Open-Ended Clinical Text Generation for Acute Care: Applying Reinforcement Learning with Clinically Grounded Rewards

Minjia Wang* MINJIAWANG@G.HARVARD.EDU
Luyang Luo* LUYANG_LUO@HMS.HARVARD.EDU
Sung Eun Kim
Harvard University, United States

Fang Cao
David A. Kim
Stanford University, United States

Pranav Rajpurkar
Harvard University, United States

Abstract

Acute care clinicians generate critical clinical text—diagnoses, treatment plans, discharge instructions—under time pressure where errors can be life-threatening. Large proprietary AI models raise privacy concerns, while smaller models lack clinical quality. We extend reinforcement learning with verifiable rewards (RLVR) to open-ended clinical text generation using two generalizable reward patterns: equivalence-based rewards for medical synonymy and diagnosis matching, as well as rubric-based rewards for multi-dimensional quality assessment. Using group relative policy optimization, we trained compact 7–8 billion parameter models on diagnosis generation (MIMIC-III), discharge instructions (DischargeMe), and treatment planning (MTSamples). Trained models achieve clinical quality across tasks (best results: F1 0.48, 4.28/5.0, 4.47/5.0 respectively), matching or surpassing the performance of large proprietary GPT-based models, while enabling on-premise deployment, sub-second inference, and full privacy. Physician review confirmed superior content comprehensiveness and fewer dangerous errors versus base models. This demonstrates a practical pathway for deploying clinical text generation in acute care with generalizable reward design patterns.

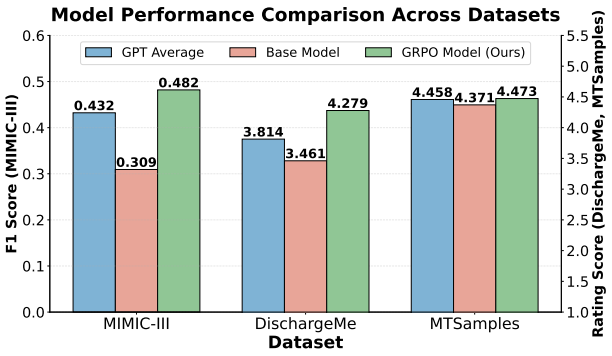

Figure 1: **Performance comparison across three medical text datasets.** MIMIC-III uses F1 score for information extraction (left y-axis); DischargeMe and MTSamples use rating scores for text generation quality (right y-axis). GRPO training consistently improves over base models, approaching or exceeding GPT performance.

Data and Code Availability All datasets are publicly available: MIMIC-III (Johnson et al., 2016) and DischargeMe (Xu et al., 2024) via PhysioNet (CITI credentialing required); MTSamples at https://mtsamples.com. Code: https://github.com/rajpurkarlab/Clinical-RLVR.

Institutional Review Board (IRB) This research uses only publicly available, de-identified datasets (MIMIC-III, DischargeMe, and MTSamples) and does not involve direct human subjects research. Hence, IRB approval was not required for this study.

* These authors contributed equally

1. Introduction

Emergency department (ED) and intensive care units (ICU) require clinicians to rapidly synthesize complex patient data into accurate diagnoses, actionable treatment plans, and clear discharge instructions—cognitive tasks where errors can be immediately life-threatening. Large proprietary models (100B+ parameters) that power current state-of-the-art systems require cloud infrastructure, impose per-query costs prohibitive for high-volume clinical use, and raise data privacy concerns (Tordjman et al., 2025; Sandmann et al., 2025). These factors have limited AI integration into clinical workflows despite clear potential for benefit in high-acuity settings like the ED and ICU.

Open-Ended Clinical Text Generation as a Practical AI Task. Clinical reasoning outputs in acute care are inherently open-ended: diagnosis lists must capture the right conditions from infinite possibilities, treatment recommendations must be tailored to individual patient circumstances, and discharge instructions must communicate complex medical information accessibly. While reinforcement learning with verifiable rewards (RLVR) has driven gains in domains with programmatically checkable correctness—mathematics (Wang et al., 2025; Zhang and Zuo, 2025), coding (Robeyns and Aitchison, 2025), and classification (Lu et al., 2025; Dai et al., 2025)—and in closed-ended medical tasks like multiple-choice questions (Zhang et al., 2025a; Qiu et al., 2025; Pan et al., 2025; Su et al., 2025; Zhang et al., 2025b) or structured classification (Lin et al., 2025; Kim et al., 2025; Liao et al., 2025; Jhaveri et al., 2025), extending these methods to open-ended clinical text has been limited. The core challenge is verification: free-text clinical outputs lack ground-truth correctness signals, yet this free-text reasoning constitutes the cognitive core of acute care practice.

Contribution: Applying RLVR to Open-Ended Clinical Text Generation. We demonstrate a practical application of reinforcement learning to open-ended clinical text generation in acute care settings. Our contribution establishes that compact models (7–8B parameters) trained with clinically grounded rewards can generate diagnostic lists, treatment plans, and discharge instructions that meet clinical quality standards while enabling on-premise deployment. Specifically, we address three key challenges: (1) *verification for open-ended outputs*—how to provide training signals when clinical text has no single correct answer; (2) *practical deployment*—achieving clinical quality with models 10× smaller than proprietary alternatives; and (3) *generalizability*—reward design patterns that extend beyond specific tasks to the broader space of clinical documentation.

We developed two reward approaches matched to different output types: equivalence-based rewards for diagnosis generation that evaluate whether predicted and reference diagnoses represent the same clinical concepts, accounting for synonyms (e.g., "heart attack" vs. "myocardial infarction"), abbreviations (e.g., "COPD" vs. "Chronic Obstructive Pulmonary Disease"), and varying specificity (e.g., "pneumonia" matching "bacterial pneumonia"); and rubric-based rewards for discharge instruction generation and treatment planning that score outputs on accuracy, completeness, and clarity using criteria similar to how attending physicians evaluate trainee notes. Using Group Relative Policy Optimization (GRPO)—an RL algorithm that learns by comparing multiple candidate outputs for the same input—we trained 7–8 billion parameter models on two acute care tasks.

The resulting models achieve clinical quality standards while enabling practical deployment: best model variants reach F1 of 0.48 on diagnosis generation, 4.28/5.0 on discharge instructions, and 4.47/5.0 on treatment plans, with sub-second inference on standard GPU infrastructure. Trained models show no statistically significant differences in performance compared to state-of-the-art GPT models on two of three tasks. Physician review confirmed more comprehensive clinical content and fewer dangerous errors compared to base models. This work establishes open-ended clinical text generation as a viable application for compact models in time-critical acute care settings.

2. Preliminaries: Group Relative Policy Optimization

We briefly review Group Relative Policy Optimization (GRPO) (Shao et al., 2024), the RL algorithm we build on. GRPO is a standard post-training component; our primary contribution is the clinically grounded reward design that makes it applicable to open-ended clinical documentation (Section 3.1).

2.1. Notation and Problem Setup

Throughout, π_θ denotes the LLM policy—where π is the probability distribution over output sequences and θ the trainable model weights—x the input clinical record, y the generated output (diagnosis list or clinical document), and $r(x, y)$ the task-specific reward function evaluating output quality. We cast each clinical documentation task as conditional sequence generation. Given an input clinical context x (e.g., compiled notes, labs, medications), the policy model π_θ—the LLM—produces an output sequence $y = (y_1, \ldots, y_T)$ (e.g., a diagnosis list or a treatment plan) with probability

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})\,. \tag{1}$$

We optimize a Kullback-Leibler-regularized (KL-regularized) expected reward objective to maximize $r(x, y)$ while limiting policy drift from a fixed reference model π_{ref} (the base model) via a coefficient β; the full objective is given in Section 2.2.

2.2. GRPO Training

GRPO is a Proximal Policy Optimization (PPO)-style (Schulman et al., 2017) method that avoids training an explicit value function by estimating advantages via within-prompt group comparisons. For each input x in a minibatch, we draw a group of G candidate completions from the current (frozen) behavior policy $\pi_{\theta_{\text{old}}}$:

$$y^{(1)}, \ldots, y^{(G)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \qquad r_i = r\Big(x, y^{(i)}\Big)\,. \tag{2}$$

We compute a group-relative advantage by normalizing rewards within the group. Here $\hat{A}_i$ is the advantage of completion i—how much better or worse it is relative to the group—and μ_r and σ_r are the group mean and standard deviation of rewards, respectively:

$$\hat{A}_i = \frac{r_i - \mu_r}{\sigma_r}, \qquad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j, \tag{3}$$

$$\sigma_r = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_r)^2}. \tag{4}$$

Because our verifiers provide outcome-level rewards, we apply the same advantage to all tokens of a sampled completion:

$$\hat{A}_{i,t} = \hat{A}_i, \qquad \forall\, t \in \{1, \ldots, T_i\}. \tag{5}$$

Clipped Policy Gradient Objective GRPO performs a clipped policy-gradient update analogous to PPO. For each sampled token $y_t^{(i)}$ under the prefix $s_{i,t} = (x, y_{<t}^{(i)})$, we form the per-token likelihood ratio

$$\rho_{i,t}(\theta) = \frac{\pi_\theta\Big(y_t^{(i)} \mid s_{i,t}\Big)}{\pi_{\theta_{\text{old}}}\Big(y_t^{(i)} \mid s_{i,t}\Big)}. \tag{6}$$

The clipped surrogate objective is

$$L_{\text{clip}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\Big(\rho_{i,t}(\theta)\,\hat{A}_{i,t}, \; \text{clip}(\rho_{i,t}(\theta),\, 1-\epsilon,\, 1+\epsilon)\,\hat{A}_{i,t}\Big) \tag{7}$$

where ϵ is the PPO clipping parameter.

KL Regularization To regularize against undesired policy drift, we add a KL penalty to the reference model computed at the same token-level conditioning states $s_{i,t}$:

$$L_{\text{KL}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\text{KL}(\pi_\theta(\cdot \mid s_{i,t}) \,\|\, \pi_{\text{ref}}(\cdot \mid s_{i,t}))\,. \tag{8}$$

The final GRPO objective is

$$\max_\theta \; L_{\text{clip}}(\theta) - \beta\, L_{\text{KL}}(\theta), \tag{9}$$

which we optimize using gradient-based updates. The group-relative normalization makes each update depend on the relative quality among sampled candidates for the same clinical input, improving training stability without requiring a separate critic or value head like in Wu et al. (2025).

3. Methods

We present a methodology for applying reinforcement learning to open-ended clinical text generation in acute care settings. The GRPO training formulation is described in Section 2; this section focuses on the contribution: clinically grounded reward design patterns (Section 3.1), model training setup, and validation across three acute care documentation tasks (diagnosis generation, discharge instructions, treatment planning) using public datasets with both automated metrics and physician review. This methodology demonstrates how reinforcement

learning can be extended from closed-ended medical question-answering tasks to the free-text generation central to clinical practice. Figure 2 illustrates the complete pipeline from problem setup through reward design to model training.

3.1. Reward Design

Diagnosis generation For disease diagnosis, the model outputs a free-form list of conditions that is compared with the ground-truth diagnosis set derived from ICD codes. Because clinically equivalent diagnoses may differ in wording, abbreviation, or granularity, we used an LLM judge (GPT-4.1)[1] to determine equivalence between predicted and reference diagnoses. The judge was instructed to account for medical synonyms, abbreviation expansion, and reasonable variations in specificity (for example, allowing a general diagnosis to match a more specific subtype when clinically appropriate). Using the judged matches, we computed set-overlap metrics (including Jaccard similarity and F1) and used the overlap signal as the reward, which penalizes both over-generation and omissions. To discourage reward hacking, we included validation checks on the judged matches; invalid match structures resulted in zero reward for the example. The prompt for the LLM judge can be found in Appendix C.

Discharge instruction and treatment plan generation For discharge instructions and treatment plan generation, outputs are longer-form prose where exact string matching is neither possible nor desirable. We therefore defined rubric-based rewards scored by GPT-4.1 on a five-point scale. Discharge instructions were scored for completeness (coverage of clinically important information), correctness (absence of clinically harmful or misleading content), and readability (clarity relative to the reference). Treatment plans were scored for accuracy (alignment with clinical guidelines), completeness (inclusion of salient details), and clarity (ease of comprehension for clinicians). The scalar reward for RLVR was the mean of the rubric dimension scores for the given task. These dimensions follow the multidimensional clinical evaluation framework of MedHELM (Bedi et al., 2025): diagnosis rewards are entirely accuracy-based (semantic equivalence against ICD ground truth), and two-thirds of the generative task reward dimensions directly penalize clinically harmful content—inaccuracies, contraindications, and omissions—making the reward signal clinically grounded rather than stylistic. The prompts for the LLM judge can be found in Appendix C.

3.2. Model Architecture and Training

We trained compact instruction-tuned language models in the 7–8B parameter range using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Backbone models included Qwen2.5-7B-Instruct (Qwen Team et al., 2024), Qwen3-8B (Yang et al., 2025), and Ministral-8B-Instruct (Jiang et al., 2023)[2]. These architectures were selected for their strong instruction-following capabilities and feasibility for on-premise deployment on standard GPU infrastructure (e.g., single A100 or H100 GPU).

To validate whether compact models can achieve clinical quality standards for open-ended text generation, we compared three conditions under identical prompts: (1) base backbone models before GRPO training, (2) GRPO-post-trained models using our clinically grounded rewards, and (3) substantially larger proprietary models (GPT-4o, GPT-4.1, and GPT-5, estimated at 100B+ parameters (Li, 2026)) representing current state-of-the-art systems.[3] This design isolates the effect of reward-based training while contextualizing performance against systems widely deployed in healthcare.

3.3. Training configuration

GRPO training used a learning rate of 1×10^{-6} with a constant schedule and a warmup ratio of 0.1. We sampled 12 completions per prompt for group comparisons, used an effective batch size of 64, and trained for one epoch. Training used bfloat16 mixed precision and gradient checkpointing. We set the KL regularization coefficient to $\beta = 0.04$ (Shao et al., 2024) and the PPO clipping parameter to $\epsilon = 0.2$. We set a maximum prompt length of 8192 tokens and a maximum completion length of 2048 tokens to accommodate long clinical contexts.

1. API version `2025-03-01-preview`.

2. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

3. All three proprietary models were accessed via API version `2025-03-01-preview`.

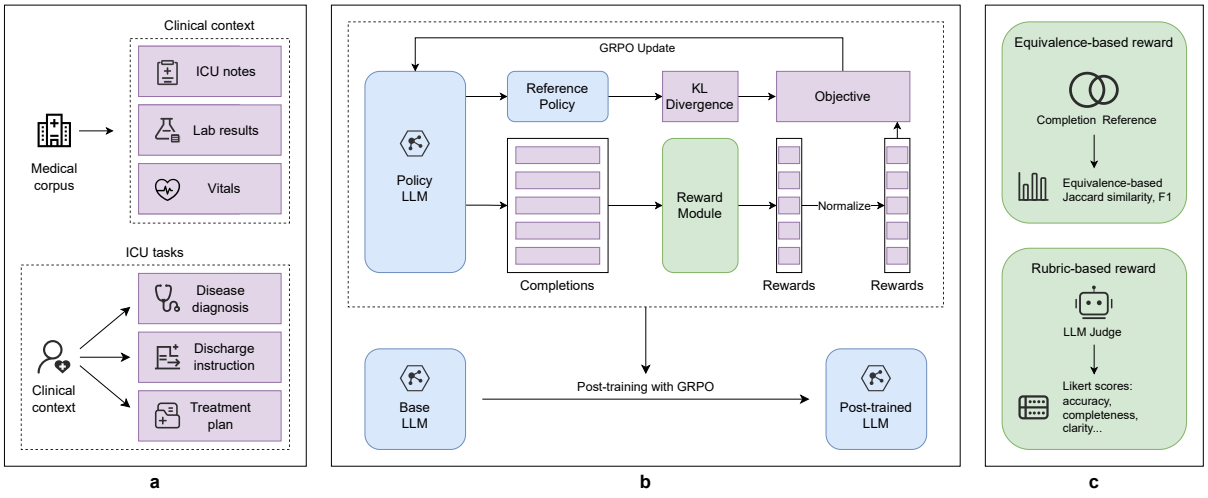

Figure 2: **RLVR training pipeline with GRPO for free-form acute-care clinical text generation. a. Problem Setup.** Clinical context (ICU notes, laboratory results, and vital signs) is constructed from a large medical corpus and processed into ICU tasks, including disease diagnosis and treatment plan generation. **b. Post-training framework with GRPO.** Training first samples a group of candidate free-text completions per prompt. A task-specific reward module scores each completion, and rewards are normalized within the group to produce relative advantages. GRPO then updates the policy using a clipped policy-gradient objective with KL regularization to a fixed reference policy. **c. Reward designs.** Equivalence-based rewards are used for diagnosis generation, in which an LLM judge aligns predicted and reference diagnosis items and scores set-level overlap (e.g., Jaccard or F1). Rubric-based rewards are used for longer-form outputs (discharge instructions and treatment plans), in which an LLM judge assigns Likert-style ratings across clinical quality dimensions (e.g., accuracy, completeness, and clarity) and aggregates them into a scalar reward.

4. Results

4.1. Datasets

We applied our methodology to three open-ended clinical text generation tasks spanning the acute care workflow. All datasets are publicly available, supporting reproducibility.

MIMIC-III (Johnson et al., 2016) is a de-identified critical care database with ICU notes and ICD-coded diagnoses available through PhysioNet; we use it for the *disease diagnosis* task. We randomly sampled 10,000 hospital admissions, split into 5,000 training and 5,000 test samples. For each admission, we compiled clinical notes, laboratory events, microbiology events, and prescriptions into comprehensive contexts. Ground truth diagnosis lists were derived from ICD codes, mapped to standard terms.

DischargeMe (Xu et al., 2024) links comprehensive MIMIC-III clinical contexts with expert-authored discharge instructions; we use it for the *discharge instruction generation* task. We randomly sampled 10,000 context–instruction pairs, split into 5,000 training and 5,000 test instances. To prevent data leakage, discharge instructions and hospital course sections were excluded from inputs.

MTSamples [4] is a publicly available collection of medical transcription samples covering multiple specialties; we use it for the *treatment plan generation* task. The processed dataset consists of 215 training and 214 test samples, where inputs are compiled clinical notes and outputs are treatment plans.

4. https://mtsamples.com

4.2. Evaluation Protocol

Models were evaluated on the aforementioned held-out test sets using the same LLM judge-based metrics employed during training, ensuring consistency between training objectives and evaluation criteria. We compared model performance using Wilcoxon signed-rank two-sided tests on per-example scores. A significance level of $p \leq 0.05$ was used. To address potential judge bias, we applied identical prompts across all models and included the judge model (GPT-4.1) as a generation baseline, enabling detection of self-preference effects. To further assess whether improvements reflect genuine quality gains rather than optimization toward GPT-4.1's scoring style, we repeated the full evaluation using GPT-5 as an independent judge; results and analysis are reported in Appendix A. We also conducted a manual review of 30 randomly sampled test cases per task (90 pairs in total), performed by a licensed physician with clinical experience in acute care. The evaluation used a non-blinded, side-by-side comparison format; the physician provided free-form clinical comments focusing on completeness, accuracy, and identification of potentially dangerous errors. This review was designed as a qualitative clinical sanity check rather than a formal inter-rater reliability study.

4.3. Quantitative Results

Figures 3 and 4 show performance for open-ended diagnosis generation, discharge instruction generation, and treatment plan generation, respectively. Figure 3 uses bars with error bars showing 95% confidence intervals; Figure 4 uses lollipop plots with points indicating mean scores across evaluation dimensions.

Generating diagnosis lists from ICU data In the ICU, critically ill patients often present with multiple concurrent conditions. Clinicians must synthesize diverse clinical data—notes, laboratory values, microbiology results, medications—into comprehensive problem lists, representing diagnostic reasoning under uncertainty with incomplete and evolving data where verification requires understanding clinical semantics rather than exact string matching.

GRPO training with equivalence-based rewards produced substantial improvements. Ministral-8B-Instruct improved from 0.31 to 0.48 F1 (+0.17), significantly outperforming GPT-4.1 (0.43), GPT-4o (0.44), and GPT-5 (0.43). The improvement was driven primarily by increased precision while maintaining reasonable recall. Base models like Qwen3-8B exhibited high recall but very low precision, indicating over-generation of spurious diagnoses; GRPO training corrected this imbalance. This precision gain is clinically important—false-positive diagnoses could trigger unnecessary interventions or delayed treatment of actual conditions.

Generating patient-facing discharge instructions Safe transitions from hospital to home require translating complex medical information into actionable patient guidance. Poor discharge communication contributes to adverse events and preventable readmissions, making clarity, completeness, and safety paramount for patient-facing communication.

GRPO training with rubric-based rewards substantially improved discharge instruction quality across all base models. Ministral-8B-Instruct improved from 3.46 to 4.28 average score (+0.82), approaching GPT-4.1 (4.42) while exceeding GPT-4o (4.04) and GPT-5 (2.99). The improvement was particularly striking given base model performance: base Ministral-8B scored only 3.46, but GRPO training elevated it to near-GPT-4.1 levels. The largest gains appeared in completeness (2.77→3.83) and correctness (3.65→4.35), with readability improving from 3.96 to 4.66—exceeding GPT-4.1's readability score of 4.02. This indicates that reward-guided training encouraged models to include critical safety information while improving clarity and patient-centered communication.

Generating treatment plans for clinical decision-making Therapeutic decision-making requires synthesizing patient-specific factors—medical history, current presentation, comorbidities, and contraindications—into actionable interventions that integrate evidence-based guidelines with individual patient circumstances such as medication dosages, monitoring parameters, and follow-up intervals.

GRPO training with rubric-based rewards yielded consistent improvements across all base models. Qwen3-8B improved from 4.37 to 4.47 average score (+0.10), matching GPT-4.1 (4.47), GPT-5 (4.46), and GPT-4o (4.44). The largest relative gains appeared in completeness (4.33→4.49), encouraging models to include clinically salient details that base models tend to omit. Accuracy also improved (4.05→4.19), shaping outputs toward guideline-concordant recommendations, while clarity remained stable (4.74→4.75).

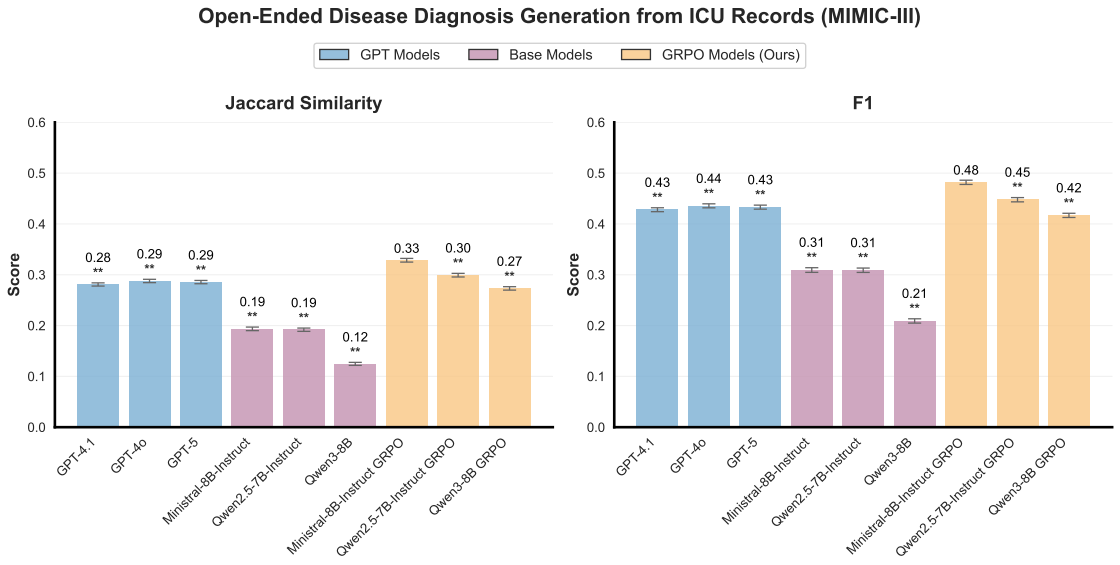

Figure 3: **Open-Ended Diagnosis Generation from ICU Records (MIMIC-III).** Bar plot showing performance on generating diagnosis lists from comprehensive ICU admission data. Bars indicate mean scores for Jaccard similarity and F1 across GPT models, base open-weight models, and GRPO-trained models; error bars show 95% confidence intervals (higher is better). Statistical annotations compare models against the best GRPO-trained model; ** indicates statistically significant difference ($p<0.01$), ns indicates no significant difference.

4.4. Qualitative Analysis

Automated metrics capture aggregate performance, but clinical deployment requires that improvements manifest in individual cases as better recommendations and fewer dangerous errors. We reviewed model outputs to assess whether quantitative gains reflected meaningful clinical differences. *Representative cases with base model and GRPO model outputs are presented in Appendix B.*

More specific and comprehensive recommendations GRPO models consistently offered more specific details for treatment recommendations, including additional therapies, doses, and side effects. In case 21, a patient with STEMI admitted to the ICU, the GRPO model identified HFmrEF and correctly recommended GDMT (Heidenreich et al., 2022), including B-blocker initiation within 24-48 hours and cardiac rehabilitation on discharge—interventions with proven mortality benefit (Rouleau et al., 2024). The base model failed to identify these evidence-based treatment options or provide timing guidance. In case 16 (acne vulgaris), the GRPO model correctly identified prescription options including tetracyclines and isotretinoin, while the base model suggested only over-the-counter therapies. In case 17 (severe onychomycosis), only the GRPO model identified nail debridement.

The models also diverged in medication management: in cases 13 (psychiatric medication interaction) and 6 (propranolol causing bradycardia), the base model recommended stopping medications without alternatives, while the GRPO model provided appropriate substitutions (bupropion and gabapentin, respectively). For discharge documentation, GRPO models demonstrated superior organization: in case 74 (elderly patient with brain hemorrhage), the GRPO model included Discharge Diagnosis and Discharge Condition subsections before discharge instructions, correctly identifying primary and secondary diagnoses. The base model lacked these crucial structural elements—a pattern observed across cases 67, 80, 81, and 86.

More accurate diagnostic capture GRPO training with equivalence-based rewards improved diagnostic accuracy. In case 22, GRPO decisively iden-

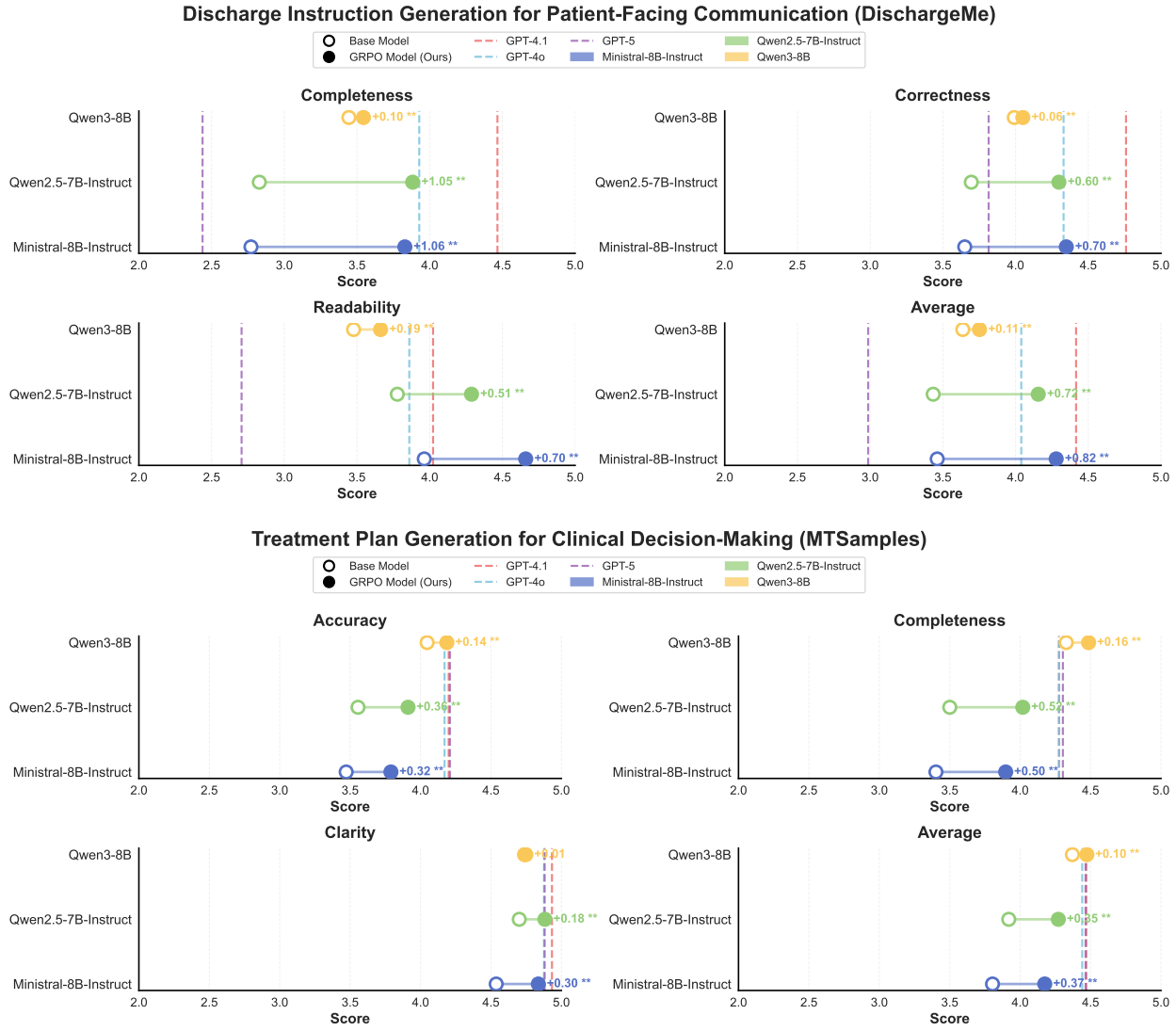

Figure 4: **Open-Ended Clinical Text Generation. Top:** Lollipop plot showing performance on discharge instruction generation (DischargeMe). **Bottom:** Lollipop plot showing performance on treatment plan generation (MTSamples). Points indicate mean scores across evaluation dimensions for base open-weight models, and GRPO-trained models (higher is better). Dotted lines show the performance of the GPT models. Statistical annotations compare models against the best GRPO-trained model; ** indicates statistically significant difference ($p<0.01$), ns indicates no significant difference.

tified acute diverticulitis as the working diagnosis, whereas the base model struggled to identify any leading diagnosis. Similarly, in case 31, the base model did not identify the root problem underlying the patient's symptoms (ALL), whereas the GRPO model was able to list this as the first problem in its problem list. In case 33, a case of new onset gait abnormalities, the GRPO model demonstrated a closer resemblance to the ground truth. The GRPO model identified mood disorders and ataxic gait as symptoms/diagnoses, which were not present with the base model's conclusions. Case 43 demonstrates the com-

prehensiveness of the GRPO model, with an accurate list of 21 diagnoses as compared to 6 in the base model for a critically ill ICU patient, including capturing respiratory failure, bowel perforation, and coagulopathy, all of which were omitted by the base model. The visible improvements of the GRPO model also extend to preparing discharge paperwork. For the discharge instructions for a patient that had died in the hospital after transitioning to comfort care (case 61), the base model failed to identify the patient was deceased and instead provided recommendations on discharge medications on a patient that had died. In contrast, the GRPO model successfully identified the patient was deceased and tailored discharge instructions to focus on supporting the family emotionally ("Your loved one passed away while under our care. We deeply regret this loss and extend our condolences to you and your family").

Fewer potentially catastrophic errors Most critically, GRPO training reduced dangerous recommendations. In case 22, the GRPO model, having identified acute sigmoid diverticulitis as the leading diagnosis, did not recommend colonoscopy, as this has a higher risk of acute bowel perforation (Peery et al., 2021) in the acute setting. In contrast the base model did recommend colonoscopy without specifying timing. This is a recommendation with potentially fatal consequences if colonoscopy was done acutely. Similarly, in case 12, in which a young man had a substantial hand laceration requiring suturing, the models diverged in their recommendation as to the timeframe in which subsequent soaking/swimming would be appropriate. The base model recommended waiting 48-72 hours before soaking, whereas the GRPO model recommended awaiting 7-10 days. Expert opinion states that after 48 hours, traumatic lacerations should not be soaked (Singer et al., 1997), again demonstrating a factual error in which the base model's recommendations could have catastrophic consequences (e.g. loss of hand function in a young patient).

5. Discussion

Generalizable Reward Design for Clinical Documentation The key contribution is demonstrating that clinically grounded reward patterns (equivalence-based for lists, rubric-based for prose) generalize beyond the specific tasks studied. Equivalence-based approaches extend to any list-based clinical output (medication lists, procedure lists, problem lists), while rubric-based approaches adapt to any prose documentation task (progress notes, procedure notes, clinical summaries). These patterns provide a reusable framework for applying reinforcement learning to clinical documentation where verification requires domain expertise rather than exact correctness.

Notably, trained models matching or exceeding the judge model (GPT-4.1) on two of three tasks under identical scoring suggests genuine capability gains rather than stylistic alignment to judge preferences (Wataoka et al., 2025; Panickssery et al., 2024)—the models learned better clinical content generation, not judge mimicry. To verify this directly, we re-evaluated all models using GPT-5 as an independent judge: GRPO improvements over base models persist across all tasks and model families, and Qwen2.5-7B GRPO surpasses GPT-5 itself on DischargeMe even under GPT-5 evaluation (Appendix A).

Deployment Advantages for Resource-Constrained Settings On-premise deployment of compact models offers three key advantages over proprietary cloud APIs. First, patient data never leaves institutional servers, ensuring compliance with HIPAA and institutional data governance policies without requiring business associate agreements or data processing contracts with third parties. Second, local models can be customized—fine-tuned on institution-specific documentation styles, formularies, or patient populations—whereas proprietary APIs offer no such adaptation. Third, fixed infrastructure costs replace unpredictable per-query pricing, and inference remains available regardless of external API changes, rate limits, or service discontinuations. These factors make compact local models a practical choice for health systems prioritizing data sovereignty and operational independence. Concretely, 7–8B models deployed on a single A100 with an inference engine such as vLLM (Kwon et al., 2023) achieve a mean latency of 0.4057 ± 0.1694 s/sample across all three tasks, enabling interactive, non-cloud-dependent clinical documentation generation.

Deployment Safeguards Practical deployment requires multiple safeguards: transparent AI attribution, streamlined EHR-integrated review workflows, comprehensive audit logging, and conservative defaults prioritizing safety. Future reward

designs could explicitly penalize high-risk failure modes—hallucinated emergent diagnoses, contraindicated medications, omitted return precautions, or overconfident language promoting automation bias.

Limitations and Future Directions Our evaluation relies on retrospective datasets and automated judging. The current study focuses on demonstrating feasibility and measurable improvements using widely used public retrospective datasets; prospective studies in live ED and ICU workflows are needed to establish clinical validity, characterize failure modes, and quantify real-world time savings and clinician acceptance. We view this as an important next step for future work.

Current reward functions treat all outputs equally, but clinical importance varies: missing sepsis matters more than missing chronic hypertension. Future designs could incorporate clinical severity weighting and differentially penalize high-risk errors; systematic reward sensitivity studies and exploration of alternative rubric formulations are also needed to clarify robustness and safety trade-offs across varied clinical domains.

Additionally, extending RLVR to multimodal inputs (vital signs, imaging, waveforms) and real-time documentation generation represents a natural direction. Evaluation protocols should be stratified by demographic and clinical subgroups to detect disparities before deployment; we plan to incorporate subgroup and fairness analyses in future work to better assess equity and safety across populations. The influence of judge model selection on both training signals and evaluation outcomes also warrants systematic study.

6. Conclusion

In this work, we demonstrate that compact 7–8B parameter models trained with clinically grounded rewards can generate clinical-quality diagnoses, discharge instructions, and treatment plans. Two generalizable reward patterns—equivalence-based for list-based outputs, rubric-based for clinical prose—address the core verification challenge in open-ended clinical text generation, achieving performance matching or exceeding substantially larger proprietary systems. The resulting models enable on-premise deployment with full data sovereignty, institution-specific customization, and sub-second inference on standard GPU infrastructure. This establishes a practical pathway for integrating AI-assisted documentation into acute care workflows while maintaining the privacy and operational independence that health systems require.

Acknowledgments

Research reported in this publication was supported by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number R01HL172794 (PIs: D.K. and P.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong-Han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S Chiou, Christy Hong, Mohana Roy, Michael F Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen-Wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H Shah. MedHELM: Holistic evaluation of large language models for medical tasks. *arXiv [cs.CL]*, May 2025.

Runpeng Dai, Tong Zheng, Run Yang, Kaixian Yu, and Hongtu Zhu. R1-RE: Cross-domain relation extraction with RLVR. *arXiv [cs.CL]*, August 2025.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin,

Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, James C Fang, Savitri E Fedson, Gregg C Fonarow, Salim S Hayek, Adrian F Hernandez, Prateeti Khazanie, Michelle M Kittleson, Christopher S Lee, Mark S Link, Carmelo A Milano, Lorraine C Nnacheta, Alexander T Sandhu, Lynne Warner Stevenson, Orly Vardeny, Amanda R Vest, Clyde W Yancy, and ACC/AHA Joint Committee Members. 2022 AHA/ACC/HFSA guideline for the management of heart failure: A report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Circulation*, 145(18):e895–e1032, May 2022.

Samyak Jhaveri, Praphul Singh, Jangwon Kim, Tara Taghavi, and Krishnaram Kenthapadi. Optimizing long-form clinical text generation with claim-based rewards. *arXiv [cs.CL]*, September 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: https://doi.org/10.1038/sdata.2016.35.

Junu Kim, Chaeeun Shim, Sungjin Park, Su Yeon Lee, Gee Young Suh, Chae-Man Lim, Seong Jin Choi, Song Mi Moon, Kyoung-Ho Song, Eu Suk Kim, Hong Bin Kim, Sejoong Kim, Chami Im, Dong-Wan Kang, Yong Soo Kim, Hee-Joon Bae, Sung Yoon Lim, Han-Gil Jeong, and Edward Choi. Enhancing LLMs' clinical reasoning with real-world data from a nationwide sepsis registry. *arXiv [cs.AI]*, May 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Bojie Li. Incompressible knowledge probes: Estimating black-box llm parameter counts via factual capacity. *arXiv preprint arXiv:2604.24827*, 2026.

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, Jinjie Gu, Ya Zhang, Yanfeng Wang, Yu Wang, and Weidi Xie. EHR-R1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv [cs.CL]*, November 2025.

Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. Training LLMs for EHR-based reasoning tasks via reinforcement learning. *arXiv [cs.CL]*, May 2025.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Guanjing Xiong, and Hongsheng Li. UI-R1: Enhancing efficient action prediction of GUI agents by reinforcement learning. *arXiv [cs.AI]*, March 2025.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. *arXiv [cs.CV]*, February 2025.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *arXiv [cs.CL]*, pages 68772–68802, April 2024.

Anne F Peery, Aasma Shaukat, and Lisa L Strate. AGA clinical practice update on medical management of colonic diverticulitis: Expert review. *Gastroenterology*, 160(3):906–911.e1, February 2021.

Zhongxi Qiu, Zhang Zhang, Yan Hu, Heng Li, and Jiang Liu. Open-medical-R1: How to choose data for RLVR training at medicine domain. *arXiv [cs.LG]*, April 2025.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei

Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv [cs.CL]*, December 2024.

Maxime Robeyns and Laurence Aitchison. Improving LLM-generated code quality with GRPO. *arXiv [cs.AI]*, June 2025.

Codie R Rouleau, Daniele Chirico, Stephen B Wilton, Matthew K MacDonald, Tianqi Tao, Ross Arena, Tavis Campbell, and Sandeep Aggarwal. Mortality benefits of cardiac rehabilitation in coronary artery disease are mediated by comprehensive risk factor modification: A retrospective cohort study. *J. Am. Heart Assoc.*, 13(10):e033568, May 2024.

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y K Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv [cs.CL]*, February 2024.

A J Singer, J E Hollander, and J V Quinn. Evaluation and management of traumatic lacerations. *N. Engl. J. Med.*, 337(16):1142–1148, October 1997.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, Shixiang Tang, Lihao Liu, Bin Fu, Wenqi Shao, Xiaowei Hu, Xiangwen Liao, Yuanfeng Ji, and Junjun He. GMAI-VL-R1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv [cs.CV]*, April 2025.

Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, et al. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. *Nature medicine*, pages 1–1, 2025.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example. *arXiv [cs.LG]*, April 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. *arXiv [cs.CL]*, June 2025.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. Longwriter-zero: Mastering ultra-long text generation via reinforcement learning. *arXiv preprint arXiv:2506.18841*, 2025.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, et al. Overview of the first shared task on clinical text generation: Rrg24 and" discharge me!". *arXiv preprint arXiv:2409.16603*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv [cs.CL]*, May 2025.

Jixiao Zhang and Chunsheng Zuo. GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. *arXiv [cs.CL]*, September 2025.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-RLVR: Emerging medical reasoning from a 3B base model via reinforcement learning. *arXiv [cs.CL]*, February 2025a.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. Med-U1: Incentivizing unified medical reasoning in LLMs via large-scale reinforcement learning. *arXiv [cs.CL]*, June 2025b.

Appendix A. Cross-Judge Robustness Analysis

A fundamental concern in LLM-as-judge evaluation is whether observed improvements reflect genuine output quality or instead reflect optimization toward a particular judge's stylistic preferences. LLM evaluators have been shown to recognize and favor their own generations (Panickssery et al., 2024; Wataoka et al., 2025), introducing a *self-preference bias* that can confound evaluation when the same model family is used for both training rewards and scoring. Since our GRPO reward functions use GPT-4.1 to score outputs during training, models could in principle learn to produce text that exploits GPT-4.1's latent scoring tendencies (e.g., preferred document structure, verbosity, or phrasing) rather than achieving authentic clinical quality gains. Such an artifact would inflate scores under GPT-4.1 evaluation while providing no real benefit under an independent evaluator.

To test this, we re-evaluated all models using GPT-5 as an independent judge, applying the identical prompts described in Appendix C without modification. GPT-5 represents a qualitatively distinct generation of language model with a different training lineage, and is therefore unlikely to share the same latent evaluation biases as GPT-4.1. For MIMIC-III, GPT-5 re-performed the semantic equivalence judgments used to count matched diagnoses; the overlap-based metrics (F1, Jaccard similarity) were then recomputed deterministically over the resulting intersection sets, making this task the least susceptible to judge preference effects. For DischargeMe and MTSamples, all rubric scores were generated de novo by GPT-5.

Table 1 reports performance under both judges across all three tasks, with per-task averages for the two rubric-graded tasks and F1 for the diagnosis task. Per-dimension breakdowns are provided in Table 2.

GRPO consistently improves every base model under both judges. Across all three tasks and all three model families, each GRPO-trained model outperforms its corresponding base model regardless of which judge is used. For MIMIC-III, Ministral-8B improves from F1 0.309 to 0.482 (+0.173) under GPT-4.1 and from 0.297 to 0.437 (+0.140) under GPT-5; Qwen2.5-7B improves from 0.309 to 0.448 under GPT-4.1 and from 0.299 to 0.418 under GPT-5; Qwen3-8B improves from 0.209 to 0.417 under GPT-4.1 and from 0.182 to 0.386 under GPT-5. The near-identical magnitude of these gains under both judges is expected: F1 is computed deterministically once diagnoses are matched, so judge influence is limited to semantic equivalence decisions that are largely consistent across model generations. For DischargeMe, average rubric scores improve for all model families under both judges. For MTSamples, improvements are positive but more modest, and hold under both judges. The universality of this pattern across all nine model–task combinations provides strong evidence that GRPO gains are not judge-specific.

GRPO models remain competitive with frontier models under independent evaluation. On MIMIC-III, the best GRPO model (Ministral-8B GRPO, F1 = 0.437) outperforms the best GPT baseline (GPT-5, F1 = 0.411) even when evaluated by GPT-5. On DischargeMe, Qwen2.5-7B GRPO achieves an average score of 3.434 under GPT-5 evaluation—exceeding GPT-5 itself (3.392), despite GPT-5 serving as both the evaluation judge and a competing generation model. A judge with self-preference would be expected to rate its own outputs highly; the fact that an open-weight GRPO model surpasses GPT-5 under GPT-5 evaluation makes judge gaming an implausible explanation. On MTSamples, GRPO models do not match GPT frontier performance under either judge; however, they consistently improve over base models under both, indicating that the gap to frontier models on this task reflects a genuine quality ceiling rather than a judge artifact.

Judge calibration differs; the direction of improvement does not. The two judges assign systematically different absolute scores. On DischargeMe, GPT-4.1 rates its own outputs highest (4.416), suggesting a mild self-preference effect; GPT-5 does not—it rates GPT-5 outputs at 3.392, below Qwen2.5-7B GRPO (3.434). Per-dimension analysis reveals that GPT-5 penalizes completeness more harshly than GPT-4.1 does (e.g., GPT-5 model scores 1.009 on completeness under GPT-5 judge, compared

Table 1: Performance under GPT-4.1 and GPT-5 judges. DischargeMe and MTSamples report average rubric scores (scale 1–5); MIMIC-III reports F1. Bold indicates the best score within each judge–task column.

	MIMIC-III (F1 ↑)		**DischargeMe** (Avg ↑)		**MTSamples** (Avg ↑)	
Model	GPT-4.1	GPT-5	GPT-4.1	GPT-5	GPT-4.1	GPT-5
GPT-4.1	0.428	0.402	**4.416**	2.982	4.470	3.821
GPT-4o	0.436	0.409	4.039	2.681	4.441	**3.848**
GPT-5	0.433	0.411	2.986	3.392	4.464	3.834
Ministral-8B-Instruct	0.309	0.297	3.461	2.532	3.804	2.938
Qwen2.5-7B-Instruct	0.309	0.299	3.434	2.515	3.919	3.014
Qwen3-8B	0.209	0.182	3.637	2.331	4.371	3.324
Ministral-8B-Instruct GRPO	**0.482**	**0.437**	4.279	2.622	4.174	3.215
Qwen2.5-7B-Instruct GRPO	0.448	0.418	4.156	**3.434**	4.271	3.168
Qwen3-8B GRPO	0.417	0.386	3.752	2.386	**4.474**	3.402

Table 2: Per-dimension rubric scores under GPT-4.1 and GPT-5 judges. DischargeMe dimensions: Completeness (Comp.), Correctness (Corr.), Readability (Read.). MTSamples dimensions: Accuracy (Acc.), Clarity (Clar.), Completeness (Comp.).

	DischargeMe (GPT-4.1 judge)				**DischargeMe** (GPT-5 judge)			
Model	Comp.	Corr.	Read.	Avg	Comp.	Corr.	Read.	Avg
GPT-4.1	4.466	4.759	4.023	4.416	3.427	2.523	2.996	2.982
GPT-4o	3.928	4.330	3.860	4.039	3.059	2.191	2.793	2.681
GPT-5	2.438	3.815	2.706	2.986	1.009	4.986	4.181	3.392
Ministral-8B-Instruct	2.771	3.650	3.963	3.461	2.320	2.086	3.191	2.532
Qwen2.5-7B-Instruct	2.828	3.696	3.777	3.434	2.339	2.018	3.189	2.515
Qwen3-8B	3.444	3.991	3.476	3.637	2.704	1.864	2.425	2.331
Ministral-8B-Instruct GRPO	3.829	4.349	4.660	4.279	2.716	2.050	3.098	2.622
Qwen2.5-7B-Instruct GRPO	3.883	4.298	4.287	4.156	2.828	3.696	3.777	3.434
Qwen3-8B GRPO	3.544	4.050	3.662	3.752	2.712	1.886	2.561	2.386

	MTSamples (GPT-4.1 judge)				**MTSamples** (GPT-5 judge)			
Model	Acc.	Clar.	Comp.	Avg	Acc.	Clar.	Comp.	Avg
GPT-4.1	4.200	4.933	4.276	4.470	3.542	4.626	3.294	3.821
GPT-4o	4.171	4.877	4.275	4.441	3.629	4.563	3.352	3.848
GPT-5	4.209	4.882	4.303	4.464	3.592	4.620	3.291	3.834
Ministral-8B-Instruct	3.472	4.537	3.402	3.804	2.664	3.850	2.299	2.938
Qwen2.5-7B-Instruct	3.556	4.701	3.500	3.919	2.720	2.393	3.930	3.014
Qwen3-8B	4.047	4.738	4.327	4.371	2.892	4.117	2.962	3.324
Ministral-8B-Instruct GRPO	3.790	4.836	3.897	4.174	2.864	4.168	2.612	3.215
Qwen2.5-7B-Instruct GRPO	3.911	4.883	4.019	4.271	2.822	4.033	2.650	3.168
Qwen3-8B GRPO	4.187	4.748	4.486	4.474	2.864	4.238	3.103	3.402

to 2.438 under GPT-4.1 judge), while rewarding correctness more generously. These calibration differences mean that absolute score comparisons across judges are not directly meaningful, but they do not affect the direction-of-improvement analysis, which is consistent under both judges.

Conclusion. GRPO improvements are not artifacts of optimizing for GPT-4.1's scoring style. Gains over base models persist when evaluated by GPT-5 across all tasks and all model families, GRPO models remain competitive with frontier models under independent assessment, and no evidence of judge gaming is detectable from the cross-judge comparison. We conclude that the observed improvements reflect genuine advances in clinical text generation quality rather than exploitation of a particular judge's preferences. Future work should nonetheless systematically examine reward functions derived from diverse LLM families to further strengthen robustness claims.

Appendix B. Representative Qualitative Examples

The following five cases illustrate clinical improvements produced by GRPO training. Each case shows relevant excerpts from the base model and GRPO-trained model outputs.

Case 12 — Hand Laceration (MTSamples) A 19-year-old male with hand lacerations repaired with 17 sutures after a fight. The critical discharge instruction concerns wound soaking. Expert guidelines state traumatic lacerations should not be soaked after 48 hours (Singer et al., 1997).

Base model (wound care excerpt)

Advise the patient to keep the wounds clean and dry, avoiding submersion in water (e.g., swimming, baths) for **48–72 hours** post-repair.

GRPO model (wound care excerpt)

Advise the patient to keep the hand elevated to reduce swelling and avoid submerging the wound in water (e.g., no swimming or soaking) for **7–10 days**.

Case 21 — STEMI with Reduced EF (MTSamples) A patient with acute STEMI successfully treated with PCI and stenting of the LAD. LV ejection fraction was 45–50%, indicating mild-to-moderate impairment. The GRPO model identified the need for guideline-directed medical therapy (GDMT) and cardiac rehabilitation; the base model omitted these.

Base model (post-PCI management excerpt)

Continue anticoagulation with dual antiplatelet therapy... Administer beta-blockers (e.g., metoprolol) to reduce myocardial oxygen demand... Schedule serial echocardiograms to monitor LV function. [No mention of timing, GDMT for HFmrEF, or cardiac rehabilitation.]

GRPO model (post-PCI management excerpt)

Ensure beta-blocker initiation **within 24–48 hours** (if no contraindications)... For LV dysfunction (EF 45–50%), use **guideline-directed medical therapy (GDMT)**: ACE inhibitors/ARNIs, mineralocorticoid receptor antagonists, SGLT2 inhibitors (e.g., dapagliflozin) if indicated... **Cardiac Rehabilitation:** Initiate early after stabilization to improve cardiovascular fitness and reduce mortality.

Case 22 — Acute Sigmoid Diverticulitis (MTSamples) A 71-year-old female with left lower quadrant pain. The ground truth identifies sigmoid diverticulitis as the working diagnosis and explicitly defers colonoscopy (acute perforation risk (Peery et al., 2021)). The base model did not commit to a leading diagnosis and listed colonoscopy as a workup option without timing caution.

Base model (workup excerpt)

Colonoscopy: Given her age, family history, and persistent constipation, consider colonoscopic evaluation to rule out neoplastic or inflammatory bowel disease... [No working diagnosis specified; colonoscopy recommended without acute-setting safety caveat.]

GRPO model (workup excerpt)

Repeat Imaging: Given persistent LLQ pain and a palpable mass, obtain a contrast-enhanced CT abdomen/pelvis to reassess for **diverticulitis**, abscess, or other pathology... Keep patient NPO empirically. [Colonoscopy not recommended in the acute setting.]

Case 43 — Critically Ill ICU Patient (MIMIC-III) A complex ICU patient with multi-organ fail-

ure, septic shock, bowel perforation, coagulopathy, and multiple comorbidities. The ground truth contains 22 diagnoses. The base model produced 6; the GRPO model produced 21.

Base model (6 diagnoses)

Sepsis, Septic Shock, Liver Failure, Kidney Failure, Hepatic Encephalopathy, Multi-Organ Failure

GRPO model (21 diagnoses)

Sepsis / Multi-organ failure; Bowel perforation with complications (ileus, renal failure, hepatic failure); Respiratory failure / hypoxemia; Hepatic encephalopathy; Acute kidney injury; Coagulopathy and electrolyte disturbances; Liver dysfunction / acute necrosis; CAD and atrial fibrillation; History of colon cancer and prior irradiation; Alcoholic cirrhosis; Diabetes mellitus; Metabolic acidosis / hypotension; Thrombocytopenia; Bacteroides infection; Hypertension; Hypercholesterolemia; GERD; Prior CABG; Abdominal surgical complications; Severe sepsis / palliative care; Pulmonary insufficiency

Case 61 — Deceased Patient (DischargeMe) The patient died in hospital after transitioning to comfort care. The ground truth is a single sentence acknowledging the death. The base model failed to register this outcome and generated standard living-patient discharge instructions. The GRPO model correctly identified the patient was deceased and addressed the family.

Base model (excerpt)

Cardiac Management: Continue medications as prescribed. . . especially those related to blood pressure control, lipid management, and heart rate regulation (lisinopril, carvedilol, sotalol).

Diabetes Management: Continue metformin and insulin therapy as prescribed. Monitor blood sugar levels regularly. . .

Follow-Up: Schedule regular follow-ups with the primary care physician and specialists as necessary.

GRPO model (excerpt)

Discharge Condition: Patient passed away during hospitalization after being transitioned to comfort care measures.

Dear Family, We are writing to inform you about your loved one's treatment and outcome during their recent stay. . .

Current Status: Your loved one passed away while under our care. We deeply regret this loss and extend our condolences to you and your family.

Appendix C. LLM Judge Prompts

This appendix provides the complete prompts used for the LLM judge (GPT-4.1) in reward computation and evaluation for all three clinical tasks.

C.1. Equivalence-Based Reward: Diagnosis Generation

The following prompt was used to determine semantic equivalence between predicted and ground-truth diagnoses, accounting for medical synonymy, abbreviations, and varying specificity (Figure 5).

C.2. Rubric-Based Reward: Discharge Instructions

The following prompt was used to evaluate discharge instructions on completeness, correctness, and readability (Figure 6).

C.3. Rubric-Based Reward: Treatment Plans

The following prompt was used to evaluate treatment plans on accuracy, completeness, and clarity (Figure 7).

```
You are a medical evaluation expert assessing the accuracy of a predicted diagnosis list
against an official diagnosis list.
Student's Predicted Diagnoses: {generated_output}
Official Diagnosis List (Ground Truth): {ground_truth_output}
Evaluation Guidelines for Matching:
1. Consider medical synonyms and alternative terminology for the same condition (e.g.,
"myocardial infarction" and "heart attack")
2. Account for different levels of specificity (e.g., "pneumonia" matching "bacterial
pneumonia")
3. Accept abbreviated forms and full names (e.g., "COPD" and "Chronic Obstructive Pulmonary
Disease")
4. Evaluate if the predicted diagnosis is clinically equivalent to any diagnosis in the
ground truth list
5. Consider ICD code matches if applicable
Task:
1. Identify all predicted diagnoses that match any diagnosis in the ground truth list (using
the evaluation guidelines above)
2. Count the total number of matching diagnoses
3. List each matching diagnosis pair (predicted -> ground truth)
Return your evaluation as a single JSON object in the following format:
{
"intersection":
{
"score": <number of matching diagnoses>,
"matched_diagnoses": [ {"predicted": "predicted diagnosis 1", "ground_truth": "matched
ground truth diagnosis"}, {"predicted": "predicted diagnosis 2", "ground_truth": "matched
ground truth diagnosis"} ],
"explanation": "Explain which diagnoses matched and why, and which did not match."
}
}
Ensure the output is valid JSON:
- Use double quotes (") for all keys and string values.
- When quoting text or sections inside the explanations, use escaped double quotes (\") to
maintain valid JSON formatting.
- Do not include any additional information in the output.
```

Figure 5: LLM judge prompt for equivalence-based reward in diagnosis generation task. The prompt handles medical synonymy, abbreviations, and varying specificity levels.

```
You are a medical expert tasked with evaluating the quality of generated discharge
instructions for a patient case.  Your goal is to assess how well the instructions address
patient needs, follow clinical best practices, and compare to the reference instructions.
The generated discharge instructions will be provided in these tags:
<response> {RESPONSE} </response>
Reference discharge instructions will be provided in these tags:
<gold_response> {GOLD_RESPONSE} </gold_response>
Carefully analyze the <response>.  For each of the following categories, rate the response on
a scale of 1 to 5 (1 = very poor, 5 = excellent), and provide a short justification for your
score.
Evaluation Criteria:
Completeness (1-5)
- Does the instruction cover all clinically important information including medications,
warning signs requiring immediate attention, activity restrictions, wound care (if
applicable), and follow-up appointments?
- Are return precautions clearly specified?
- Is medication reconciliation addressed?
Correctness (1-5)
- Is the instruction medically accurate and free from harmful or misleading information?
- Are there any contraindicated recommendations?
- Is dosing information accurate when provided?
Readability (1-5)
- Is the instruction clear, well-organized, and appropriate for patient comprehension?
- Is medical jargon explained or avoided?
- Are instructions actionable and specific?
Output Format:  Output the evaluation as a single valid JSON object matching the following
structure:
{
"completeness":  { "score":  0, "explanation":  "Explain why this score was given." },
"correctness":  { "score":  0, "explanation":  "Explain why this score was given." },
"readability":  { "score":  0, "explanation":  "Explain why this score was given." }
}
Ensure the output is valid JSON:
- Use double quotes (") for all keys and string values.
- When quoting text or sections inside the explanations, use escaped double quotes (\") to
maintain valid JSON formatting.
- Do not include any additional information in the output.
```

Figure 6: LLM judge prompt for rubric-based reward in discharge instruction generation task. The prompt evaluates completeness, correctness, and readability.

```
You are a medical expert tasked with evaluating the quality of a generated treatment plan for
a clinical scenario. Your goal is to assess how well the treatment plan addresses the patient
case, follows clinical best practices, and compares to the reference treatment plan.
The generated treatment plan will be provided in these tags:
<response> {RESPONSE} </response>
The reference treatment plan will be provided in these tags:
<gold_response> {GOLD_RESPONSE} </gold_response>
Carefully analyze the <response>. For each of the following categories, rate the response on
a scale of 1 to 5 (1 = very poor, 5 = excellent), and provide a short justification for your
score.
Evaluation Criteria:
Accuracy (1-5)
- Does the treatment plan align with current clinical guidelines and medical evidence?
- Are all recommended treatments appropriate for the patient's condition?
- Are there any contraindicated medications or interventions?
Completeness (1-5)
- Does the plan include all clinically important details such as medication dosages,
administration routes, monitoring parameters, duration, and follow-up?
- Are important aspects of care (medication, lifestyle modifications, follow-up) addressed?
- Are potential complications or side effects mentioned when appropriate?
Clarity (1-5)
- Is the plan logically organized and easy for the treating team to understand and implement?
- Are instructions specific and actionable?
- Is the reasoning for treatment choices explained when necessary?
Output Format: Output the evaluation as a single valid JSON object matching the following
structure:
{
"accuracy": { "score": 0, "explanation": "Explain why this score was given." },
"completeness": { "score": 0, "explanation": "Explain why this score was given." },
"clarity": { "score": 0, "explanation": "Explain why this score was given." }
}
Ensure the output is valid JSON:
- Use double quotes (") for all keys and string values.
- When quoting text or sections inside the explanations, use escaped double quotes (\") to
maintain valid JSON formatting.
- Do not include any additional information in the output.
```

Figure 7: LLM judge prompt for rubric-based reward in treatment plan generation task. The prompt evaluates accuracy, completeness, and clarity.