Fine-Tuning LLaMA for Academic Reasoning Evaluation

Simon Green[1]

School of Computing, University of Leeds, UK
[1] MSc, Artificial Intelligence
{od21sg}@leeds.ac.uk

Abstract. [Abstract Placeholder]

Keywords: GRPO, LLaMA, QLoRA, Fine-Tuning, Reinforcement Learning, HLE, Evaluation Metrics, Policy Optimization

1 Introduction

The rapid evolution of large language models has significantly advanced natural language processing applications. However, the substantial computational resources required for fine-tuning these models remain a critical challenge. Recent approaches, such as Quantized Low-Rank Adaptation [1], demonstrate that quantization and low-rank adaptation can effectively reduce memory consumption during fine-tuning. Concurrently, reinforcement learning techniques are gaining traction. For instance, the DeepSeek R1 release introduced the Grouped Relative Policy Optimization algorithm, which optimizes policy performance without relying on explicit value networks [8, 7].

Large language models can be viewed as highly efficient forms of lossy compression. As such, they excel at solving problems present in their training data—for example, in mathematics or physics—by leveraging patterns rather than engaging in human-like creative reasoning and developing new explatory knowledge, despite what commercial model developers might suggest [2]. Their architectures—whether autoregressive, predicting the next token through conditional probabilities [9], bidirectional, estimating probabilities for masked tokens, or non-autoregressive, modeling joint distributions over full sequences—rely on probabilistic token prediction, rooted in statistical associations rather than inventive thought [3, 5]. This limitation is evident in works like Phan et al., where top-tier models reportedly achieved no more than 14% accuracy on Humanity's Last Exam, a dataset designed to test creativity and intelligence with novel questions [6]. Such poor performance underscores the need to explore whether fine-tuning on datasets like HLE can push these models beyond mere interpolation of learned distributions toward genuine creative reasoning.

In this study, we investigate the impact of fine-tuning a model on the HLE dataset to determine whether exposure to similar, previously unseen problems

can enhance its intelligence. We integrate the Grouped Relative Policy Optimization algorithm into the fine-tuning pipeline of the LLaMA model, aiming to improve both efficiency and academic reasoning performance. Our approach is evaluated using the Humanity's Last Exam dataset [6], a comprehensive benchmark comprising over 2,700 rigorously developed academic questions. By combining a unique mix of algorithms and datasets, this work seeks to offer a novel perspective on fine-tuning large language models. We compare the performance of our fine-tuned LLaMA model against the DeepSeek R1 results and the baseline LLaMA model.

We hypothesize that fine-tuning the LLaMA Vision model on a single graphics processing unit, using quantized low-rank adaptation and group-relative policy optimization, will enhance its performance on unseen mathematical problems within the test subset of the Humanity's Last Exam dataset, which the model has not previously encountered [6]. Furthermore, we propose that this combined approach—leveraging quantized low-rank adaptation for memory efficiency and group-relative policy optimization as a reinforcement learning strategy—offers a sound method to achieve improved results compared to the baseline LLaMA model [1, 8]. Given the absence of preliminary results in this proposal, we focus on establishing the potential for enhanced reasoning through these techniques rather than specifying a numerical improvement threshold.

2 Background and Literature Review

Recent advancements in large language model optimization have introduced innovative methods that enhance performance on complex datasets. Studies such as those by Shao et al. (2024) and Phan et al. (2025) have gained attention for their contributions, including Group-Relative Policy Optimization and the Humanity's Last Exam dataset, respectively, accumulating citations and public implementations shortly after release [8, 6]. A notable technique, quantized low-rank adaptation, enables efficient fine-tuning of large language models on consumer-grade hardware by quantizing model weights and employing low-rank adapters. This quantization process is mathematically defined as:

$$Q(\alpha)\& = \left\lfloor \frac{\alpha}{\delta} \right\rfloor, D(\kappa)\& = \kappa \cdot \delta, \epsilon_q\& = \alpha - D\Big(Q(\alpha)\Big) = \alpha - \left\lfloor \frac{\alpha}{\delta} \right\rfloor \cdot \delta. \quad (1)$$

where a floating-point value α is quantized to the nearest multiple of a fixed precision δ. Low-rank adaptation minimizes memory usage by training a small set of adapter parameters while keeping most model weights fixed, with the updated projection expressed as:

$$\boldsymbol{\Phi}\,\boldsymbol{\Omega}\& = \boldsymbol{\Upsilon}, \boldsymbol{\Upsilon}\& = \boldsymbol{\Phi}\,\boldsymbol{\Omega} + \sigma\,\boldsymbol{\Phi}\,\Lambda_1\,\Lambda_2, \quad (2)$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{\mu \times \eta}$, $\boldsymbol{\Omega} \in \mathbb{R}^{\eta \times \omega}$, $\Lambda_1 \in \mathbb{R}^{\eta \times \rho}$, $\Lambda_2 \in \mathbb{R}^{\rho \times \omega}$, and σ is a scalar. This combination of 4-bit quantization and low-rank adapters reduces the memory footprint, allowing models with tens of billions of parameters to be fine-tuned

on a single graphics processing unit with modest memory, recovering full 16-bit performance. Dettmers et al. (2023) demonstrated this with a 65-billion-parameter model fine-tuned on a 48-gigabyte graphics processing unit [1].

Another advancement, Group-Relative Policy Optimization, extends the proximal policy optimization framework by using group-relative rewards instead of an independent value network [7]. The advantage of an action is calculated by comparing its reward to the average of peer actions within a group:

$$\Delta\rho_t = \rho_t - \frac{1}{|\Gamma_t|} \sum_{\varphi \in \Gamma_t} \rho_\varphi, \tag{3}$$

$$\hat{A}_t = \sum_{k=0}^{\infty} \gamma^k \, \Delta\rho_{t+k}, \tag{4}$$

$$L^{\mathrm{GRPO}}(\vartheta) = \hat{\mathbb{E}}_t \left[\min\Big(r_t(\vartheta)\, \hat{A}_t,\ \mathrm{clip}\Big(r_t(\vartheta), 1-\epsilon, 1+\epsilon \Big)\, \hat{A}_t \Big) \right]. \tag{5}$$

This approach simplifies the model, reduces estimation errors, and improves computational efficiency during fine-tuning [8]. The Humanity's Last Exam dataset, a multimodal benchmark with 2,700 challenging questions across mathematics, physics, biology, and natural sciences, was developed through a rigorous multi-stage review by subject matter experts [6]. Designed to resist simple internet retrieval, it evaluates genuine reasoning and problem-solving abilities, serving as a robust academic benchmark for large language models.

Influential work by Shao et al. (2024) on DeepSeekMath integrates these methods, demonstrating that a 7-billion-parameter model, optimized with high-quality mathematical data, can rival larger models like Minerva (540 billion parameters), while Group-Relative Policy Optimization improved MATH benchmark scores from 46.8% to 51.7% [8]. These findings underscore the efficacy of parameter-efficient fine-tuning and reward-based optimization, providing a foundation for our study.

3 Methodology

This study fine-tunes a large language model with vision capabilities to enhance mathematical reasoning, utilizing the Humanity's Last Exam dataset from the Center for AI Safety [6]. We employ quantized low-rank adaptation and Group-Relative Policy Optimization, as introduced in the background, to optimize the LLaMA-3.2-11B-Vision-Instruct model on a single graphics processing unit.

3.1 Dataset

The Humanity's Last Exam dataset provides a diverse corpus of 2,700 question-answer pairs across multiple domains, with our focus on the mathematics category (1,106 examples) to evaluate reasoning capabilities. This dataset includes

both text-based and image-augmented questions, necessitating a multimodal model. Text token lengths in the mathematics category range from 15 to 13,518 (mean 216.81, standard deviation 483.85), while 4.07% of questions include images averaging 0.145 megabytes (standard deviation 0.211 megabytes), requiring vision processing despite the predominantly textual nature of the tasks. Its complexity aligns with benchmarks like MATH, which informed high-performing models through extensive data curation [4, 8].

We prepared the dataset by stratifying and splitting it to ensure balanced representation, particularly for mathematics. Using a 90% confidence interval and 5% error margin, the mathematics category was divided into 80.29% training (888 samples) and 19.71% testing (218 samples), prioritizing a robust training set. Of the test questions, 4.13% contain images, and 91.28% feature exact-match answers, emphasizing precision. Figure 1 illustrates the question distribution, with mathematics dominating at 1,106 examples.

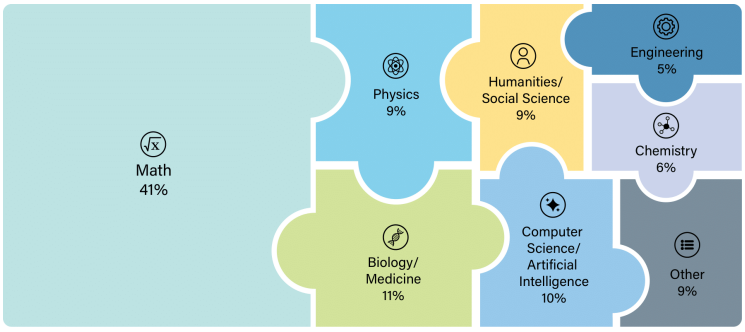

Fig. 1. Distribution of Questions in the Humanity's Last Exam Dataset [6]

3.2 Fine-Tuning

We selected the LLaMA Vision model for its ability to process both text and images, critical for the 4.13% of mathematics questions with visual elements. With 11 billion parameters, it balances computational feasibility and performance, offering an open-source, resource-efficient solution for single graphics processing unit fine-tuning, unlike text-only models. Its vision-language integration aligns with trends in multimodal artificial intelligence.

Fine-tuning employs quantized low-rank adaptation, using 4-bit quantization with a normal float type and brain float 16 compute precision to reduce memory usage while preserving performance akin to full-precision models. Low-rank adapters, configured with a rank of 4, alpha of 32, and 5% dropout, target query and value projection modules. Inputs are tokenized to a maximum length of 32 tokens, padded, and truncated as needed. To enhance mathematical reasoning, we integrate Group-Relative Policy Optimization, which eliminates the

critic model and estimates baselines from group scores of multiple outputs (two completions per prompt). Rewards are computed via cosine similarity between generated completions and ground-truth answers, normalized to [0, 1], with a Kullback-Leibler penalty (beta = 0.1) ensuring alignment with the reference model.

Quantization of the 11-billion-parameter model to 4 bits with rank-4 low-rank adaptation reduces its memory footprint M_{total} as:

$$M_{\text{total}} = (P \times B/8) + A + T + D, \tag{6}$$

where $P = 11 \times 10^9$ parameters, $B = 4$ bits (0.5 bytes), $A = 0.008$ gigabytes (low-rank adapters), $T = 1.5$ gigabytes (training overhead), and $D = 0.16$ gigabytes (dataset). This yields $M_{\text{total}} = (11 \times 0.5) + 0.008 + 1.5 + 0.16 = 7.168$ gigabytes [1].

Training occurs over 2 epochs with a batch size of 2 and a learning rate of 1×10^{-4}, generating two completions per prompt (temperature 0.7, 32-token limit) using online sampling from the real-time policy model. Validated on the test set, this combined approach enhances reasoning and problem-solving efficiently, drawing on prior work demonstrating parameter-efficient gains and reward-based optimization improvements [8].

4 Expected Contributions and Metrics

This study contributes by fine-tuning the LLaMA model with Group-Relative Policy Optimization and evaluating its ability to generalize to unseen problems in the Humanity's Last Exam dataset. We adopt the same evaluation metrics and methodology as the original Humanity's Last Exam study to ensure consistency and comparability [6]. Specifically, we focus on model accuracy, which measures the percentage of correct answers to test questions, and calibration error, which assesses the alignment between the model's predicted confidence and actual performance. These metrics are sound for evaluating reasoning and reliability, maintaining continuity with prior work while addressing our hypothesis. As this is a research proposal, we do not present results but aim to establish a framework for quantifying improvements in predictive accuracy and confidence calibration, critical for assessing adaptation to novel academic challenges.

5 Results

Table 1 presents the comparative evaluation of our approach (LLaMA + GRPO) alongside models from the Humanity's Last Exam study. The table includes two key metrics: model accuracy, which measures the ability to correctly solve unseen problems, and calibration error, which assesses the reliability of the model's confidence estimates. Our evaluation follows a split train-evaluation framework, focusing on the model's capacity to generalize to new and challenging problem domains.

Model	Accuracy (%) ↑	Calibration Error (%) ↓	Source
GPT-4o	3.1	92.3	[6]
Grok-2	3.9	90.8	[6]
Sonnet 3.5	4.8	88.5	[6]
Gemini Flash Thinking	7.2	90.6	[6]
o1	8.8	92.8	[6]
DeepSeek-R1*	8.6	81.4	[6]
o3-mini (medium)*	11.1	91.5	[6]
o3-mini (high)*	14.0	92.8	[6]
LLaMA	-	-	
LLaMA + GRPO (ours)	-	-	

Table 1. Comparative Evaluation of Models on the Humanity's Last Exam Benchmark. *Model is not multi-modal, evaluated on text-only subset.

6 Discussion

TBD

7 Conclusion and Future Work

TBD

8 Acknowledgements

The author acknowledges the support of the School of Computing at the University of Leeds and extends gratitude to the developers of QLoRA, GRPO, and the HLE dataset for making their resources publicly available.

9 Data Access Statement

The code and results used in this study are openly accessible at *this repository*. It contains all the scripts, configurations, and datasets analysis necessary to reproduce the experiments and validate the findings presented in this work.

References

[1] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG]. URL: https://arxiv.org/abs/2305.14314.

[2] David Deutsch. *The Beginning of Infinity: Explanations That Transform the World*. New York: Viking Press, 2011.

[3] Jiatao Gu et al. *Non-Autoregressive Neural Machine Translation*. 2018. arXiv: 1711.02281 [cs.CL]. URL: https://arxiv.org/abs/1711.02281.

[4] Dan Hendrycks et al. "Measuring Mathematical Problem Solving With the MATH Dataset". In: *NeurIPS* (2021).
[5] Jason Lee, Elman Mansimov, and Kyunghyun Cho. *Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement*. 2018. arXiv: 1802.06901 [cs.LG]. URL: https://arxiv.org/abs/1802.06901.
[6] Long Phan et al. *Humanity's Last Exam*. 2025. arXiv: 2501.14249 [cs.LG]. URL: https://arxiv.org/abs/2501.14249.
[7] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.
[8] Zhihong Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. 2024. arXiv: 2402.03300 [cs.CL]. URL: https://arxiv.org/abs/2402.03300.
[9] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

10 Appendix

10.1 Appendix A: Dataset Description

Category	Questions	Min Test Size	Min Test %	Train %
Biology/Medicine	303	143	47.260000	52.740000
Chemistry	170	105	61.560000	38.440000
Computer Science/AI	258	132	51.290000	48.710000
Engineering	130	88	67.720000	32.280000
Humanities/Social Science	235	126	53.630000	46.370000
Math	1106	218	19.670000	80.330000
Other	258	132	51.290000	48.710000
Physics	240	127	53.100000	46.900000

Table 2. Statistics by Category for cais/hle Test Split (90% CI, 5% Error)

Category	Questions	Train %	Test %	% Image	% Exact Match
Biology/Medicine	303	52.81	47.19	19.58	48.25
Chemistry	170	38.24	61.76	38.10	75.24
Computer Science/AI	258	48.84	51.16	6.82	69.70
Engineering	130	32.31	67.69	46.59	72.73
Humanities/Social Science	235	46.38	53.62	11.11	59.52
Math	1106	80.29	19.71	4.13	91.28
Other	258	48.84	51.16	22.73	75.00
Physics	240	47.08	52.92	6.30	84.25

Table 3. Dataset Train/Test Split by Category

Category	Text Tokens	Image Size (MB)	Images
Biology/Medicine	$\downarrow$11 - $\uparrow$3008 257.13μ $\pm$344.08σ	0.3582μ $\pm$0.3318σ	19.47%
Chemistry	$\downarrow$16 - $\uparrow$1699 186.07μ $\pm$259.06σ	0.0655μ $\pm$0.0967σ	36.47%
Computer Science/AI	$\downarrow$15 - $\uparrow$5052 420.34μ $\pm$648.82σ	0.2335μ $\pm$0.3776σ	5.81%
Engineering	$\downarrow$14 - $\uparrow$8972 396.45μ $\pm$919.56σ	0.1733μ $\pm$0.2037σ	40.0%
Humanities/Social Science	$\downarrow$15 - $\uparrow$1294 201.03μ $\pm$227.64σ	0.3205μ $\pm$0.3624σ	10.64%
Math	$\downarrow$15 - $\uparrow$13518 216.81μ $\pm$483.85σ	0.145μ $\pm$0.211σ	4.07%
Other	$\downarrow$11 - $\uparrow$7156 174.16μ $\pm$478.07σ	0.3883μ $\pm$0.4474σ	22.48%
Physics	$\downarrow$11 - $\uparrow$7313 248.2μ $\pm$520.54σ	0.1423μ $\pm$0.253σ	5.83%

Table 4. Dataset Description by Category for cais/hle Test Split