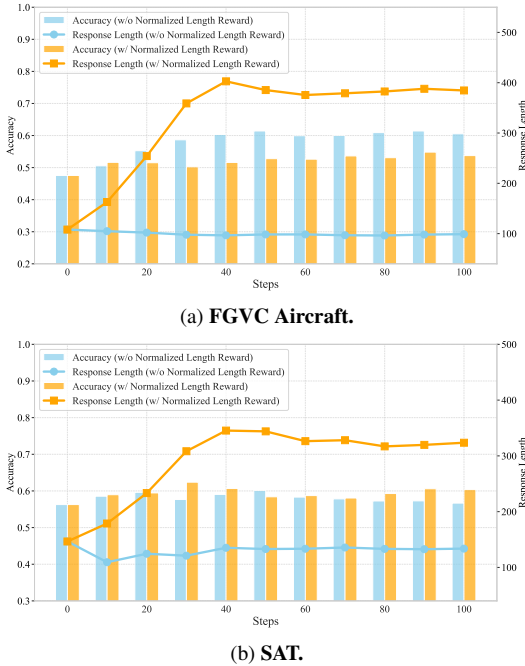


Figure 2. Accuracy and Response Length over Steps.

enhance the model’s structured reasoning capabilities.

These findings suggest that RFT is a more effective approach than SFT for these CV benchmarks. Additionally, explicit reasoning proves beneficial for complex benchmarks but offers limited advantages for simpler tasks.

4.3. Performance and Response Length

In DeepSeek-R1 [4], researchers identified an “aha moment” in Large Language Models (LLMs) from the model performance on text-only benchmarks: as response length increases, accuracy improves, suggesting that the model exhibits a capacity for self-reflection. However, it raises the question of whether computer vision (CV) task necessitates an extensive reasoning process. For example, most CV classification tasks primarily evaluate perceptual capabilities. And such tasks may not invariably benefit from prolonged explicit reasoning. Just as humans classify objects by recognizing salient features without elaborate reasoning, MLLMs may also gain little benefit from an extended reasoning pro-

cess.

We incorporated a normalized length reward into the Reinforcement Fine-Tuning (RFT) process to investigate this hypothesis. In Figure 2, subfigures (a) illustrate a relatively straightforward visual task, FGVC Airplane classification that only involves object classification based on the images provided. For such task, where visual perception is paramount, an increase in response length detrimentally affects performance compared to experiments conducted without the length reward. In contrast, within the SAT benchmark in Figure 2b, where reasoning is required after object recognition, longer responses correlate with improved performance.

Although the qualitative analysis presented in Figure 1 indicates that the application of the normalized length reward improves the reasoning process and logical coherence, the benchmark accuracy results reveal that the impact of increased response length on performance varies as a function of the difficulty and complexity of the task.

We found that encouraging Multimodal Large Language Models (MLLMs) to think longer during the Reinforcement Fine-Tuning (RFT) process does not lead to significant performance improvements. This suggests that current reasoning approaches are of limited help for traditional vision tasks. The reason may be that traditional Computer Vision (CV) tasks focus more on perceptual capabilities rather than complex reasoning processes.

5. Conclusions

In this work, we perform experimental examination over the suitabilities and limitations of Reinforcement Fine-Tuning in visual tasks. Through quantitative comparisons, ablation studies and qualitative cases, we find that RFT generally works better than SFT for MLLMs over traditional CV tasks. Furthermore, we find that though the thinking process from RFT is essential for MLLMs, encouraging them to think longer and producing more intermediate results is not always helpful, possibly due to the nature of these tasks for less “reasoning” than “recognition”. We hope this study will serve as a pilot study for the MLLM community where RL is only at the stage of early exploration.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3
- [2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2
- [3] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2
- [4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 2, 4
- [5] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1
- [6] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2
- [7] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2
- [8] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 2
- [9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [10] Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. *arXiv preprint arXiv:2412.01720*, 2024. 2
- [11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2
- [12] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rlt: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 1, 2
- [13] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3
- [14] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1
- [15] Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. *arXiv preprint arXiv:2412.00142*, 2024. 2
- [16] Yassine Ouali, Adrian Bulat, Alexandros Xenos, Anestis Zaganidis, Ioannis Maniadis Metaxas, Brais Martinez, and Georgios Tzimiropoulos. Discriminative fine-tuning of lvlms. *arXiv preprint arXiv:2412.04378*, 2024. 2
- [17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2
- [18] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023. 2
- [19] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 3
- [20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2
- [21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2
- [22] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 2
- [23] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3
- [24] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero’s” aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025. 1, 2