

# DeepICD-R1: Medical Reasoning through Hierarchical Rewards and Unsupervised Distillation

Tom Röhr<sup>1</sup>, Thomas Steffek<sup>1</sup>, Roman Teucher<sup>2</sup>, Keno Bressemer<sup>3</sup>,  
Alexei Figueroa<sup>1</sup>, Paul Grundmann<sup>1</sup>, Peter Troeger<sup>1</sup>, Felix Gers<sup>1</sup>, Alexander Löser<sup>1</sup>

<sup>1</sup> Berlin University of Applied Sciences (BHT), Data Science and Text-based Information Systems Group

<sup>2</sup> Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

<sup>3</sup> University Hospital Rechts der Isar, Technical University of Munich

{tom.roehr, thomas.steffek, alexei.figueroaroser, paul.grundmann, peter.troeger,  
felixalexander.gers, aloeser}@bht-berlin.de  
roman.teucher@iais.fraunhofer.de, keno.bressemer@tum.de

## Abstract

Large language models (LLMs) show strong reasoning abilities, but full retraining for the medical domain is often infeasible because of lacking data or compute resources. We present DeepICD-R1, a framework for efficient medical reasoning fine-tuning that unites hierarchical rewards with distilled supervision. We reformulate ICD-10-CM prediction as a reinforcement learning problem and design a hierarchical outcome-based reward that reflects the ICD code structure across chapter, category, and full-code levels. In parallel, we publish a large-scale distilled dataset of over 90k reasoning traces derived from MIMIC-IV admission notes, integrating clinical validation and official coding guidelines. Fine-tuning smaller instruction-tuned LLMs with this data and GRPO reinforcement yields consistent gains in diagnostic accuracy and reasoning coherence. Extensive ablations confirm that hierarchical supervision and verifiable outcome rewards enable competitive, domain-specialized reasoning models without additional pretraining, providing a reproducible foundation for clinical NLP research.

**Keywords:** Clinical NLP, Large Reasoning Model, GRPO, Supervised Fine-Tuning

## 1. Introduction

Large language models (LLMs) currently define the state of the art across a wide range of NLP tasks. However, only a few institutions possess the computational resources required to train such monolithic foundation models. As an alternative, *reasoning- and test-time compute strategies* offer scalable pathways for domain specialization. These methods allow adaptation of existing LLMs, leveraging the extensive world knowledge already embedded in their parameters. Consequently, *large reasoning models (LRMs)* have become a new frontier in language model intelligence, achieving breakthroughs in domains such as mathematics and programming (Ye et al., 2025; Muennighoff et al., 2025; Zhou et al., 2023).

Reinforcement-based post-training methods represent a promising direction, shifting the focus from static pretraining data to tasks with *verifiable rewards*. Verifiable rewards allow automatic correctness checking, enabling models to iteratively explore solution spaces and overcome the limitations of fixed datasets. A central open question is whether such reasoning capabilities can generalize beyond formal domains toward *clinical reasoning* (Wu et al., 2025; Chen et al., 2025; Wang et al., 2024), in which interpretability and verifiability are crucial. In clinical settings, diagnostic reasoning processes typically follow a hierarchical structure, progressing from the identification of a general dis-

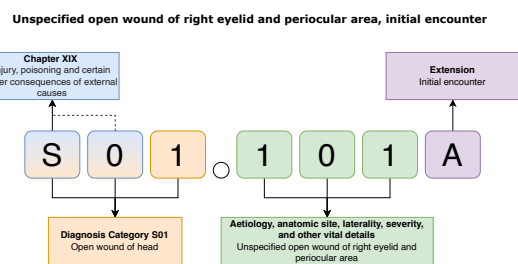


Figure 1: Structure of the ICD-10-CM code hierarchy. Each code can contain up to seven characters, with additional digits indicating progressively higher diagnostic specificity.

ease area to increasingly specific diagnostic categories.

Building on this clinical reasoning hierarchy, this work introduces **DeepICD-R1**, a framework for efficient medical reasoning using fine-tuning and reinforcement learning with verifiable hierarchical rewards. We demonstrate, for the first time, that without additional domain-specific pretraining and using verifiable rewards derived from structured diagnostic codes, reasoning-oriented post-training successfully improves a model's capabilities on the *clinical core task of outcome prediction* (van Aken et al., 2021; Röhr et al., 2024). Figure 2 shows an overview of our complete framework. We make three key contributions:

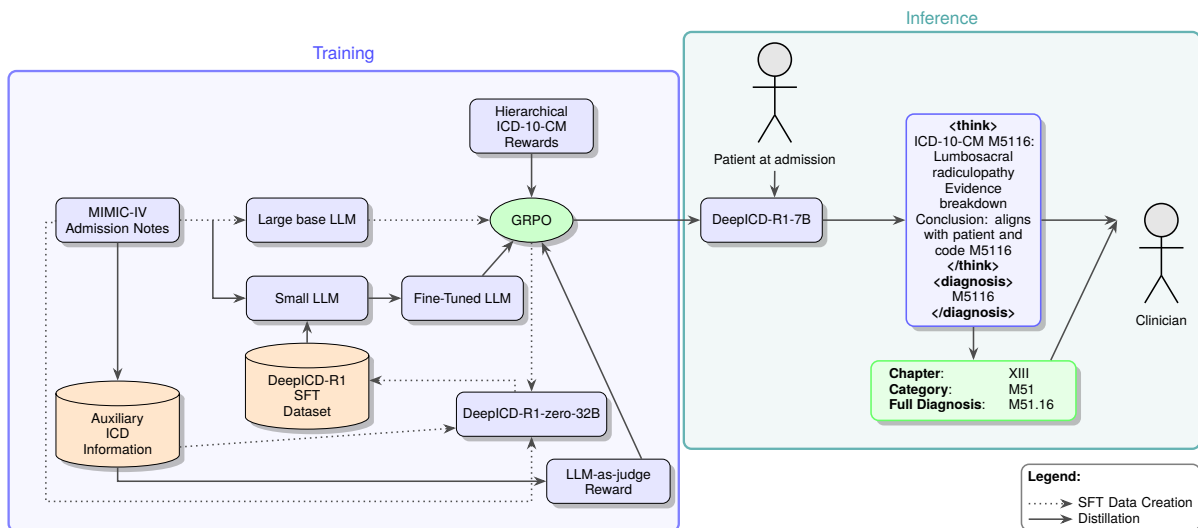


Figure 2: Overview of the full framework. We train a large base LLM on MIMIC-IV admission notes using GRPO, guided by dedicated reward functions and evaluation tasks. This process yields **DeepICD-R1-zero-32B**. Leveraging this model and external ICD knowledge, we construct a novel reasoning dataset. Smaller models are then fine-tuned on this dataset and further optimized with our reward functions, resulting in **DeepICD-R1-7B**.

### 1: Hierarchical outcome-based reward design.

We design a hierarchical reward function that mirrors the structured nature of ICD codes. It reflects partial correctness across chapter, category, and full code levels, enabling interpretable performance analysis along the ICD hierarchy, see also Figure 1.

### 2: Distilled dataset for reasoning fine-tuning.

We construct a large-scale reasoning dataset derived from MIMIC-IV admission notes, combining hierarchical reasoning with clinical validation and additional ICD coding information. This MIMIC-Reason dataset provides 93k reasoning samples over 6k unique codes and supports supervised fine-tuning that substantially improves both diagnostic accuracy and reasoning coherence.

### 3: Reinforcement learning design study for clinical outcome prediction.

We systematically evaluate reinforcement-based fine-tuning for clinical reasoning using Group Relative Policy Optimization (GRPO) (Shao et al., 2024b). Through ablation experiments on reward components and response length, combined with qualitative and descriptive analyses, we identify key design challenges specific to outcome prediction from admission notes. All code, including data processing pipelines and reward functions, as well as trained models and the SFT dataset, will be publicly released via GitHub, HuggingFace, and PhysioNet to ensure reproducibility.<sup>1</sup>

<sup>1</sup><https://github.com/DATAXIS/DeepICD-R1>

## 2. Related Work

**Large Reasoning Models (LRMs).** Large language models have recently evolved into large reasoning models (LRMs) that explicitly optimize for multi-step reasoning. Chain-of-thought prompting improves problem solving by eliciting intermediate steps (Wei et al., 2022). Reinforcement-based post-training methods, including Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and Proximal Policy Optimization (PPO) (Schulman et al., 2017), further align models with reasoning-oriented objectives. While PPO remains a standard approach, its actor-critic design incurs high computational cost, limiting scalability for reasoning-heavy objectives. Recent work such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024a) and DeepSeek-R1 (DeepSeek-AI et al., 2025) reduces the cost of policy optimization while maintaining reasoning fidelity, demonstrating that structured reasoning traces and verifiable rewards can replace full re-training. GRPO substitutes the critic with a group-relative reward estimator. Given a prompt, the actor generates multiple candidate outputs that are scored using static reward functions. Each output’s advantage is computed relative to the group mean, and the model is updated toward higher-performing outputs. Thus, GRPO eliminates the need for a value network, substantially reducing computational demand while maintaining stable learning. These developments form the foundation for scalable reasoning adaptation, motivating our use of GRPO-based fine-tuning for clinical reasoning with verifiable rewards.

**Medical Reasoning and Clinical Alignment.** Adapting LRMs to medicine requires factual grounding, interpretability, and safety. Foundation models such as Med-PaLM (Singhal et al., 2023, 2025) achieve strong QA performance after supervised and feedback-based tuning but remain limited by static supervision. More recent efforts like HuatuoGPT-o1 (Chen et al., 2025), MedReason (Wu et al., 2025), and DiReCT (Wang et al., 2024) incorporate reasoning traces through LLM generation, medical knowledge graphs, or human annotation. While these works improve interpretability for multi-hop question answering, they rely on curated reasoning data or heuristic verification. In contrast, our approach introduces *hierarchical, automatically verifiable rewards* based on ICD-10-CM structure, enabling reinforcement learning without domain-specific pretraining or costly human evaluation for a single label differential diagnosis setting (Pearn, 2022).

**Outcome Prediction vs. ICD Coding.** Predicting discharge diagnoses from admission information extends beyond automated ICD coding (Mullenbach et al., 2018). Unlike ICD Coding, which maps completed clinical notes to known diagnoses, outcome prediction requires reasoning about uncertain disease trajectories (van Aken et al., 2021; Röhr et al., 2024). Existing approaches rely on encoder-based or metric-learning architectures (van Aken et al., 2022; Figueroa et al., 2024) but lack interpretable reasoning traces. Grundmann et al. (2025) recently assessed generative models for this task, yet without post-training. By framing outcome prediction as a reinforcement problem aligned with the ICD hierarchy, we provide structured intermediate supervision that fosters both interpretability and verifiable evaluation.

**Positioning DeepICD-R1.** Our framework, **DeepICD-R1**, unifies for the first time reasoning-based post-training with ICD-10-CM clinical outcome prediction. It introduces (i) *hierarchical outcome rewards* reflecting ICD structure and (ii) a *distilled reasoning dataset* built from admission notes and external medical knowledge. Unlike prior medical LRMs that depend on static QA datasets or heuristic verifiers, DeepICD-R1 leverages *verifiable, structured signals* to guide reasoning. Therefore, it represents the first integration of reasoning-oriented reinforcement learning and hierarchical outcome-grounded supervision in the clinical domain.

### 3. Diagnostic Reasoning Model

This section introduces the methods and resources used to study reasoning-oriented ICD-10-CM di-

agnosis prediction at admission time. We first outline, the objective and reward functions used during post-training, followed by the hierarchical prediction tasks, and finally describe our novel data resources.

#### 3.1. Reward Functions and Tasks

We extend the work from van Aken et al. (2021) and reformulate clinical outcome prediction as a multi-class generation task in the ICD-10-CM code space using MIMIC-IV admission notes (Johnson et al., 2024, 2023; Goldberger et al., 2000). We use only records from the *hosp* split and exclude sections containing diagnostic or treatment information to prevent label leakage. The first annotated diagnosis code is taken as the outcome target, as the ordering of codes in MIMIC-IV approximately reflects their relative clinical relevance (Johnson et al., 2023).

To enable GRPO post-training, we implement three complementary rewards: a format reward, a hierarchical outcome reward, and a LLM-as-a-judge reward.

**Format reward.** The model is rewarded for producing reasoning traces enclosed in `<think>` tags, followed by a diagnosis enclosed in `<diagnosis>` tags. To obtain a positive reward the outputs needs to satisfy two regular expressions: The first regular expression checks for exactly one `<think>...</think>` block and one `<diagnosis>...</diagnosis>` block, allowing arbitrary text inside each tag. Second, text within the `<diagnosis>` tags must further match the ICD-10-CM format pattern  $^{\wedge}[A-Z][0-9]\{2\}[A-Z0-9]\{0,4\}\$$ .

**Hierarchical outcome reward.** As Equation 1 shows, we leverage the ICD hierarchy (Figure 1) to compute partial credit based on prefix digit overlap  $O_{prefix}$  between prediction  $p$  and ground truth  $g$ :

$$r = \frac{O_{prefix}(g, p)}{|g|} + |O_{prefix}(g, p) - 2| \quad (1)$$

Note that we award special bonuses for reaching three and more correct digits, since the first three digits already encapsulate important diagnostic information in an ICD code (Hirsch et al., 2016). To encourage the model to predict more than three digits, we linearly scale this bonus for each additional digit after.

**LLM-as-a-judge reward.** To improve reasoning clarity and consistency, an external LLM from a different model family scores outputs on a 1–5 scale based on the admission note, predicted diagnosis, and our auxiliary ICD-10-CM information. Instead of the ground-truth diagnosis, we use the predicted

diagnosis to retrieve the symptom explanation and specific guidelines chapter. This ensures that the auxiliary signal encourages concise, interpretable reasoning aligned with clinical expectations, in respect to the predicted diagnosis.

**Hierarchical ICD-10-CM prediction tasks.** To evaluate reasoning quality, we introduce three hierarchical prediction tasks based on the same MIMIC-IV admission notes: **chapter-**, **category-**, and **full-code** prediction (Table 1). Each task is multi-class, with an added *NIL* label for out-of-distribution predictions at non-chapter levels.

Task	ICD Level	# Classes
Chapter	1–2 digits	21
Category	3 digits	1115
Full Diagnosis	4–7 digits	6369

Table 1: ICD-10-CM hierarchical prediction tasks.

### 3.2. Data Resources

General instruction-tuned models lack the clinical grounding, hierarchical structure, and domain-specific reasoning patterns required for reliable outcome prediction. To overcome these limitations, we construct dedicated data resources that provide factual grounding and verifiable supervision: enriched ICD-10-CM metadata and a supervised fine-tuning (SFT) dataset of reasoning traces derived from the MIMIC-IV admission notes. These resources enable GRPO to operate on structured, clinically meaningful signals.

**ICD-10-CM symptom descriptions and coding guidelines.** We compile public symptom explanations and official coding guidelines to enrich reasoning and reward functions. Symptom descriptions were scraped from a verified national health portal<sup>2</sup> after confirming scraping permissions. These cover **11,788** ICD-10-CM codes (**163** three-digit, **6,562** four-digit, and **5,063** five-digit entries) with an average of 132 words per description. While there is at least one description for each ICD chapter, Figure 3 shows a distribution skewed toward Chapter XIII (musculoskeletal diseases) and Chapter XIX (injuries and poisoning). Official ICD-10-CM coding guidelines<sup>3</sup> were parsed into structured JSON format by chapter for efficient retrieval during reasoning and data generation.

**Supervised Fine-Tuning (SFT) dataset.** Standalone GRPO training may degrade reasoning flu-

<sup>2</sup><https://gesund.bund.de/>

<sup>3</sup><https://www.cms.gov/files/document/fy-2025-icd-10-cm-coding-guidelines.pdf>

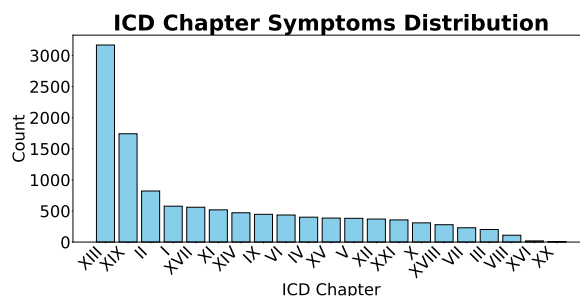


Figure 3: Distribution of scraped symptom descriptions across ICD-10-CM chapters. The highest variance occurs in Chapters XIII and XIX, indicating uneven coverage of diagnostic domains.

ency or produce malformed traces (DeepSeek-AI et al., 2025). To alleviate these problems, DeepSeek-AI et al. (2025) propose to initialize GRPO fine-tuning from a supervised model trained for the same reasoning objective. Following the same approach, we train an initial GRPO model to produce structured reasoning and predictions as described in subsection 3.1. We then use this checkpoint to generate reasoning traces for admission notes paired with their first annotated diagnosis code. To increase validity and consistency, we incorporate the code specific scraped symptom description and chapter guidelines to the prompt during generation. The dotted lines within the *Training* block in Figure 2 indicate this process. The resulting dataset contains **93,142** samples across **6,368** unique codes, with an average reasoning length of 477 words, as shown in Table 2.

Metric	Value
Samples	93,142
Unique codes	6,368
Mean trace length	477.3 words
Min / Max trace length	31 / 883 words

Table 2: Statistics of the SFT dataset for reasoning-based ICD-10-CM prediction.

## 4. Experimental Setup

This section outlines the experimental configuration for the ICD-10-CM diagnosis prediction (subsection 3.1) using both reinforcement learning (GRPO) and supervised fine-tuning (SFT). We apply stratified sampling (Parsons, 2017) to produce train, test, and validation splits and report 65,228, 18,654, and 9,260 samples within each split, respectively.

### 4.1. Training Environment

**GRPO.** Reinforcement learning experiments are conducted with the open-source framework *Verl*

(Sheng et al., 2025), following hyperparameter recommendations from its official repository.<sup>4</sup> Each model is trained with an effective batch size of 64 and 8 rollouts per update, with the temperature set to 0.9. The KL regularization term is disabled, consistent with the open-source DAPO setup (Yu et al., 2025). We retain the default entropy coefficient and learning rate schedule provided by *Verl*.

**SFT.** Supervised fine-tuning is performed using the *trl* library (von Werra et al., 2020). Models are trained for one epoch with an effective batch size of 32 and a learning rate of  $2 \times 10^{-5}$  using the AdamW optimizer (Loshchilov and Hutter, 2017). A linear warm-up of 100 steps is applied.

To ensure consistent generation behavior across settings, training and inference are executed with *vLLM* (Kwon et al., 2023) using identical decoding parameters.

## 4.2. Models

We evaluate models from two widely used open-source families, namely *Qwen2.5* and *Llama3.1*, alongside derivatives fine-tuned for medical reasoning and clinical question answering. All models are instruction-tuned, since base versions were found unable to follow the required reasoning–diagnosis output format.

From the Qwen series, we use **Qwen2.5-7B-Instruct** and **Qwen2.5-32B-Instruct** (Qwen et al., 2025). We omit Qwen3 due to computational constraints and to maintain comparability across experiments conducted before its release. Qwen2.5 remains a strong baseline for reasoning-oriented tasks and has extensive open-source support.

Within the Llama family, we include **Llama3.1-8B-Instruct** (Grattafiori et al., 2024).

In addition, we evaluate two reasoning-specialized medical models: **HuatuogPT-o1-7B** (Chen et al. (2025), based on Qwen2.5-7B) and **MedReason-8B** (Wu et al., 2025), based on HuatuogPT-o1-8B, which in turn is based on Llama3.1-8B. Although these models were not designed for ICD prediction, their training on clinical reasoning tasks provides an informative test of transfer to structured diagnostic prediction.

## 5. Experimental Results and Discussion

In this section we report the quantitative results for hierarchical ICD-10-CM prediction and analyze how reasoning-oriented fine-tuning affects diagnostic performance. Additional studies examine hier-

archical performance trends, the impact of reasoning traces in supervised-finetuning, and qualitative properties of generated traces.

### 5.1. Quantitative Results

We evaluate performance at chapter, category, and full code level of the ICD-10-CM hierarchy using macro-averaged precision, recall, and F1.

Table 3 summarizes all metrics across all tasks. Overall, reasoning-oriented post-training substantially improves diagnostic accuracy compared to instruction-tuned baselines.

**Baseline comparison.** Among instruction-tuned models, **Qwen2.5-32B-Instruct** achieves the highest baseline F1 across all hierarchy levels, followed by **Qwen2.5-7B-Instruct**. **HuatoGPT-o1-7B**, which is based on Qwen2.5-7B, ranks third. **Llama3.1-8B-Instruct** and **MedReason-8B** perform considerably worse, since they often do not adhere to the required `<think><diagnosis>` structure.

**Joint SFT + GRPO yields highest performance.** The **Llama3.1-8B-Instruct (SFT + GRPO)** model achieves the best results across all hierarchy levels, surpassing its baseline by +42.4 F1 at the chapter level, +14.2 F1 at category level, and +4.3 F1 at the full-code level. Once fine-tuned with reasoning supervision, the Llama model learns to follow the structured output format and excels. The **Qwen2.5-7B-Instruct (SFT + GRPO)** model ranks second, confirming the consistency of improvements across architectures.

**SFT is essential for detailed code prediction.** Within the Qwen2.5 family, shown here as a representative example, we observe that the 7B model performs moderately at the chapter level (39.4 F1) but nearly fails for category (3.4 F1) and full-code (0.2 F1) prediction. The larger 32B variant offers only limited gains for chapter- and full-code level (10.4 F1 and 0.9 F1). Supervised fine-tuning achieves gains of up to 17.5× relative to instruction-tuned baselines and markedly outperforms GRPO-only training (+255.8% vs. +44.1% F1 at the Category level), underscoring that fine-grained diagnostic prediction requires explicit supervision beyond instruction or reward optimization.

**Effect of domain-specific pretraining.** Medical reasoning models such as **MedReason-8B** and **HuatuogPT-o1-7B** perform worse than their non-medically-fine-tuned counterparts Qwen2.5-7B and Llama3.1-8B, despite their strength in clinical QA tasks. Their focus on factual question answering rather than forward reasoning limits transferability

<sup>4</sup>[https://github.com/volcengine/verl/tree/main/examples/grpo\\_trainer](https://github.com/volcengine/verl/tree/main/examples/grpo_trainer)

Model Name	Chapter			Category			Full Code		
	R	P	F1	R	P	F1	R	P	F1
Baselines									
Qwen2.5-7B-Instruct	39.0	44.8	39.4	4.0	4.7	3.4	0.3	0.5	0.2
Qwen2.5-32B-Instruct	42.2	49.1	43.5	12.0	13.4	10.4	1.3	1.4	0.9
Llama3.1-8B-Instruct	14.6	43.0	17.1	1.2	4.2	1.4	0.1	0.2	0.1
MedReason-8B	15.1	43.3	17.8	1.9	6.5	2.5	0.1	0.2	0.1
HuatuoGPT-01-7B	37.2	41.9	37.6	3.9	4.6	3.3	0.2	0.3	0.2
GRPO									
Qwen2.5-7B-Instruct	45.6	49.6	43.9	6.4	6.0	4.9	0.5	0.6	0.4
Qwen2.5-32B-Instruct	49.6	56.1	49.8	15.1	13.9	12.6	1.5	1.4	1.1
SFT									
Qwen2.5-7B-Instruct	48.3	57.3	48.9	13.1	13.0	12.1	4.3	3.6	3.5
Llama3.1-8B-Instruct	51.1	61.5	52.9	14.6	16.2	14.0	5.1	4.0	3.9
SFT + GRPO									
Qwen2.5-7B-Instruct	55.0	61.4	54.5	16.1	15.6	14.4	5.6	3.9	3.9
Llama3.1-8B-Instruct	59.1	64.6	59.5	17.1	18.0	15.6	6.4	4.2	4.3

Table 3: Recall-, Precision-, and F1-macro results for non-finetuned models, as well as models fine-tuned in a supervised manner (SFT) and through reinforcement learning (GRPO). Training with both SFT and GRPO outperforms models using only one training scheme. Both models trained with SFT consistently match or outperform a model 4x their size trained with GRPO. SFT is especially helpful for full code prediction, where GRPO lacks. Overall, the more specific the ICD code, the more performance declines over all settings.

to diagnostic forecasting. This emphasizes the need for specialized diagnostic large reasoning models.

**Ablation study results.** We conducted ablation experiments to assess the contribution of individual reward components and response length settings. The following observations summarize their effects:

**Outcome and format rewards are essential.** GRPO training fails to converge when either component is removed. Without the format reward, predictions cannot be properly structured or evaluated; without the outcome reward, the model receives no meaningful learning signal for accurate ICD-10-CM code generation.

**LLM-as-a-judge enhances reasoning quality.** Models trained solely with outcome and format rewards achieve competitive quantitative performance but exhibit a collapse in reasoning quality. Their reasoning traces often degenerate into superficial or even nonsensical outputs, offering no interpretive value for medical professionals. Incorporating the augmented LLM-as-a-judge reward restores coherent and structured reasoning while maintaining predictive accuracy, as the model learns to satisfy both outcome- and reasoning-oriented objectives.

**Longer traces add redundancy rather than diversity.** Increasing the maximum response length leads to repetitive reasoning rather than novel insights, reducing both readability and performance on evaluation tasks. These observations are con-

sistent with similar findings in Zhang and Zuo (2025), Huang et al. (2025), and Hassid et al. (2025).

## 5.2. Discussion

**Impact of Chapter Imbalance on Model Performance.** Figure 4 shows the distribution of samples across ICD-10-CM chapters in our generated reasoning dataset and the corresponding model performance. As in the original MIMIC-IV dataset, we observe a pronounced class imbalance: the two most frequent chapters account for one third of all samples, while the least frequent tercile spans 13 chapters. Model performance mirrors this distribution, with markedly lower F1 for less frequent chapters, like rare congenital or genetic conditions (Chapter XVII). This pattern aligns with findings from previous encoder-based approaches (Röhr et al., 2024; Figueroa et al., 2024) and indicates that the extensive pre-training of LLMs alone does not mitigate long-tail effects. We conjecture that a modular architecture of models trained on targeted subsets may outperform the single, centralized model of current approaches.

**Reasoning traces improve supervised fine-tuning.** To test whether performance gains stem from reasoning supervision or outcome-only tuning, we fine-tuned baseline models on (i) reasoning + outcome and (ii) outcome-only targets. Table 4 shows that removing reasoning traces leads to large F1 declines, in some cases below baseline performance.

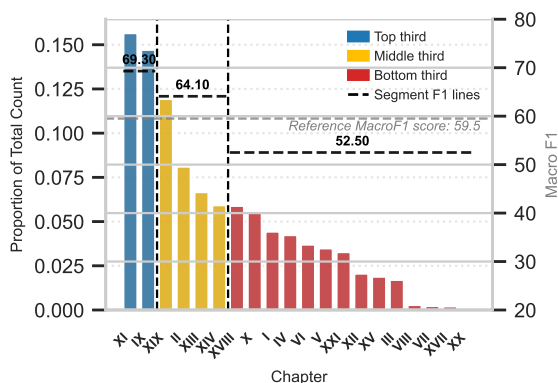


Figure 4: Distribution of samples per ICD-10-CM chapter in the MIMIC-Reason dataset. The imbalance is evident: one third of all cases stem from only two chapters, while the final third spans 13 chapters with considerably worse performance.

Model	Chapter	Category	Full Code
	F1	F1	F1
Baselines			
Qwen	39.0	3.4	0.2
Llama	17.1	1.4	0.1
SFT - Trace + Outcome			
Qwen	48.9	12.6	3.5
Llama	<b>52.9</b>	<b>14.0</b>	<b>3.9</b>
SFT - Outcome only			
Qwen	19.6	1.1	0.1
Llama	23.9	1.0	0.1

Table 4: Comparison of Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct as baseline, supervised fine-tuned on outcome with traces, and on outcome only. The results strongly suggest that models benefit from our traces in supervised fine-tuning. In the case of Qwen, training without traces even harms the model.

This indicates that explicit reasoning steps guide models toward more reliable diagnostic representations and calibration.

**Descriptive review of reasoning coverage.** Figure 5 illustrates the relation between an input admission note and the reasoning trace produced by **DeepICD-R1-7B** for a synthetic case from MTSamples.<sup>5</sup>

The model captures most patient information but omits several details relevant for clinical assessment. We see in Figure 5 that the reasoning follows a structured progression resembling human diagnostic logic (Röhr et al., 2025): initial aggregation of subjective symptoms (steps 1-2 in Figure 5), integration of objective examination findings

<sup>5</sup><https://www.mtsamples.com>

(3), and final synthesis into a diagnostic conclusion (4). Descriptively, this output demonstrates strong grounding between the extracted evidence and the predicted diagnosis. However, without external verification or expert review, some reasoning segments may only appear plausible rather than clinically valuable.

**Qualitative clinician assessment.** A qualitative evaluation with clinicians compared **DeepICD-R1-7B** (SFT+GRPO), **DeepICD-R1-zero-32B** (GRPO only), and the **Qwen2.5-32B-Instruct** baseline, focusing on cases in which DeepICD-R1-7B underperformed quantitatively. Clinicians provided structured feedback on reasoning quality, interpretability, and clinical plausibility. Key observations include:

**Premature diagnostic closure.** As illustrated in Figure 5, DeepICD-R1-7B frequently infers a diagnosis at the beginning of its reasoning process. Clinicians confirmed that in incorrect cases, subsequent reasoning tends to retroactively justify this early hypothesis, occasionally introducing minor hallucinations or inconsistent causal links.

**Emergent clinical reasoning in DeepICD-R1-zero-32B.** Clinicians indicated that DeepICD-R1-zero-32B often produces coherent reasoning traces that identify and contextualize symptoms, despite lacking supervision. In contrast, the Qwen2.5-32B baseline generates denser textual summaries but provides less interpretive linkage between symptoms and diagnosis.

**Lack of awareness of diagnostic significance.** Across all reviewed cases, clinicians observed that the models often failed to recognize the clinical severity or implications of a diagnosis. For instance, the model may predict *angina* when the described symptoms clearly indicate *unstable angina*, a more critical condition requiring urgent care.

**Preference for concise reasoning.** Clinicians consistently favored shorter, well-structured traces, citing improved interpretability and a lower risk of hallucinated content.

Overall, while **DeepICD-R1-7B** achieves the highest quantitative accuracy, it performs less favorably in qualitative evaluation due to its tendency toward premature diagnostic closure and limited awareness of diagnostic severity. Clinicians generally preferred the **DeepICD-R1-zero-32B** model for its concise, contextually grounded reasoning, despite its more limited diagnostic depth. These findings suggest that future reasoning-oriented clinical models must not only improve interpretability and factual grounding but also develop a calibrated understanding of disease severity and clinical significance.

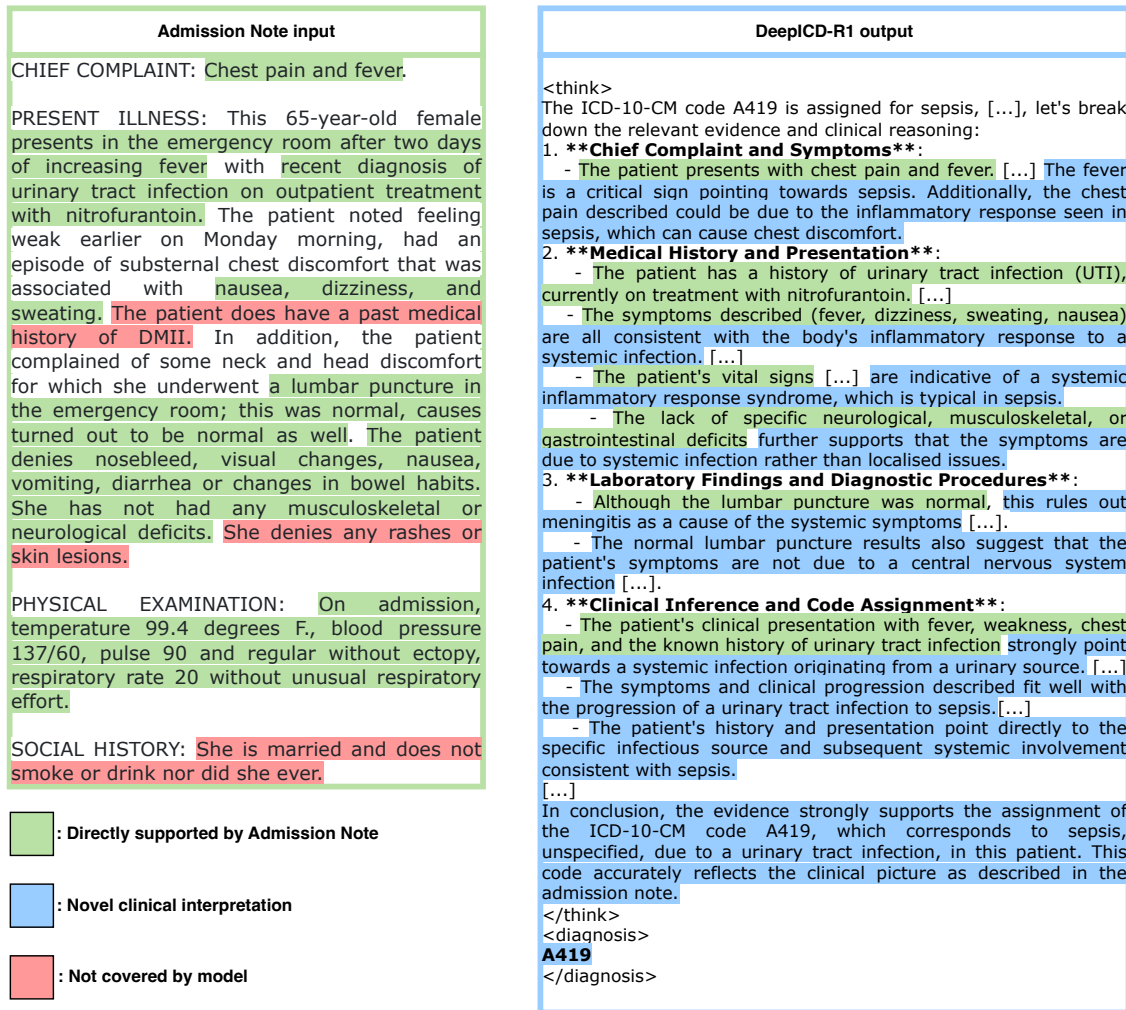


Figure 5: Analysis of a synthetic admission note and the corresponding reasoning trace generated by **DeepICD-R1-7B**. Text directly supported by the admission note is highlighted in green, model-generated clinical inferences in blue, and information from the note not reflected in the reasoning trace in red. The model successfully grounds most extracted evidence in the source text, though certain clinically relevant details remain unaddressed.

## 6. Conclusion

We introduced **DeepICD-R1**, a dataset and framework for training large reasoning models (LRMs) on hierarchical ICD-10-CM outcome prediction. Our approach combines **Group Relative Policy Optimization (GRPO)** with outcome- and process-based reward functions to enable verifiable, clinically grounded reasoning without domain-specific pretraining. We further curated complementary resources to enhance reward computation and supervised fine-tuning (SFT) data generation.

Through extensive experiments, we demonstrated that models only trained with both, our SFT data and hierarchical rewards, outperform all baselines in both reasoning quality and diagnostic accuracy. We analyzed the effects of class imbalance

on performance, highlighting the need for data-aware training strategies. Our results further show that reasoning traces, rather than fine-tuning alone, drive substantial performance improvements. Finally, we presented examples of model-generated reasoning traces and discussed qualitative characteristics, like premature diagnostic closure and lack of awareness for diagnostic significance.

**Future Work** Future research on DeepICD-R1 should address several remaining challenges in diagnostic reasoning.

- Clinician feedback revealed that generated explanations can appear plausible while omitting critical differential diagnoses; incorporating in-reasoning verification modules to assess diag-

nostic validity during generation could enhance trustworthiness.

- While the model performs strongly at the chapter level, performance declines for fine-grained category and full-code prediction due to data imbalance. Addressing this limitation may require improved supervised fine-tuning initialization, more diverse reward formulations, and further data augmentation.
- Extending the framework toward multi-label reasoning, where models infer interdependent comorbidities rather than a single diagnosis, would better reflect real-world clinical complexity and move clinical LRMs closer to practical utility.

## 7. Ethical Considerations

All data used in DeepICD-R1 come from de-identified MIMIC-IV records under the PhysioNet data-use agreement. No identifiable patient information was accessed or generated. The model is a research prototype and must not be used for real-world diagnosis or clinical decision-making. Generated reasoning traces can appear plausible but may contain clinically incorrect statements. Biases inherent in MIMIC-IV—including demographic and disease-frequency skew—may propagate to model outputs. Future work should include bias auditing and fairness-aware fine-tuning. All code, reward functions, and processing scripts will be released for transparency and reproducibility, while derived reasoning data will remain under research-only terms to ensure privacy compliance.

## 8. Limitations

Despite its convincing results, DeepICD-R1 has several limitations. First, all experiments rely on English-language MIMIC-IV admission notes. While this corpus is widely used, its institutional and demographic bias limits generalizability to other healthcare settings and languages. Second, the ICD-10-CM label space is highly imbalanced. Rare codes receive limited supervision, and although hierarchical rewards mitigate this effect, long-tail conditions remain difficult to model reliably. Third, the reasoning traces, while coherent and interpretable, are not guaranteed to be clinically correct. Some explanations may contain subtle factual errors or reflect confirmation bias toward early diagnostic hypotheses. Fourth, the reinforcement learning signal is derived from automatically verifiable rewards and an LLM-as-a-judge metric. These scalable proxies lack the depth of expert clinical feedback and may overlook safety-critical reasoning flaws. Finally, GRPO fine-tuning, though more efficient than

full retraining, still demands significant compute and parameter tuning, which may limit accessibility for smaller research groups.

## 9. Acknowledgements

We would like to thank the reviewers for their helpful suggestions and comments.

Our work is funded by the German Federal Ministry of Education and Research (BMBF) 01|S23015A (AI4SCM) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 528483508 - FIP 12, the European Union under the grant project 101079894 (COMFORT - Improving Urologic Cancer Care with Artificial Intelligence Solutions) and Federal Ministry of Economic Affairs and Energy under grant agreement 13IPC040D (SOOFI – Sovereign Open Source Foundation Models).

## 10. Bibliographical References

- Junying Chen, Zhenyang Cai, Ke Ji, et al. 2025. [Towards medical complex reasoning with LLMs through medical verifiable problems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#).
- Alexei Figueroa, Jens-Michalis Papaioannou, Conor Fallon, Alexandra Bekiaridou, Keno Bressemer, Stavros Zanos, Felix Gers, Wolfgang Nejdl, and Alexander Löser. 2024. [Boosting Long-Tail Data Classification with Sparse Prototypical Networks](#). In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part VII*, pages 434–449, Berlin, Heidelberg. Springer-Verlag.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. [The llama 3 herd of models](#).
- Paul Grundmann, Dennis Fast, Jan Frick, Thomas Steffek, Felix Gers, Wolfgang Nejdl, and Alexander Löser. 2025. [CliniBench: A Clinical Outcome Prediction Benchmark for Generative and Encoder-Based Language Models](#).

- Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. 2025. [Don't Overthink it. Preferring Shorter Thinking Chains for Improved LLM Reasoning.](#)
- J.A. Hirsch, G. Nicola, G. McGinty, R.W. Liu, R.M. Barr, M.D. Chittle, and L. Manchikanti. 2016. [Icd-10: History and context.](#) *American Journal of Neuroradiology*, 37(4):596–599.
- Kerui Huang, Shuhan Liu, Xing Hu, Tongtong Xu, Lingfeng Bao, and Xin Xia. 2025. [Reasoning efficiently through adaptive chain-of-thought compression: A self-optimizing framework.](#)
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. [Mimic-iv, a freely accessible electronic health record dataset.](#) *Scientific Data*, 10(1):1.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention.](#)
- Ilya Loshchilov and Frank Hutter. 2017. [Fixing weight decay regularization in adam.](#) *CoRR*, abs/1711.05101.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling.](#)
- James Mullenbach, Sarah Wiegrefe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. [Explainable Prediction of Medical Codes from Clinical Text.](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback.](#) In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Van L. Parsons. 2017. *Stratified Sampling*, pages 1–11. John Wiley & Sons, Ltd.
- John Pearn. 2022. [Herbert french \(1875-1951\) and his differential diagnosis a “work of reference unique in medical literature”.](#) *J. Med. Biogr.*, 30(2):131–135.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. 2025. [Qwen2.5 technical report.](#)
- Tom Röhr, Soumyadeep Roy, Fares Al Mohamad, Jens-Michalis Papaioannou, Wolfgang Nejdl, Felix Gers, and Alexander Löser. 2025. [“where does it hurt?” – dataset and study on physician intent trajectories in doctor patient dialogues.](#) In *Frontiers in Artificial Intelligence and Applications*, Frontiers in artificial intelligence and applications. IOS Press.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal Policy Optimization Algorithms.](#)
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024a. [DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models.](#)
- Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. 2024b. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models.](#)
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. [Hybridflow: A flexible and efficient rlhf framework.](#) In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys '25*, page 1279–1297, New York, NY, USA. Association for Computing Machinery.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkumar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. [Large language models encode clinical knowledge.](#) *Nature*, 620(7972):172–180.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou,

- Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31(3):943–950.
- Betty van Aken, Jens-Michalis Papaioannou, Marcel Naik, Georgios Eleftheriadis, Wolfgang Nejdl, Felix Gers, and Alexander Loeser. 2022. [This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 172–184, Online only. Association for Computational Linguistics.
- Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. [Direct: Diagnostic reasoning for clinical notes via large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 74999–75011. Curran Associates, Inc.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. [Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs](#).
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. [Limo: Less is more for reasoning](#).
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiase Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. [DAPO: An Open-Source LLM Reinforcement Learning System at Scale](#).
- Jixiao Zhang and Chunsheng Zuo. 2025. [GRPO-LEAD: A Difficulty-Aware Reinforcement Learning Approach for Concise Mathematical Reasoning in Language Models](#).
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: less is more for alignment](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.

## 11. Language Resource References

- A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. 2000. [PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals](#). *Circulation*, 101(23):E215–220.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. [MIMIC-IV](#).
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV-Note: Deidentified free-text clinical notes](#).
- Tom Röhr, Alexei Figueroa, Jens-Michalis Papaioannou, Conor Fallon, Keno Bressem, Wolfgang Nejdl, and Alexander Löser. 2024. [Revisiting clinical outcome prediction for MIMIC-IV](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 208–217, Mexico City, Mexico. Association for Computational Linguistics.
- Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. [Clinical outcome prediction from admission notes using self-supervised knowledge integration](#). In *Proceedings of the 16th Conference of the European*

*Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.