43008: Reinforcement Learning

Assignent-3, Part-E Project Report

Project Title: Medical Doctor Agent

Team Name and Project Number: Q^2 + P, Project 9

Team Members:

Student Name	Student ID	Email ID
Anh Quan Tran	24908396	anhquan.tran@student.uts.edu.au
Paul Butler	26124228	paul.butler@student.uts.edu.au
Hoang Quan Dang	25076833	hoangquan.dang@student.uts.edu.au

Date: 3rd November 2025

Table of Contents

Abstract

Clinical question-answering requires verifiable reasoning and machine-readable outputs, but general-purpose LLMs often produce unstructured rationales or fragile answers. We introduce a two-stage post-training pipeline that transforms small LMs into structured medical reasoners. First, Supervised Fine-Tuning (SFT) trains the response grammar, reasoning within <THINK>…</THINK> followed by a final medical decision in <ANSWER>…</ANSWER>. Next, we implement Group Relative Policy Optimization (GRPO) with a multi-reward setup that simultaneously optimizes (i) strict format adherence, (ii) partial credit for format, and (iii) semantic answer correctness through an **LLM verifier** that manages clinical aliases and wording differences. We utilize LoRA for efficient parameter updates and a length-independent Dr.GRPO objective to prevent reward-length coupling. Evaluated on **MedQA-USMLE** (n=1,273) and **MedMCQA** (n=4,183), our top model (Qwen3-1.7B-Instruct + GRPO) attains 49.41% and 46.07% exact-match accuracy, respectively, with nearly 100% format compliance; GRPO also surpasses PPO on both datasets. These findings demonstrate that verifier-guided, multi-signal GRPO consistently enhances factual accuracy while ensuring outputs are interpretable and conform to templates, offering a practical route toward reliable, compact medical reasoning systems.

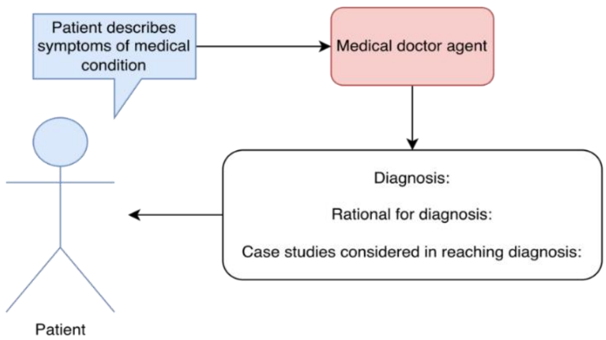
1. Introduction and Background

There is a large unmet need for health services in many parts of the world. Existing health systems do not have sufficient access to human, infrastructure and financial resources to address the unmet need. Innovative solutions are likely the only way to significantly improve this situation. In this project we develop a medical doctor agent that seeks to diagnose a patient's medical condition based on a description of the patient's symptoms. This agent is accessed through a chatbot. This has potential to increase availability of primary medical care services. The medical doctor agent is developed through training a Qwen3 LLM via supervised fine-tuning followed by reinforcement learning with group relative policy optimization.

1.1. The problem you tried to solve

Humans obsess about their health condition and in many instances have insatiable demand for medical assessment and advice. Health systems in most regions of the world are unable to meet all this demand due to budget, human and infrastructure constraints. WHO estimates indicate over half the world's population (~4.5 billion people) lack access to essential health services and >2 billion face severe financial hardship paying for health care (WHO, 2023). There are many aspects to the lack of accessible health services. Our project is focused on addressing the lack of access to primary medical care (i.e. first point of contact for non-hospital healthcare, including advice on prevention, diagnosis, and treatment of illnesses and injuries).

The model environment consists of a patient describing the symptoms of their medical condition to the medical doctor agent. The medical doctor agent then selects relevant case studies, develops a complex chain of thought to find a diagnosis and then presents the diagnosis and relevant rational to the patient.



1.2. Motivation

We believe AI has potential to increase the productivity of the healthcare systems and to reduce the cost of healthcare. The focus of the project is it develop an AI medical doctor agent, that is a step toward delivering primary medical care. In some situations, use of the doctor agent means patients receive primary medical care where it would not be possible otherwise. In rural Australia up to 20% of the population don't have access to a doctor (Australian Institute of Health and Welfare, 2016). In other situations, the doctor agent may increase the productivity of health systems and medical practitioners. This increases the availability or reduces the cost of primary medical care. The doctor agent may improve the quality of medical diagnosis, by providing a second opinion either for healthcare professionals or patients, thereby improving health outcomes for patients. The project motivation is to enable more of the unmet global healthcare demand to be satisfied.

1.3. Application

The medical doctor agent can be accessed through a chatbot. There are many potential applications, including:

- Providing health advice to patients without the need to visit a human doctor. In many instances this means patients receive medical attention, where it would not otherwise be possible without extreme cost or inconvenience. In other instances, the convenience of using a medical doctor agent through a chatbot means patients may seek advice for medical conditions they erroneously consider trivial but require urgent attention.
- At home screening/triage of patients to identify those with highest priority for acute medical care (e.g. ambulance priority).
- At medical clinic screening/triage of patients to increase medical practitioners' productivity.
- In some situations, patients may be more comfortable sharing their medical condition through a chatbot rather than directly to a human doctor. This could improve the likelihood of the patient receiving the appropriate medical care.
- Use by medical practitioners for second checking of a patient's diagnosis.
- Role playing in medical student education delivery.

1.5. Dataset

The dataset for supervised fine-tuning comprises of 19,704 instances of a medical question, complex chain of thought and a response, and was obtained from Chen et al. (2024). An example of data instance is show below.

Example of data for SFT	
Question	Given the symptoms of sudden weakness in the left arm and leg, recent long-distance travel, and the presence of swollen and tender right lower leg, what specific cardiac abnormality is most likely to be found upon further evaluation that could explain these findings?
Complex chain of thought	Okay, let's see what's going on here. We've got sudden weakness in the person's left arm and leg - and that screams something neuro-related, maybe a stroke? But wait, there's more. The right lower leg is swollen and tender, which is like waving a big flag for deep vein thrombosis, especially after a long flight or sitting around a lot. So, now I'm thinking, how could a clot in the leg end up causing issues like weakness or stroke symptoms? Oh, right! There's this thing called a paradoxical embolism. It can happen if there's some kind of short circuit in the heart - like a hole that shouldn't be there. Let's put this together: if a blood clot from the leg somehow travels to the left side of the heart, it could shoot off to the brain and cause that sudden weakness by blocking blood flow there. Hmm, but how would the clot get from the right side of the heart to the left without going through the lungs and getting filtered out? Here's where our cardiac anomaly comes in: a patent foramen ovale or PFO. That's like a sneaky little shortcut in the heart between the right and left atria. And it's actually pretty common, found in about a quarter of adults, which definitely makes it the top suspect here. So with all these pieces - long travel, leg clot, sudden weakness - a PFO fits the bill perfectly, letting a clot cross over and cause all this. Everything fits together pretty neatly, so I'd bet PFO is the heart issue waiting to be discovered. Yeah, that really clicks into place!
Response	The specific cardiac abnormality most likely to be found in this scenario is a patent foramen ovale (PFO). This condition could allow a blood clot from the venous system, such as one from a deep vein thrombosis in the leg, to bypass the lungs and pass directly into the arterial circulation. This can occur when the clot moves from the right atrium to the left atrium through the PFO. Once in the arterial system, the clot can travel to the brain, potentially causing an embolic stroke, which would explain the sudden weakness in the left arm and leg. The connection between the recent travel, which increases the risk of deep vein thrombosis, and the neurological symptoms suggests the presence of a PFO facilitating a paradoxical embolism.

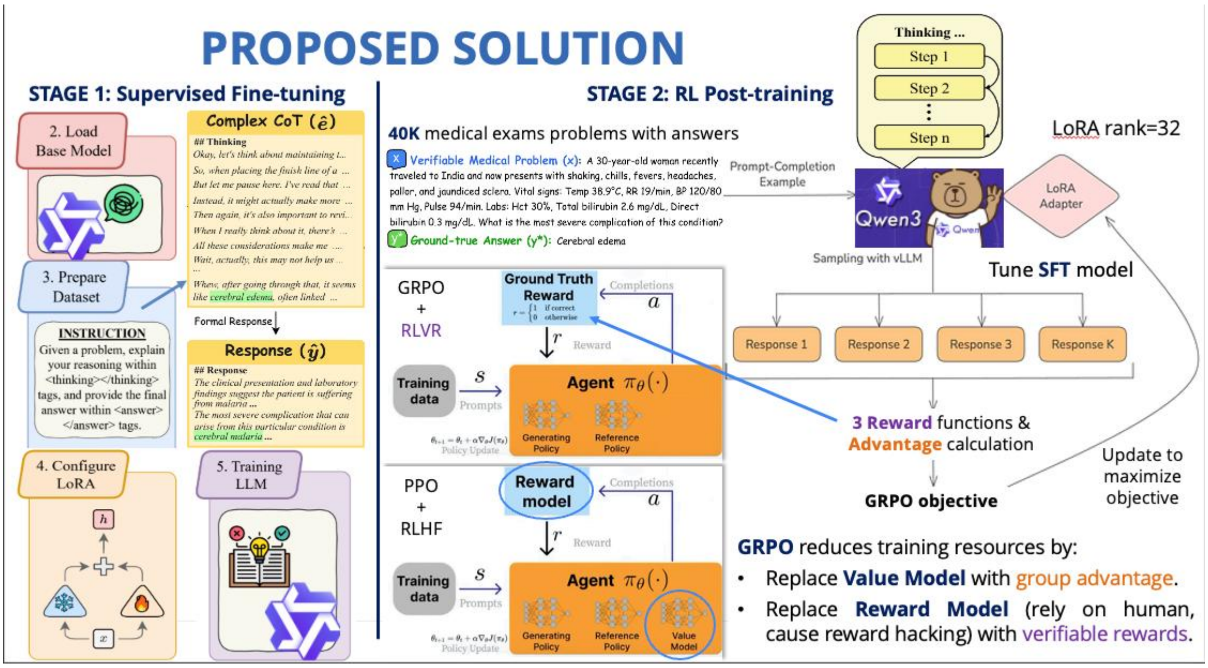
For reinforcement learning, we obtained a dataset of 40,644 open ended & verifiable medical questions and answers (Chen et al., 2024). The dataset was produced by taking multiple choice questions from medical exams and rearranging these into open ended questions and answers. Examples of data instances are shown below.

Example of data for PPO and GRPO	
Question	**Answer**
An 88-year-old woman with osteoarthritis is experiencing mild epigastric discomfort and has vomited material resembling coffee grounds multiple times. Considering her use of naproxen, what is the most likely cause of her gastrointestinal blood loss?	Gastric ulcer
In the context of disseminated intravascular coagulation (DIC), which blood component is expected to show an increase due to the excessive breakdown of fibrin?	Fibrin degradation products
In a 3-year-old boy with severe diarrhea, vomiting, fever, and dry mucous membranes, who is unvaccinated and has been in contact with other similarly affected children at daycare, what structural features are characteristic of the RNA virus likely causing his illness?	Double-stranded, icosahedral, non-enveloped
Based on the chest radiograph and abdominal CT scan of a middle-aged male complaining of nagging abdominal pain for the past 2 weeks, what is the probable diagnosis that should be considered?	Hydatid Cyst

2. Methodology

2.1. Pipeline Overview

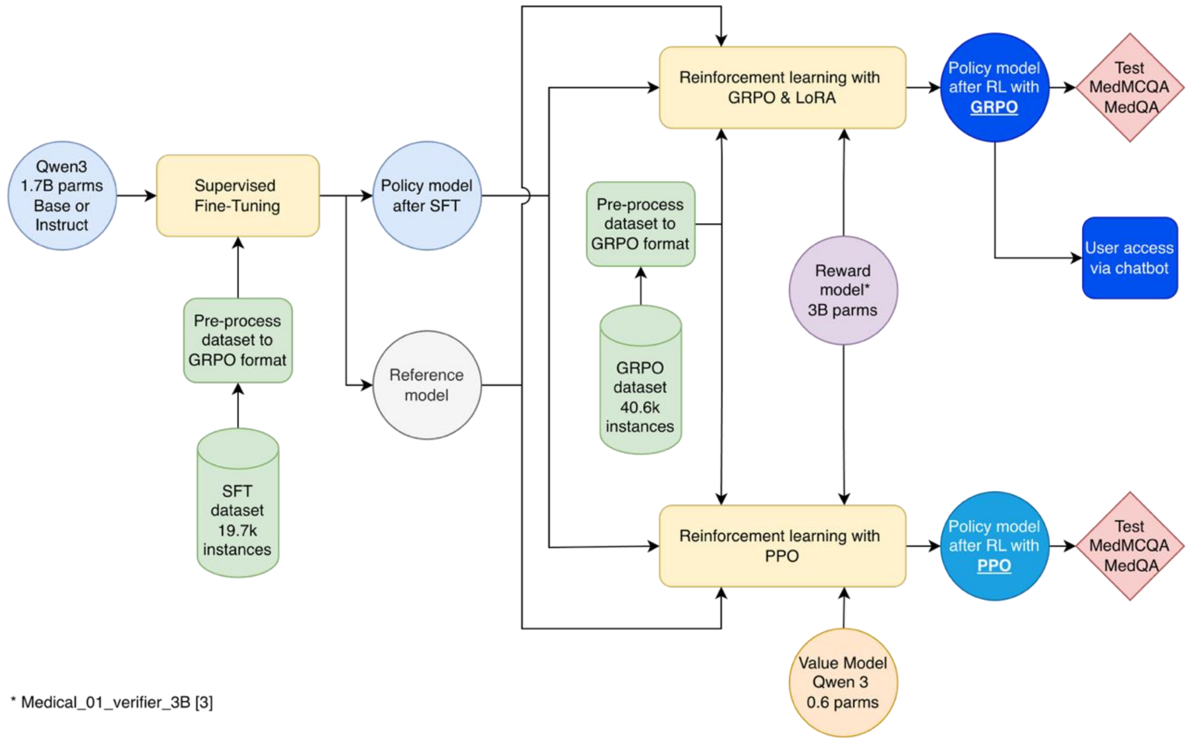
To develop a model capable of accurate, explainable, and format-constrained clinical reasoning, we propose a two-stage fine-tuning pipeline that gradually incorporates both structural consistency and reasoning quality into the model. This pipeline includes **(i) Supervised Fine-Tuning (SFT)** to establish output formatting standards, followed by **(ii) Reinforcement Learning (RL)** with **Group Relative Policy Optimization (GRPO)** to enhance clinical reasoning and answer correctness through multi-reward optimization.



In the 1st stage, we train the model to produce outputs in a structured format using high-quality examples of medical reasoning. This phase is not designed to optimize accuracy but to enforce the reasoning's grammar: embedding detailed logic within **<THINK>** tags and clear conclusions within **<ANSWER>** tags. This prepares the model to produce outputs that can be reliably parsed and evaluated by downstream reward functions.

The 2nd stage leverages **GRPO**, a recent RL method that avoids the instability and inefficiencies of PPO by not requiring value-function estimation. Instead, **GRPO** computes group-relative advantages across multiple completions for a given prompt, making it particularly suitable for tasks like medical QA where absolute scoring can be noisy or unreliable. To further address reward sparsity and instability, we introduce a multi-reward system that assesses each generated response along 3 independent axes: strict format compliance, soft structural adherence, and semantic correctness verified through an LLM-based verifier.

Together, these 2 phases work synergistically: SFT bootstraps the model's ability to communicate in a structured and auditable format, while **GRPO** optimizes clinical reasoning fidelity and factual alignment within that structure. This pipeline enables us to fine-tune general-purpose LLMs into domain-aligned medical reasoners.

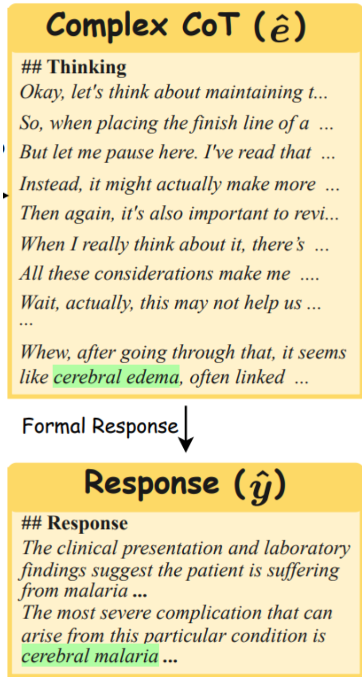


2.2. Stage 1 - Supervised Fine-Tuning (SFT)

SFT establishes a structured, auditable output format and introduces complex clinical reasoning patterns before any reinforcement signals are applied. Specifically, the model is trained to produce a two-part response: an explicit deliberation trace and a final decision. This stage enhances procedural coherence (clear **THINK/ANSWER** structure) and reasoning fluency (multi-step clinical logic), creating a stable foundation for subsequent policy improvements.

We fine-tune on the public **FreedomIntelligence/medical-o1-reasoning-SFT** dataset, which offers question–reasoning–answer triples specifically curated for medical reasoning in English (Chen et al., 2024). This dataset is suitable for teaching complex CoT trajectories that are consistent with verifiable medical problem, making it well suited to impose the desired **THINK/ANSWER** structure during SFT.

During SFT, the model **imitates** high-quality, verifier-consistent trajectories, given a clinical prompt, the target output includes (i) a multi-step rationale (symptoms -> differentials -> tests -> reconciliation) in **<THINK>…</THINK>**, followed by (ii) a concise clinical decision in **<ANSWER>…</ANSWER>**. Because these trajectories are tied to **verifiable medical problems**, where each item pairs an open-ended question with a ground-truth answer, the supervision combines structural discipline with semantic intent, preparing the policy for later scoring using verifiable signals rather than fragile string matches. This stage trains the model to **think before responding**, establishing a disciplined process where it explores, assesses, and refines its clinical reasoning before making a final decision.

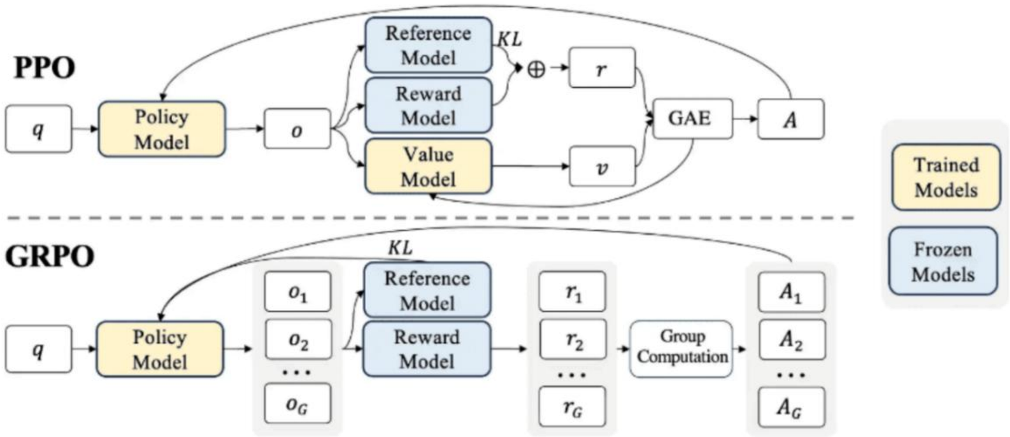


2.3. Stage 2 - Enhancing Reasoning with Reinforcement Learning

For policy optimization, we use **Group Relative Policy Optimization (GRPO)** (Shao et al., 2024), a state-of-the-art reinforcement learning technique that addresses key limitations of traditional **Proximal Policy Optimization (PPO)** (Schulman et al., 2017)**.** Specifically, **PPO** for LLMs typically trains a value network and stores long token trajectories to compute generalized advantage estimates. With **multi-hundred-token** clinical **Chain-of-Thought (CoT)**, that critic pathway becomes memory-intensive and error-prone: instability in value regression leads to noisy advantages.

GRPO is a **critic-free** variant of PPO tailored to post-training LLMs: for each prompt, the model generates a **group of completions**; each completion receives a composite reward; advantages are computed **relative to the group** and normalized; then a clipped importance-sampling surrogate (akin to PPO's) updates the policy against a frozen reference model. Formally, if r_i is the reward of completion o_i in a group of size G .

GRPO forms: $\overline{r} = \frac{1}{G}\Sigma_i r_i$; $\sigma_r = \sqrt{\frac{1}{G}\Sigma_i (r_i - \overline{r}\,)^2}$ and advantages $A_i = \frac{r_i - \overline{r}}{\sigma_r}$; the policy is updated with a clipped ratio objective while optionally regularizing to a reference policy via KL, exactly mirroring PPO's trust-region effect but **without** a learned value function (critic) or GAE rollouts for long sequences.



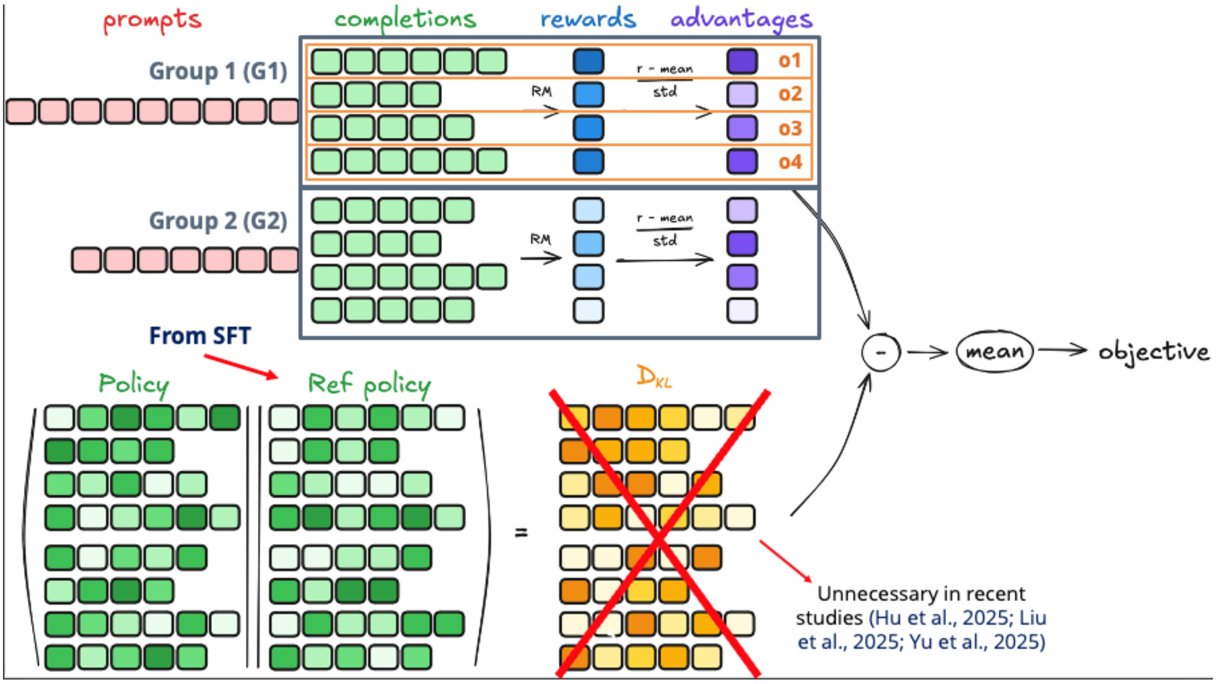
Given SFT policy π_{ref} and current policy π_θ, GRPO uses the **clipped importance-ratio surrogate** familiar from PPO while plugging in **group-relative advantages**:

$$\mathcal{L}_{GRPO}(\theta) = -\,E_{i,t}\left[\,min\left(ratio_{i,t}\hat{A}_i, clip\left(ratio_{i,t}, 1-\epsilon_c, 1+\epsilon_c\right)\hat{A}_i\right) - \beta D_{KL}(\pi_\theta || \pi_{ref})\right],$$

where $ratio_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t}|)}{\pi_{ref}(o_{i,t}|q,o_{i,<t}|)}$ is the importance sampling ratio (token/sequence-level). By eliminating the need for a learned value function, instead using group-relative advantage estimation across multiple responses, **GRPO** not only reduces computational cost but also improves training stability and scalability.

In addition, clinical problems vary in difficulty; absolute rewards fluctuate across prompts. By centering and scaling within the group (mean/std), **GRPO** converts absolute scores into relative wins/losses per prompt, yielding gradient variance that scales roughly with O(1/G). This "race within a heat" makes the signal robust to question difficulty and to reward-scale changes, which is crucial for our **multi-reward design**. Finally, **PPO** commonly couples a **single scalar reward** (often from a learned reward model) with a critic. There are 2 problems with this. First, the reward model relies on human judgments that usually lack explicit criteria and require expensive human annotation. Second, models can **exploit correlations in length or phrasing**. For example, it can generate very long completions if the length is correlated with a higher score. The solution here is to define a list of smaller verifiable rewards, not a final all-consuming singular one. **GRPO** naturally pairs with rule or verifier-based rewards, allowing us to eliminate the need for a reward model and replace subjective human evaluation with reliable, objective signals. Moreover, as it compares completions **against each other**, it focuses on which answer is better rather than how high the raw score is. This tightens the connection between optimization and our goals (strict format, soft format, medically correct answer). **GRPO** formulation explicitly combines multiple rewards/completion and updates with the clipped surrogate, resulting in strong improvements in reasoning tasks.

Conceptually, this classifies our training within the **Reinforcement Learning with Verifiable Rewards (RLVR)** framework: rewards are derived from objective checks against ground truth using a verifier that assesses correctness or incorrectness, rather than from a preference model that yields relative, subjective scores. RLVR has been emphasized as a practical alternative to preference-based RL, such as **Reinforcement Learning with Human Feedback (RLHF)**, for reasoning tasks because it lowers variance, reduces reward hacking (such as length inflation), and enables programmatic reward design that is closely aligned with task validity.



To further stabilize learning with long LLM sequences, we incorporated 3 refinements:

- **Multi-Reward System:** Each completion is scored by a programmable multi-reward system (strict format, partial format, semantic correctness via a verifier). The overall reward is then normalized within the group for GRPO advantages, making it ideal for group-relative estimation.

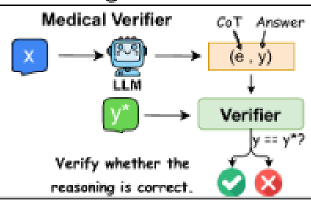
- **Batch-level scaling and efficient sampling:** We calculate group-level mean and the standard deviation at the batch level to make scales comparable across prompts, while using fast generation (vLLM-style multi-sample decoding) to implement the "many candidates per prompt" regime efficiently.

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

- **Regularization without KL:** The KL to the reference policy can be used as a penalty, however, recent studies showed that KL is often unnecessary when ratio clipping and a strong reference policy are present (Liu et al., 2025; Yu et al., 2025; Hu et al., 2025). Thus, we set $\beta = 0$ and rely on clipping and reference-policy anchoring for regularization.

- **Dr-GRPO loss removes length bias:** Standard sequence-normalized losses can favor shorter or longer outputs. We therefore adopt Dr. GRPO (Liu et al., 2025), which divides by a fixed generation budget rather than the sequence length, removing response-length bias and allowing the model to focus purely on content quality.

2.4. 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

Our core innovation is this multi-reward design. A single reward is not enough to capture the nuances of good medical reasoning. In GRPO, multiple rewards can be summed (or weighted) for each completion, forming a composite reward $r_i = \Sigma_{k=1}^{K} w_k r_{i,k}$. Therefore, we designed a 'panel of 3 expert judges' working in parallel ($K = 3$), each evaluating the model's output from a different perspective:

1. **Strict Format Compliance:** We assign a high reward **(+3.0)** only when the output perfectly follows the required structure: reasoning enclosed in **<THINK>…</THINK>** followed by a final decision in **<ANSWER>…</ANSWER>**. Giving a large positive reward here makes format a key criterion: without a parseable transcript, higher-level rewards (e.g., correctness) cannot be reliably measured. This shaping signal bootstraps stable CoT behavior early and prevents reward hacking through unstructured, verbose text.

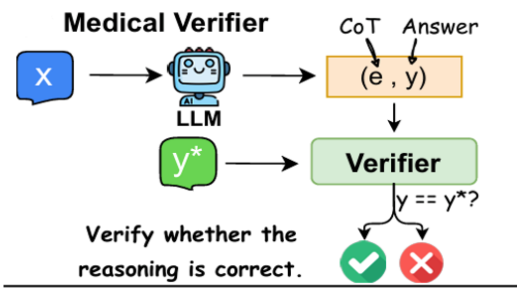
2. **Partial Format Credit:** To encourage incremental learning, especially from noisier base policies, we assign graduated credit (ranging from -1.5 to +1.5 total) based on the presence or absence of key tags, even if the full template isn't satisfied. for example, if it still includes the **</THINK>** tag correctly, it still gets a small amount of credit. This reduces advantage variance within a GRPO group (making gradients better conditioned) and discourages degenerate strategies such as duplicating tags to chase format points.

3. **Medical Answer Accuracy (Verifier-based):** Because medical answers often have aliases or paraphrases, exact string matching is fragile. We therefore use a medical verifier to evaluate the semantic alignment between the generated **<ANSWER>** and **the ground-truth label**. Rewards are graded in parts based on the verifier's confidence.

   - **+5.0** if the verifier predicts *True* with high confidence (e.g., p>0.9).
   - **+3.5 / +2.0 / +1.5** for progressively lower probabilities but still supportive (e.g., p>0.7,0.5,0.3).
   - **−2.5** if judged incorrect.

   This design provides a verifiable, model-based, rule-driven medical signal without the need for costly human preference models, following best practices to avoid absolute scoring that could encourage reward hacking.

2.5. Medical Verifier



Large-scale medical QA struggles with aliasing (e.g., drug synonyms, abbreviations, order-dependent phrasings), which makes "exact-match" reward functions fragile. To prevent rewarding length or formatting tricks while still assessing factuality, we adopt an **LLM-as-verifier**: For each generated response, a compact classifier LLM analyses the model's answer and the gold reference, providing a correctness judgment (or probability) that the two are semantically equivalent.

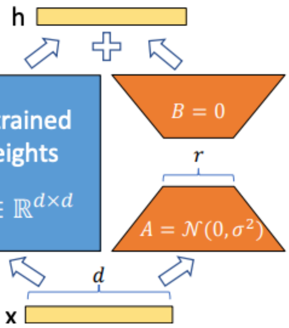
It is important to distinguish this from **Reinforcement Learning from AI Feedback (RLAIF)** (Bai et al., 2022). In RLAIF, an AI system acts as the rater, learning preferences or applying a written "constitution," and the policy is optimized to satisfy those (largely normative) judgments (e.g., harmlessness or helpfulness). By contrast, our verifier does **not** learn preferences; it checks whether an answer is factually consistent with the reference label. In other words, we use AI to measure verifiable correctness, not to define what is preferred.

We use a compact sequence-classification LLM as the verifier. It is prompted with two blocks: the model's response and the reference answer, then asked to produce a binary judgment ("True/False") or a calibrated probability of correctness. In addition, the verifier score contributes to our "Medical Accuracy" reward: high scores earn high rewards (e.g., ≥0.9), partial credit is given for near-matches (e.g., 0.5–0.9), and penalties are applied for contradictions or non-answers. Since the check is semantic, it naturally manages aliases and minor lexical differences, avoiding false negatives common with exact string matching.

Using an **LLM verifier** transforms medical QA into a **verifiable RL** problem: the reward focuses on factual alignment with gold answers, not stylistic preferences. This avoids the issues of preference models (subjectivity, instability) while providing GRPO with a clear, low-variance signal that directly optimizes what we care about: **semantically accurate, structurally correct clinical reasoning**.

2.6. Low-Rank Adaptation

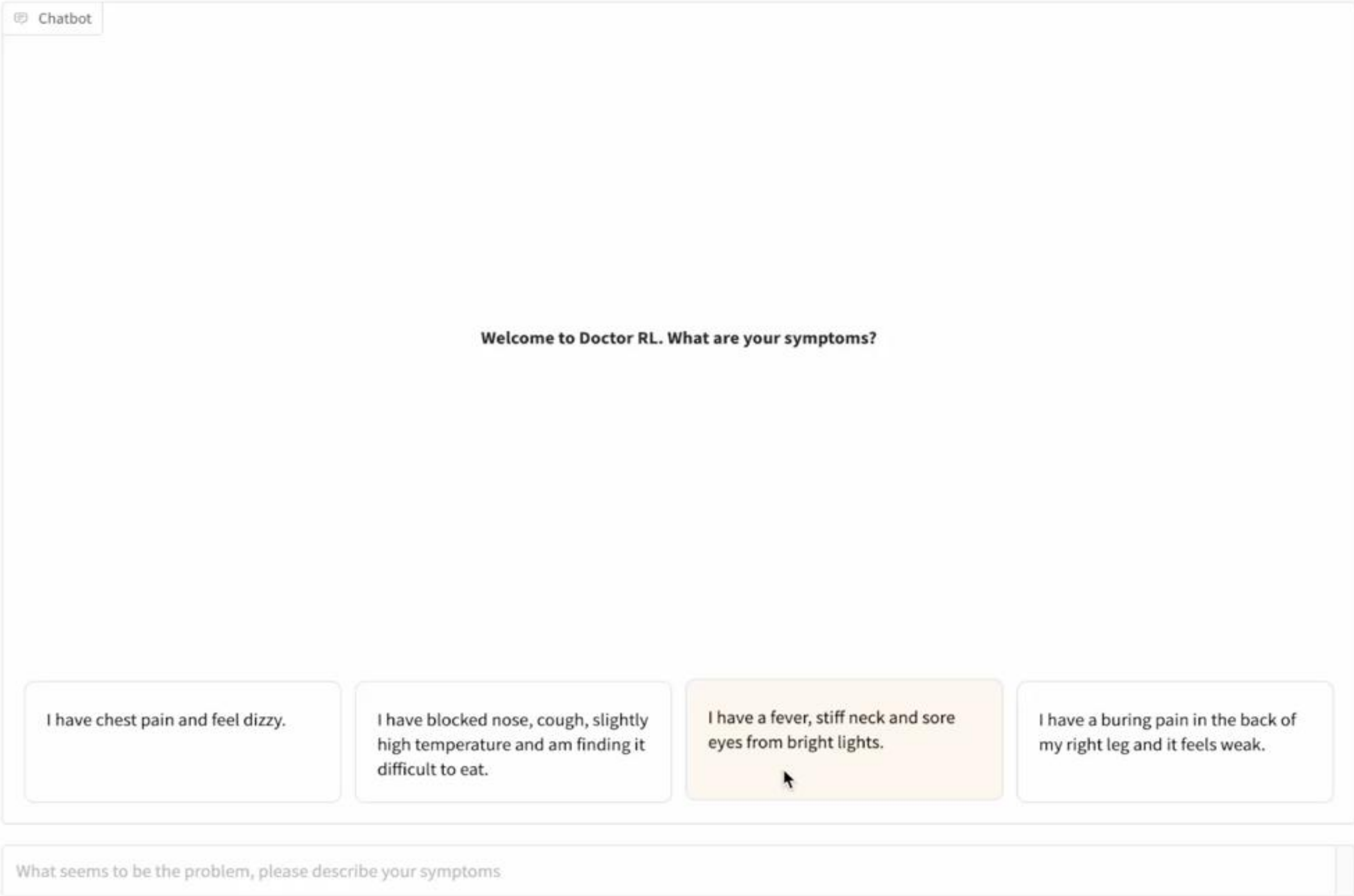
To fine-tune reasoning behaviour under tight memory budgets, we used **Low-Rank Adaptation (LoRA),** which freezes the pretrained weights and learns a pair of small low-rank matrices per selected linear layer. Formally, for base weight $W \in R^{d\times k}$, **LoRA** parameterizes the update as $\triangle W = BA$, where $A \in R^{d\times k}$, $B \in R^{d\times r}$, $r \ll \min(d,k)$, the fowad pass becomes $h = \left(W + \frac{\alpha}{r}BA\right)x$, where α scales the LoRA path. During adaptation only low-rank matrices A, B are updated while W remains frozen, yielding the same functional capacity as full fine-tuning but with orders-of-magnitude fewer trainable parameters and lower optimizer state. This reduces memory and networking overhead during training and checkpointing.



In our setup, adapters attach to attention projections (q/k/v/o) and MLP projections (gate/up/down), with rank $r = 32$ to balance capacity and speed, while$\alpha = 2r$ scaling $\alpha = 2r$ to stabilize training. We pair LoRA with gradient checkpointing and mixed precision to sustain long sequences, and with efficient kernels (e.g., Flash-style attention, Unsloth) to keep VRAM within a single-GPU budget.

2.6 GUI Design

We implemented the medical agent as a conversational web app. The interface is built on Gradio's ChatInterface. User will interact with the system by either typing their symptoms or select a suggested symptom chip.



After that, the agent steams its reasoning in real time and then presents a concise final diagnosis. Each response includes 2 collapsible panels. First, **Contexts,** which show the retrieval snippets the model considered by using **Retrieval Augmented Generation (RAG).** It grounds the model's reasoning in trusted medical sources before it answers, cutting hallucinations and improving safety. Second, **Thinking**, this part reveals the full step by step chain of thought with a timestamp indicating how long reasoning took.

For example, user's input is: "I have a fever, stiff neck and sore eyes from bright lights". The agent will then make a chain of thought e.g. Alright, so based on the... It will also use **RAG** to ground the reasoning with the medical sources e.g. Reference 2: I have been running a fever of .... After all, it will give the final diagnosis and give a recommendation about "It is crucial to consult healthcare professionals...".

Chatbot

I have a fever, stiff neck and sore eyes from bright lights.

▼ Contexts

Reference 1: Question: Hello About 2 weeks ago I went into a minute clinic bc I had pain all over(achy) and a slight fever. She told me I had a sinus infection and gave me meds to take. Just a few days ago my symptoms have come back but worse. It hurts to open my eyes and I was running a fever of 101.8 last night. Do I have the flu? Answer: Hi, see flu is not diagnosis but it's a symptom complex and different from common cold. If you're having difficulty to open eye with fever, then It could be due to conjunctivitis, keratitis usually associated with upper respiratory tract infection or sore throat, or common cold need to see Ophthalmologist if u really have hurting to open the eye with other eye symptoms otherwise physician would be better.

Reference 2:
Question: I have been running a fever of 100.6.I passed out yesterday morning and have a major headache since.It was hard to even open my eyes all the way last night. Temperature is back to normal today but my head feels like it will pop off my head from the pain.
Answer: With this limited information, the best case scenario is that this may be a bad sinus infection. The worse case scenario is that it may be an infection related to the brain or eyes. Sinus infections are much more common and much more likely an explanation. At least statistically speaking. A sound advice would be to see your doctor. After asking you a few questions and examining you, a more detailed explanation may become available. They would definitely try to rule out the worst case scenario diagnosis and let you know if you need further testing or treatment with antibiotics. If the fever continues, I am afraid you will not be in a position to sit it out. You will have to see a doctor at some point.

Reference 3:
Question: Hello I am 20 years old and the other night in bed I woke up suffering severe chills. Since then my skin mainly on my back is very sensitive/sore and is getting worse. Also my eyes have become very sensitive to light and hurt. Do you have any ideas? Its very debilitating.
Answer: As per your explanation here, you have symptoms like soreness in back, as well your eye is sensitive with light. From both this

What seems to be the problem, please describe your symptoms

Chatbot

▶ Contexts

▼ Thinking (took 21.92 seconds)

Alright, so based on what happened recently—fever, stiff neck, sore eyes when exposed to bright light—I'm thinking about something serious, possibly meningococcemia or maybe an infectious mononucleosis. Those symptoms together ring alarm bells. Let's break them down.

Firstly, those signs like fever, stiff neck, and sensitivity to light are classic indicators pointing towards bacterial meningitis, which requires immediate attention because it can lead to serious health issues quickly. Meningococcemia is another possibility linked to similar symptoms and needs prompt action too.

Sore eyes, especially with sensitivity to light, might imply inflammation around the optic nerve—a condition known as idiopathic intracranial hypertension or papilloedema—which can occur alongside these types of infections. However, without additional neurological tests confirming increased pressure inside the skull, we can't conclusively say it's causing trouble yet.

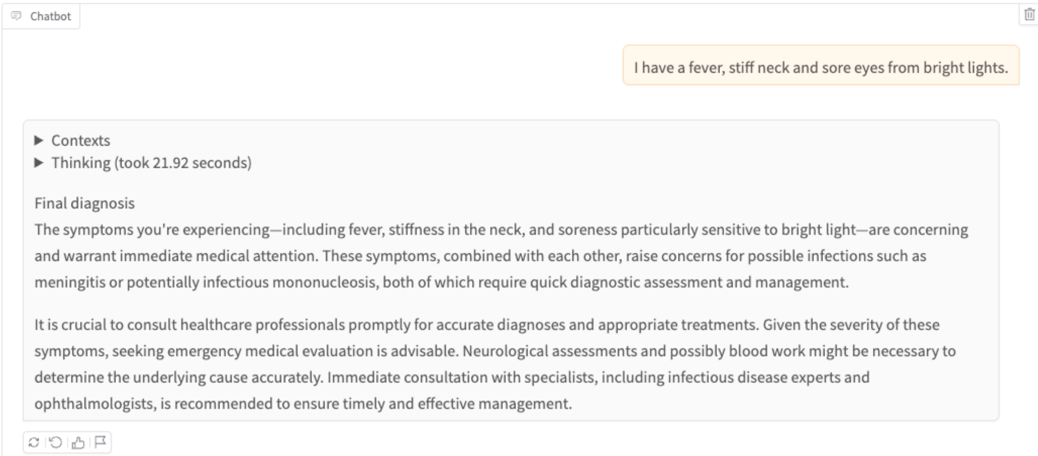
The combination makes me think it's critical to seek urgent medical evaluations immediately rather than waiting until things resolve on their own. This includes neurologists who specialize in such conditions and perhaps hematologists if suspected by blood-related causes.

Given the urgency and seriousness of these symptoms, rushing to hospital for imaging studies like an MRI/CT scan and laboratory tests seems essential. These tests will confirm whether we're dealing with an acute infection requiring intervention promptly.

Final diagnosis
The symptoms you're experiencing—including fever, stiffness in the neck, and soreness particularly sensitive to bright light—are concerning and warrant immediate medical attention. These symptoms, combined with each other, raise concerns for possible infections such as meningitis or potentially infectious mononucleosis, both of which require quick diagnostic assessment and management.

It is crucial to consult healthcare professionals promptly for accurate diagnoses and appropriate treatments. Given the severity of these

What seems to be the problem, please describe your symptoms



3. Results and Evaluation

We evaluate 2 variants of Qwen3-1.7B-Instruct and Qwen3-1.7B-Base across 3 RL setups:

1. Qwen3-1.7B-Instruct with PPO,
2. Qwen3-1.7B-Instruct with GRPO
3. Qwen3-1.7B-Base with GRPO.

This setup isolates the impact of prior instruction alignment and the choice of reinforcement learning optimizer on medical reasoning quality. The Instruct model incorporates conversational and structural priors that often promote compliance but may be bias reasoning style, whereas the Base model serves as an unaligned backbone.

3.1. Experiment settings

Configurations for PPO baseline		**Qwen3-1.7B-Instruct**
SFT Training	**No. epochs**	3
	Max sequence length	4096
	Learning rate	5e-6
PPO	**No. epochs**	3
	Temperature	1.0
	Total episodes	1024
	Learning rate	5e-7
	Warmup ratio	0.05
	KL coefficient	**0.03**
Configurations for GRPO		**Qwen3-1.7B Instruct and Base**
	Max sequence length	2048

LoRA Settings	**Rank**	32
	Alpha	64
	Dropout	0.1
	Target Modules	'q_proj', 'k_proj', 'v_proj', 'o_proj', 'gate_proj', 'up_proj', 'down_proj'
SFT Training	**Num epochs**	3
	Batch size	32
	Optimizer	AdamW with 5 warmup steps
	Learning rate scheduler	Cosine with 2e-4 learning rate
GRPO	**vLLM sampling params**	min_p = 0.1, top_p = 1.0, top_k = -1
	No. generations	4
	Temperature	1
	Num epochs	1
	Batch size	64
	Optimizer	AdamW with 5 warmup steps
	Learning rate scheduler	Cosine with 1e-5 learning rate
	Loss type	dr_grpo

3.2. Training Results

Both models successfully converged during GRPO optimization. The **train/loss** curves decrease from 0.06–0.07 at the start to **≤ 0.02** by the end for both models, with one transient spike for the Base run (~**0.105** at ~1.55k steps), followed by a quick recovery to the pre-spike trend. Meanwhile, **train/reward** steadily increases from ~**3.3 to 4.3** (Instruct) and **3.4 to 4.4** (Base), confirming that policy updates consistently enhance the reward signal rather than cause oscillations. The **train/reward_std** increases in the early phase and then plateaus around **3.05 (Instruct)** and **3.15 (Base)**. Under GRPO's group-relative normalization, this is expected: early training broadens exploration across multiple sampled completions per prompt; later, dispersion stabilizes as the policy collapses onto higher-reward modes.

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)

Format-related rewards have effectively saturated for both models. These near-deterministic ceilings indicate that **SFT + LoRA** provided the policy with a strong prior over the **<THINK>…</THINK><ANSWER>…</ANSWER>** protocol; GRPO then sustained that behavior. As a result, the **Average Reward** is primarily influenced by "always-on" structure gains, reaching **4.30 ± 3.14**(Instruct) and **4.40 ± 3.19** (Base), with only a **+0.10** difference between models.

Another key observation is the **Check Answer Correctness** metric, which measures semantic alignment between the model's answer and the ground truth using our **LLM Verifier.** Although both models achieved high structure scores, the **Correctness** scores have a small mean and large variation: **−0.19 ± 3.13** for the Instruct model and **−0.10 ± 3.18** for the Base model. A mean near 0 and **σ ~ 3.1** suggest a **bimodal distribution** of strong positives (up to +5.0) and penalties (down to −2.5), indicating that the model usually either gets the diagnosis exactly right or misses it, with fewer intermediate cases. This explains why the overall average reward increases even when the correctness mean stays near 0: the perfect structural scores boost reward, but factual grounding remains inconsistent.

Completion length differentiates between models. The **Base** model stabilizes **around 383 tokens**, while **Instruct** stabilizes around **636 tokens**. That is roughly a **253-token gap (about**

40% shorter) for Base. Since we used **Dr. GRPO** (length-agnostic objective) and none of the rewards reference sequence length, this difference reflects the model's reasoning style, not reward hacking. Later in this report, we will see that the shorter chains of Base align with its slightly weaker benchmark accuracy; shallow deliberation leaves potential unexplored.

3.3. Testing results

We evaluated both models on 2 standard medical-reasoning benchmarks:

- **MedQA-USMLE** comprises English USMLE-style questions with four options; following common practice, we report results on the held-out test split **(n=1,273)** provided in curated releases (train/test = 10,178/1,273), which target clinical reasoning and factual recall across pre-clinical and clinical subjects. This USMLE subset is widely adopted in broader medical QA suites and is treated as a standard proficiency probe for LLMs.

question	answer	options	meta_info	answer_idx
string · lengths	string · lengths	dict	string · classes	string · classes
66 3.58k	1 230		2 values	4 values
A 23-year-old pregnant woman at 22 weeks…	Nitrofurantoin	{ "A": "Ampicillin", "B": "Ceftriaxone", "C":…	step2&3	D
A 3-month-old baby died suddenly at night while…	Placing the infant in a supine…	{ "A": "Placing the infant in a supine position on a firm…	step2&3	A
A mother brings her 3-week-old infant to the…	Abnormal migration of ventral…	{ "A": "Abnormal migration of ventral pancreatic bud", "B…	step1	A
A pulmonary autopsy specimen from a 58-year-…	Thromboembolism	{ "A": "Thromboembolism", "B": "Pulmonary ischemia",…	step1	A
A 20-year-old woman presents with menorrhagia…	Von Willebrand disease	{ "A": "Hemophilia A", "B": "Lupus anticoagulant", "C":…	step1	D
A 40-year-old zookeeper	Scorpion sting	{ "A": "Aspirin", "B": "Oral	step1	C

- **MedMCQA** is a large-scale **MCQA (**Multiple-Choice Question Answering) benchmark derived from medical entrance examinations **(AIIMS & NEET-PG)**, spanning around **194k** items over **21** subjects and around **2.4k** healthcare topics; we use its official validation split **(n=4,183)** as a stable development/evaluation set.

id	question	opa	opb	opc	opd	cop
string · lengths	string · lengths	string · lengths	string · lengths	string · lengths	string · lengths	class label
36 36	1 1.57k	1 287	1 287	1 286	1 301	4 classes
e9ad821a-c438-4965-9f77-…	Chronic urethral obstruction due t…	Hyperplasia	Hyperophy	Atrophy	Dyplasia	2 c
e3d3c4e1-4fb2-45e7-9f88-…	Which vitamin is supplied from onl…	Vitamin C	Vitamin B7	Vitamin B12	Vitamin D	2 c
5c38bea6-787a-44a9-b2df-…	All of the following are…	Adjustable gastric banding	Biliopancreatic diversion	Duodenal Switch	Roux en Y Duodenal By pass	3 d
cdeedb04-fbe9-432c-937c-…	Following endaerectomy on…	Central aery of the retina	Infraorbital aery	Lacrimal aery	Nasociliary aretry	0 a
dc6794a3-b108-47c5-8b1b-…	Growth hormone has its effect on…	Directly	IG1-1	Thyroxine	Intranuclear receptors	1 b
5ab84ea8-12d1-	Scrub typhus is	Louse	Tick	Mite	Milk	2 c

cop class label	choice_type string · classes	exp string · lengths	subject_name string · classes	topic_name string · lengths
4 classes	2 values	1 22.5k	21 values	3 135
2 c	single	Chronic urethral obstruction because of urinary calculi, prostatic hyperophy, tumors, normal…	Anatomy	Urinary tract
2 c	single	Ans. (c) Vitamin B12 Ref: Harrison's 19th ed. P 640* Vitamin B12 (Cobalamin) is synthesized solel…	Biochemistry	Vitamins and Minerals
3 d	multi	Ans. is 'd' i.e., Roux en Y Duodenal Bypass Bariatric surgical procedures include:a. Vertical…	Surgery	Surgical Treatment Obesity
0 a	multi	The central aery of the retina is a branch of the ophthalmic aery. It is the sole blood supply to…	Ophthalmology	null
1 b	single	Ans. is 'b' i.e., IGI-1GH has two major functions :-i) Growth of skeletal system :- The growth is…	Physiology	null

Performance is reported as **exact-match accuracy**. To identify where improvements happen, we further break down outcomes per item into:

- **Format** (whether the THINK/ANSWER template is correctly emitted),
- **Answer** (a semantically correct final answer, verified by an LLM-based verifier that handles medical aliases), and
- **Both** (correct format and answer at the same time).

Our initial hypothesis was that the Base version might perform better, since instruction tuning can sometimes introduce generalization bias in domain-specific RL. However, the results partially reject this hypothesis.

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%)

Using the same backbone and data, **GRPO outperforms PPO** on both benchmarks, with a larger margin on the more challenging and larger MedMCQA (+6.0%). Furthermore, **Base + GRPO** approaches **exceeds Instruct + PPO** on MedMCQA (+4.38 %), suggesting that group-relative, multi-reward RL can bootstrap medical reasoning from an untuned model. The overall best results are achieved by **Instruct + GRPO** (49.41% MedQA, 46.07% MedMCQA).

Qwen3-1.7B	Correct Count	MedQA_USLME_test		
		Before SFT/RL	After SFT/RL	Improvement
Instruct	Format	646/1273 (50.75%)	**1273/1273 (100.00%)**	+627 (49.25%)
	Answer	375/1273 (29.46%)	**629/1273 (49.41%)**	+254 (19.95%)
	Both	375/1273 (29.46%)	**629/1273 (49.41%)**	+254 (19.95%)

Base	**Format**	531/1273 (41.71%)	1271/1273 (99.84%)	+740 (58.13%)
	Answer	239/1273 (18.77%)	570/1273 (44.78%)	+331 (26.00%)
	Both	239/1273 (18.77%)	570/1273 (44.78%)	+331 (26.00%)

Qwen3-1.7B	**Correct Count**	**MedMCQA_validation**		
		Before SFT/RL	**After SFT/RL**	**Improvement**
Instruct	**Format**	3152/4183 (75.35%)	**4183/4183 (100.00%)**	+1031 (24.65%)
	Answer	1601/4183 (38.27%)	**1927/4183 (46.07%)**	+326 (7.79%)
	Both	1601/4183 (38.27%)	**1927/4183 (46.07%)**	+326 (7.79%)
Base	**Format**	1543/4183 (36.89%)	4180/4183 (99.93%)	+2637 (63.04%)
	Answer	655/4183 (15.66%)	1858/4183 (44.42%)	+1203 (28.76%)
	Both	655/4183 (15.66%)	1858/4183 (44.42%)	+1203 (28.76%)

For **Qwen3-1.7B Instruct** on **MedQA**, answer accuracy climbs from **29.46%** to **49.41%** (+19.95%), and “Both” (correct format and answer jointly) rises identically from **29.46%** to **49.41%**. On **MedMCQA**, Instruct improves from **38.27%** to **46.07%** (+7.79 %), again with “Both” tracking the same gain. Because the format was already high before training (75.35% on **MedMCQA**), most of GRPO’s contribution for Instruct is in turning structured outputs into correct clinical conclusions.

The **Qwen3-1.7B Base** model starts with much lower structural priors and therefore gains more from GRPO’s group-relative signal. On **MedQA**, answer accuracy increases from **18.77%** to **44.78%** (+26.01%), while format improves from **41.71%** to **99.84%** (+58.13%), which almost reaches maximum structural competence. On **MedMCQA**, answer accuracy goes from **15.66%** to **44.42%** (+28.76) and “Format” from **36.89%** to **99.93%** (+63.04%). These changes indicate that GRPO first normalizes the structure and then improves its semantic accuracy, especially when starting from an untuned backbone.

A head-to-head comparison after training shows **Instruct+GRPO** remains slightly ahead: **49.41% vs 44.78%** on **MedQA** (+4.63%) and **46.07% vs 44.42%** on **MedMCQA** (+1.65 pts), with the format effectively tied at the ceiling. Against an equally configured **PPO** baseline on the Instruct backbone, **GRPO** yields consistent gains with **+3.93%** on **MedQA** (**49.41%** vs **45.48%**) and **+6.03 pts** on **MedMCQA** (**46.07%** vs **40.04%**). This indicates that critic-free, group-relative advantages combined with programmable, verifiable rewards address the instability and reward-hacking tendencies observed with single-score **PPO** pipelines.

In conclusion, **GRPO's impact is twofold.** First, it fixes structural errors across models, reducing a noisy source of variance in evaluation. Second, and more importantly, it improves answer correctness slightly for the instruction-tuned model where structure was already strong, and significantly for the base model where there was more room for improvement.

3.4. Limitations

Our study surfaces several constraints that limit the current conclusions. First, the approach remains **sensitive to reward design and hyperparameters** (e.g., weights across strict/soft formats and verifier rewards, scaling choices, generation budgets, temperature). Small adjustments can shift the update signal, occasionally causing regressions or length artifacts despite Dr.GRPO. Second, while GRPO consistently **achieves format compliance**, **semantic correctness remains inconsistent**: improvements are significant but vary across topics and item difficulty, indicating persistent gaps in factual grounding and domain calibration. Third, **compute costs** are substantial: group sampling (4–8 completions per prompt), verifier inference, and RL updates increase both wall-clock time and VRAM use even with LoRA and gradient checkpointing. Finally, our **training period was limited**, prioritizing feasibility over full convergence; some results suggest additional potential with longer runs and curriculum schedules.

4. Conclusion

There is a large unmet need for health services in many parts of the world, with over half the world's population lacking access to adequate healthcare, due to human, infrastructure and affordability constraints. This project seeks to contribute to improve human health outcomes using AI. Here, we developed a medical doctor agent to respond to a patient's description of medical symptoms, by 1) selecting a set of relevant contexts, 2) developing a complex chain of thought and 3) thereby reaching a diagnosis of the symptoms. The doctor agent must be capable of accurate, explainable, and format-constrained clinical reasoning.

To develop the doctor agent, we started with a small LLM (Qwen3 1.7B) due to compute constraints and used a 2-stage process of supervised fine-tuning to establish output formatting standards, followed by reinforcement learning with GRPO and LoRA to enhance clinical reasoning and answer correctness. The innovations include a multi-reward system with strict format, soft format and medical accuracy components to the reward. Medical accuracy assessment is via an LLM-based verifier.

We ran experiments comparing the results between PPO and GRPO methods and the results between the base and instruct versions of the starting model (Qwen3 1.7B). The best model (Qwen3 Instruct with GRPO) achieved a ~20% and 7.8% improvement on the MedQA and MedMCQA standard benchmark tests respectively, compared to the pre-trained model. It also outperformed the PPO trained model by 4% and 6% on the tests, respectively.

We see multiple ways to improve reliability and generality. First of all, e**xpand model families**: run the pipeline on Gemma, Mistral, and Llama variants to evaluate architecture effects and data priors, with controlled LoRA ranks and identical GRPO settings. In addition, s**trengthen the verifier**: ensemble or calibrate LLM verifiers, add rules for aliases and negations, and include uncertainty thresholds for abstention; explore lightweight retrieval-augmented verification for fact-checking before reward assignment. Finally, i**mprove rewards and training: conduct** ablations on reward weights and scaling (group vs. batch), add penalties for unsupported claims, and test curriculum GRPO (format -> reasoning -> clinical constraints).

5. References

Australian Institute of Health and Welfare. (2016). *Coordination of health care: experiences of barriers accessing health services among patients aged 45 and over.* https://www.aihw.gov.au/getmedia/d864d80b-22cb-4356-a1f6-24c2f350004b/aihw-chc-4.pdf?v=20230605165931&inline=true

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., & McKinnon, C. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Department of Health, Disability and Ageing. (2025, Oct 25). *About primary care*. Department of Health, Disability and Ageing. https://www.health.gov.au/topics/primary-care/abouts.

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. *arXiv preprint arXiv:2412.18925*.

Feng, Y., Wang, J., Zhou, L., Lei, Z., & Li, Y. (2025). Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. *arXiv preprint arXiv:2505.19630*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR, 1*(2), 3.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences, 11*(14), 6421.

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. *arXiv preprint arXiv:2503.20783*. https://doi.org/10.48550/arxiv.2503.20783

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. Conference on health, inference, and learning.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. 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., & Wu, Y. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*. https://doi.org/10.48550/arxiv.2402.03300

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

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., & Lv, C. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yu, K., Yu, C., Zhang, T., Zhao, X., Yang, S., Wang, H., Zhang, Q., & Xu, Q. (2025). Temporal separation with entropy regularization for knowledge distillation in spiking neural networks. Proceedings of the Computer Vision and Pattern Recognition Conference.