UNIVERSITY OF TECHNOLOGY SYDNEY

43008 Reinforcement Learning

ASSESSMENT 3 (Part F):

Presentation and Demo, Recorded video, Viva

Project 9 - Team: Q^2 + P

MEDICAL DOCTOR AGENT

Reinforcement Learning with Verifiable Rewards (RLVR)

using AI Verifier for Medical Reasoning Correctness

24908396 – Anh Quan Tran

26124228 – Paul Butler

25076833 – Hoang Quan Dang

Instructors:

Assoc Prof. Nabin Sharma

Shudarshan Kongkham

Rozhin Vosoughi

INTRODUCTION AND BACKGROUND

- Over half the world's population lack access to essential health services (WHO, 2023). Innovative solutions using AI can help.
- **Medical doctor agent** diagnoses medical condition based on patient's symptoms.
- But cost of failure is high, medical diagnosis requires complex and transparent reasoning, and a key challenge is verifying the reasoning.

Our goals:

- Structured output: Clear, machine-readable format adherence.
- Verifiable reasoning: Enable step-by-step validation of the model's logic.
- High accuracy: Ensure both reasoning & final answer are correct.

An example of desired output:

<THINK>Okay, let's see. We have...

Step 1: Identify the symptoms...

Step 2: Consider several factors...

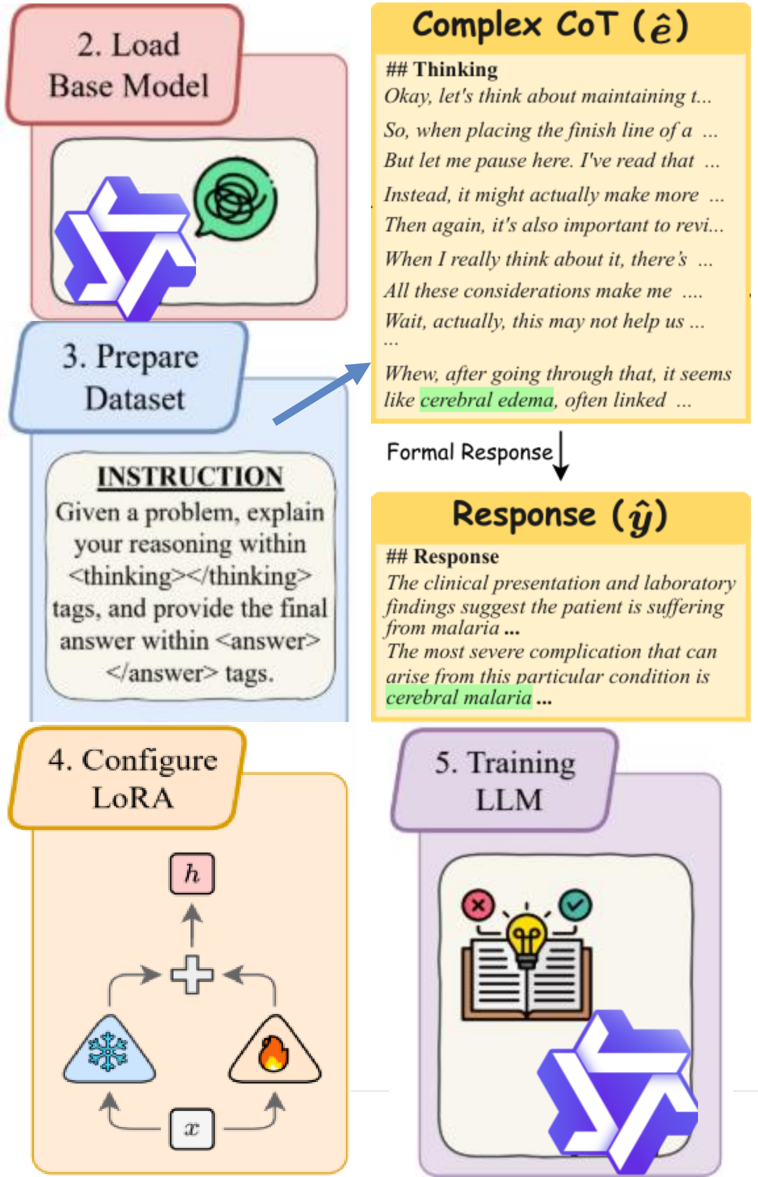
Thus, the drug should be Amlodipine... </THINK>

<ANSWER>Amlodipine</ANSWER>

- **Stage 1** - LLM learns complex medical reasoning **+ Stage 2** - Reasoning improved with RL (PPO or GRPO).
- **Applications:** AI medical care for patients unable to access a human doctor, AI triage to increase health system productivity, second medical opinion for doctors/patients.

PROPOSED SOLUTION

STAGE 1: Supervised Fine-tuning

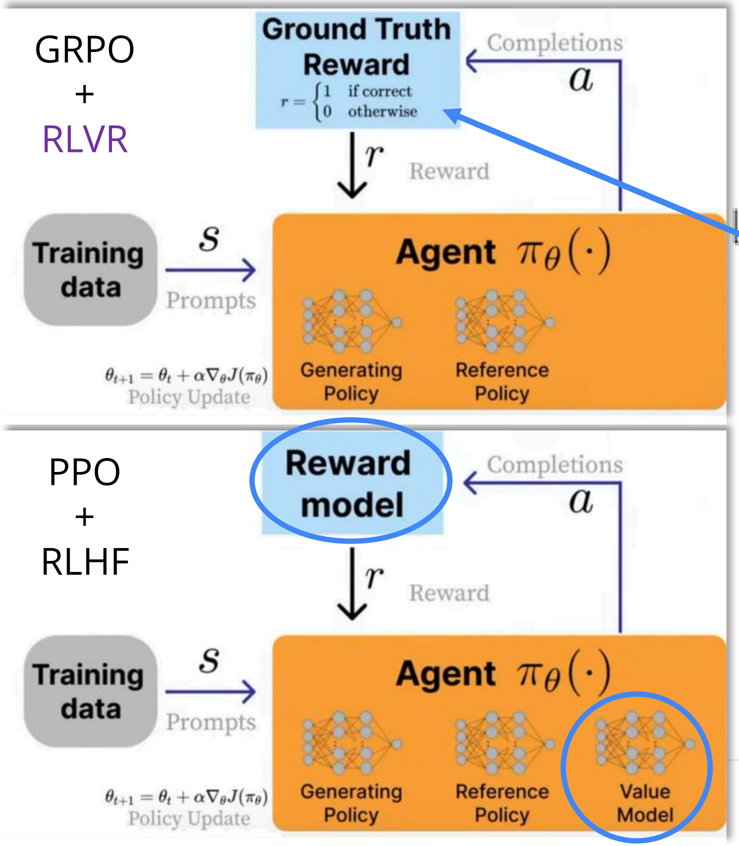


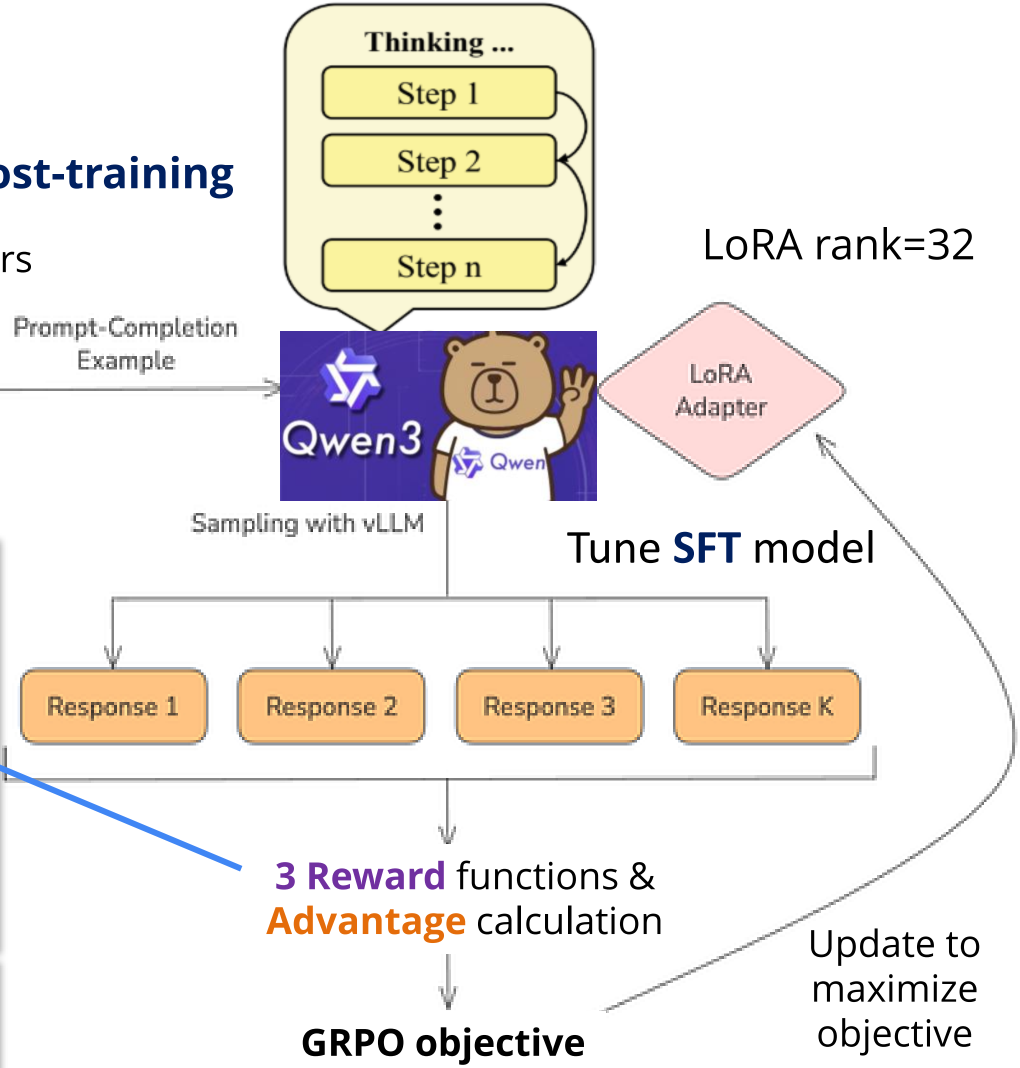
STAGE 2: RL Post-training

40K medical exams problems with answers

Verifiable Medical Problem (x): A 30-year-old woman recently traveled to India and now presents with shaking, chills, fevers, headaches, pallor, and jaundiced sclera. Vital signs: Temp 38.9°C, RR 19/min, BP 120/80 mm Hg, Pulse 94/min. Labs: Hct 30%, Total bilirubin 2.6 mg/dL, Direct bilirubin 0.3 mg/dL. What is the most severe complication of this condition?

Ground-true Answer (y*): Cerebral edema

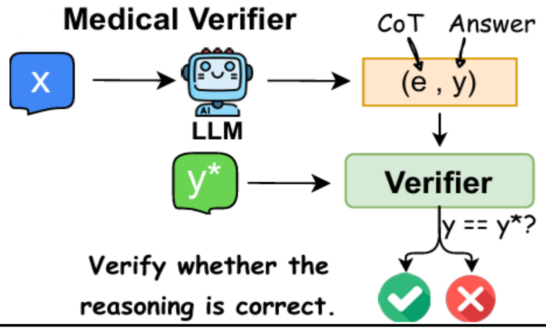




GRPO reduces training resources by:

- Replace **Value Model** with group advantage.
- Replace **Reward Model** (rely on human, cause reward hacking) with verifiable rewards.

MULTI-REWARD SYSTEM

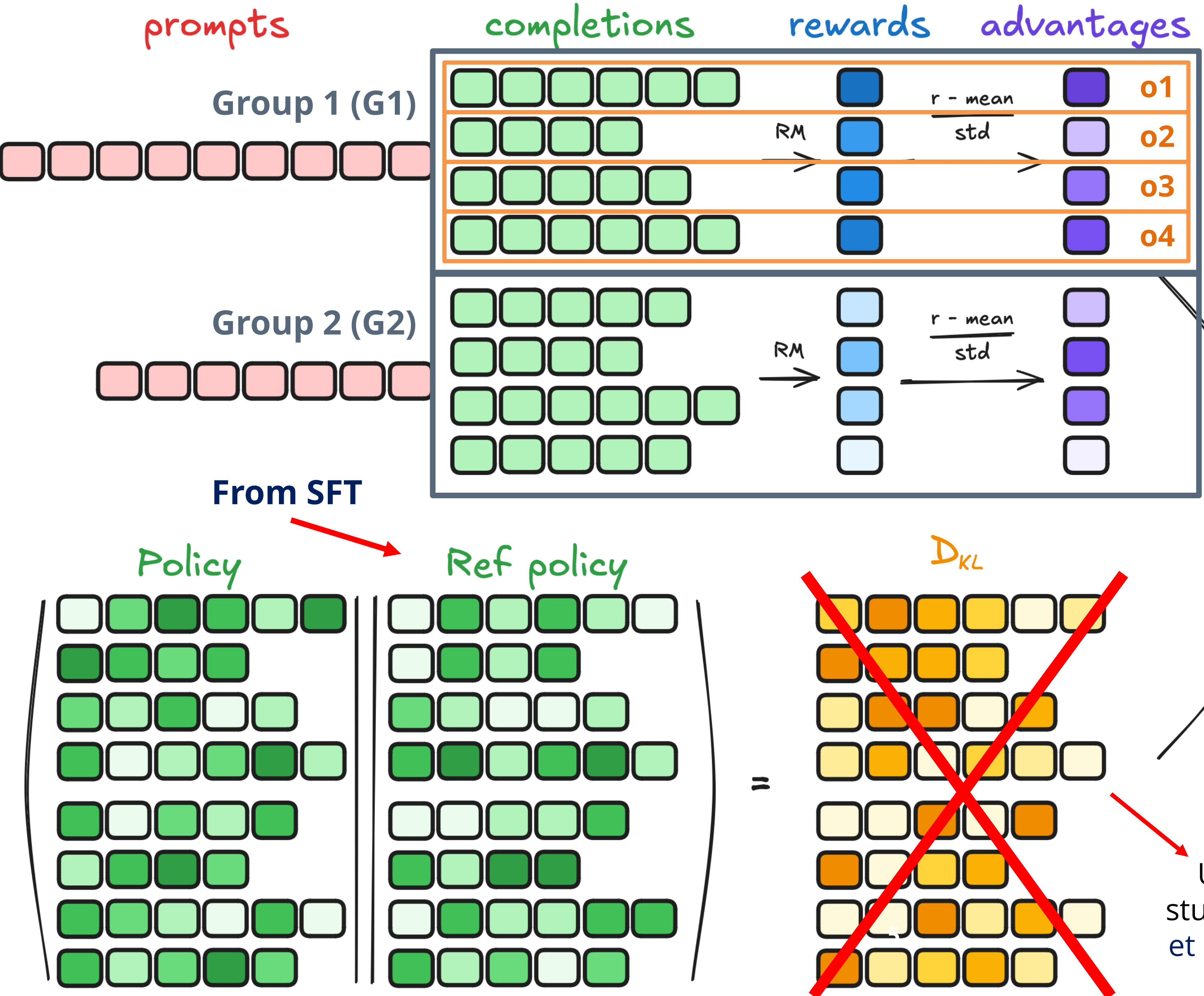
Function	Description	Scoring	Justification
1. Strict Format	Did it perfectly follow: <THINK> … </THINK> <ANSWER> … </ANSWER>	3.0 (Perfect) 0.0 (Otherwise)	- Ensure structured output. - Large reward for perfect structure.
2. Soft Format	Partial credit for tags (e.g., count of <THINK>, etc.)	0.5 (Correct tag) -0.5 (Redundant tag)	- Graduated learning. - Mitigate sparse rewards and improve convergence.
3. Medical Accuracy	How close does it align with the ground-truth answer? Handle aliases via semantic verification with LLM verifier.	>0.9: 5.0 (High confidence) >0.7: 3.5 (Strong alignment) >0.5: 2.0 (Partial/approximation) >0.3: 1.5 (Reasonable attempt) ≤0.3: -2.5 (Wrong answer)	 Return a probability

Single-reward limitations:

- ***Goodhart's law**: When a measure becomes a target, it ceases to be a good measure.*
- In standard RLHF with **PPO**, rewards are derived from 1 model **=>** high variance & slow convergence.
- Models might learn to **hack** the reward.

- **GRPO**'s group-relative ranking reduces variance and scales rewards at group/batch level (mean/std normalization).
- **Multi-Reward** prevents over-optimization on 1 aspect (e.g., correct format but wrong answer).

GRPO OBJECTIVE IMPROVEMENT



$$\hat{A} = \frac{r - \text{mean at local/group level}}{\text{std at global/batch level}}$$

(Liu et al., 2025)

$-$ → mean → objective

$$\frac{1}{\cancel{\sum_{i=1}^{G}|o_i|}}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}$$

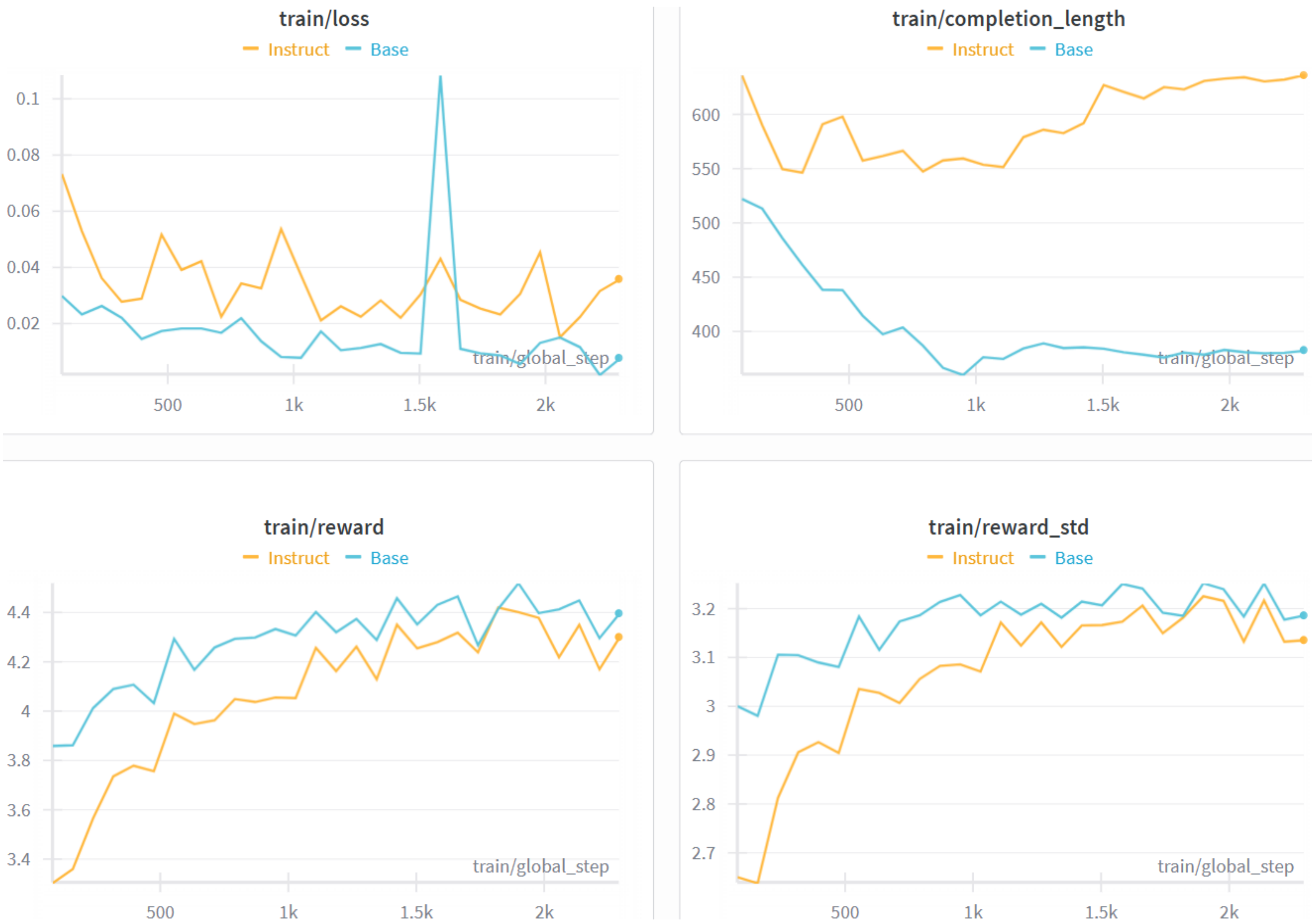
$$\frac{1}{max\_completion\_len * G}$$

in Dr. GRPO by Liu et al. (2025)

Unnecessary in recent studies (Hu et al., 2025; Liu et al., 2025; Yu et al., 2025)

PROJECT 9

EXPERIMENTAL RESULTS



Qwen3-1.7B	Instruct	Base
Match Format Strictly	3.00 ± 0.01	3.00 ± 0.00
Match Format Softly	1.49 ± 0.00	1.50 ± 0.00
Check Answer Correctness	-0.19 ± 3.13	-0.10 ± 3.18
Average Reward	4.30 ± 3.14	4.40 ± 3.19
Completion Lengths	636 tokens (336 to 1176)	383 tokens (241 to 582)

EXPERIMENTAL RESULTS

Qwen3-1.7B	RL Method	MedQA_USLME_test	MedMCQA_validation
Instruct	PPO	579/1273 (45.48%)	1674/4183 (40.04%)
	GRPO	**629/1273 (49.41%)**	**1927/4183 (46.07%)**
Base	GRPO	570/1273 (44.78%)	1858/4183 (44.42%)

Qwen3-1.7B	Correct Count	MedQA_USLME_test			MedMCQA_validation		
		Before SFT/RL	After SFT/RL	Improvement	Before SFT/RL	After SFT/RL	Improvement
Instruct	Format	646/1273 (50.75%)	**1273/1273 (100.00%)**	+627 (49.25%)	3152/4183 (75.35%)	**4183/4183 (100.00%)**	+1031 (24.65%)
	Answer	375/1273 (29.46%)	**629/1273 (49.41%)**	+254 (19.95%)	1601/4183 (38.27%)	**1927/4183 (46.07%)**	+326 (7.79%)
	Both	375/1273 (29.46%)	**629/1273 (49.41%)**	+254 (19.95%)	1601/4183 (38.27%)	**1927/4183 (46.07%)**	+326 (7.79%)
Base	Format	531/1273 (41.71%)	1271/1273 (99.84%)	+740 (58.13%)	1543/4183 (36.89%)	4180/4183 (99.93%)	+2637 (63.04%)
	Answer	239/1273 (18.77%)	570/1273 (44.78%)	+331 (26.00%)	655/4183 (15.66%)	1858/4183 (44.42%)	+1203 (28.76%)
	Both	239/1273 (18.77%)	570/1273 (44.78%)	+331 (26.00%)	655/4183 (15.66%)	1858/4183 (44.42%)	+1203 (28.76%)

CONCLUSION AND FUTURE WORK

Recap

- Built a **Medical Doctor Agent** using **SFT + GRPO + LoRA** on **Qwen3-1.7B**.
- Designed a **multi-reward RL system** with format, soft, and medical accuracy rewards.
- Integrated an **LLM-based verifier** for semantic correctness.
- Applied the <THINK> → <ANSWER> reasoning template for transparent outputs.
- Compared **Instruct vs Base** models to assess alignment impact.
- Improved reasoning stability and benchmark accuracy (MedQA, MedMCQA).

Limitations

- Sensitive to **reward design and hyperparameters**.
- **Semantic correctness** still inconsistent despite strong structure.
- **High computational cost** for RL fine-tuning.
- **Short training time** limits full convergence.

Future work

- Test GRPO on other model families (Gemma, Mistral, Llama).
- Enhance verifier robustness for better factual grounding.

DEMO

REFERENCES

Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., & Wang, B. (2024). *HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs*. https://doi.org/10.48550/arxiv.2412.18925

Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., & Shum, H.-Y. (2025). *Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model*. https://doi.org/10.48550/arxiv.2503.24290

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., & Lin, M. (2025). *Understanding R1-Zero-Like Training: A Critical Perspective*. https://doi.org/10.48550/arxiv.2503.20783

Liu, Z., Liu, J., He, Y., Wang, W., Liu, J., Pan, L., Hu, X., Xiong, S., Huang, J., Hu, J., Huang, S., Yang, S., Wang, J., Su, W., & Zheng, B. (2025). *Part I: Tricks or Traps? A Deep Dive into RL for LLM Reasoning*. https://doi.org/10.48550/arxiv.2508.08221

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. https://doi.org/10.48550/arXiv.1707.06347

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., & Guo, D. (2024). *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. https://doi.org/10.48550/arxiv.2402.03300

W. H. Organization. (2023). *Tracking universal health coverage: 2023 global monitoring report*. World Health Organization.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., ... Wang, M. (2025). *DAPO: An Open-Source LLM Reinforcement Learning System at Scale*. https://doi.org/10.48550/arxiv.2503.14476