

Eliciting Medical Reasoning with Knowledge-enhanced Data Synthesis: A Semi-Supervised Reinforcement Learning Approach

Haolin Li^{1,2}, Shuyang Jiang^{1,2}, Ruipeng Zhang⁵,
Jiangchao Yao^{3,4}, Ya Zhang^{4,2,6}, Yanfeng Wang⁴✉

¹College of Computer Science and Artificial Intelligence, Fudan University ²Shanghai AI Laboratory

³CMIC, Shanghai Jiao Tong University ⁴School of Artificial Intelligence, Shanghai Jiao Tong University

⁵Department of Radiology, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine

⁶Institute of Artificial Intelligence for Medicine, Shanghai Jiao Tong University School of Medicine

Abstract

While large language models hold promise for complex medical applications, their development is hindered by the scarcity of high-quality reasoning data. To address this issue, existing approaches typically distill chain-of-thought reasoning traces from large proprietary models via supervised fine-tuning, then conduct reinforcement learning (RL). These methods exhibit limited improvement on underrepresented domains like rare diseases while incurring substantial costs from generating complex reasoning chains. To efficiently enhance medical reasoning, we propose MedSSR, a Medical Knowledge-enhanced data Synthesis and Semi-supervised Reinforcement learning framework. Our framework first employs rare disease knowledge to synthesize distribution-controllable reasoning questions. We then utilize the policy model itself to generate high-quality pseudo-labels. This enables a two-stage, intrinsic-to-extrinsic training paradigm: self-supervised RL on the pseudo-labeled synthetic data, followed by supervised RL on the human-annotated real data. MedSSR scales model training efficiently without relying on costly trace distillation. Extensive experiments on Qwen and Llama demonstrate that our method outperforms existing methods across ten medical benchmarks, achieving up to **+5.93%** gain on rare-disease tasks. Our code is available at <https://github.com/tdlh1/MedSSR>.

1 Introduction

Large language models (LLMs) have demonstrated remarkable reasoning capabilities (Brown et al., 2020; Wei et al., 2022; Achiam et al., 2023; Chiang et al., 2023), spurring the development of diverse post-training methods (Ouyang et al., 2022; Rafailov et al., 2023). Among them, various methods have been designed to enhance the reasoning

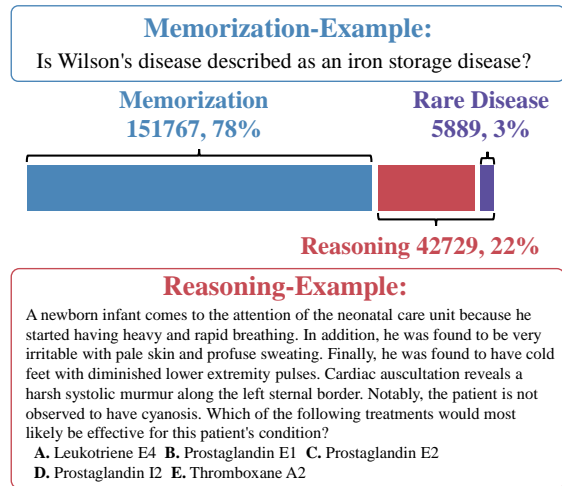


Figure 1: Distribution of question types in existing medical datasets. Only 22% are reasoning-intensive, and just 3% among them concern rare diseases.

ability of LLMs in the medical field, aiming to address complex medical problems (Zhang et al., 2023, 2024). However, unlike natural domains where large amounts of reasoning-intensive data are available, downstream domains like healthcare inherently suffer from a scarcity of reasoning-intensive data (Thapa et al., 2025; Chen et al., 2025d; Li et al., 2025). Most existing medical benchmarks consist primarily of memorization-oriented questions, rather than complex reasoning cases. This issue is especially acute for underrepresented subfields like rare diseases, where labeled data is extremely limited (see Figure 1). Due to privacy constraints and the specialized expertise required, acquiring complex medical reasoning data remains exceptionally challenging, yet critical for real-world clinical applications.

Building on this challenge of data scarcity, existing methods for enhancing medical reasoning primarily use large-scale proprietary models to generate chain-of-thought (CoT) reasoning traces (Chen et al., 2025b). The generated traces are then used in supervised fine-tuning to initialize the reason-

✉: Corresponding author

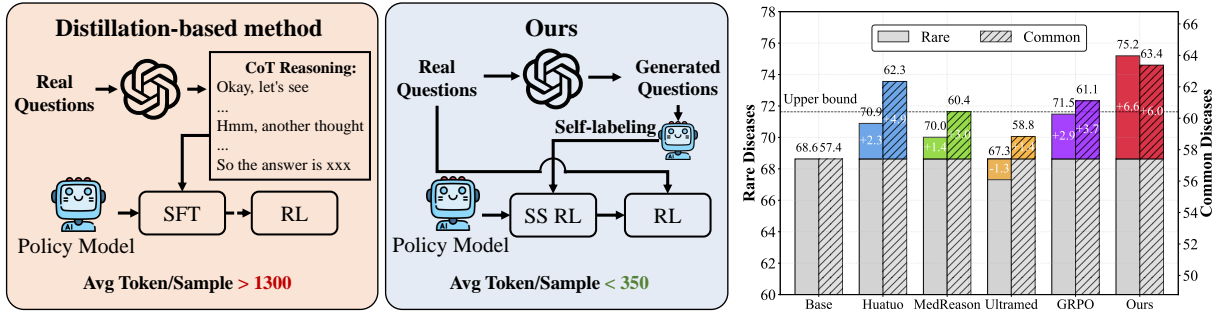


Figure 2: **Left:** Comparison of existing methods for medical reasoning. “SS RL” is short for self-supervised RL. “Avg Token/Sample” denotes the average number of tokens consumed from an API model to generate one sample. **Right:** Performance improvement comparison based on Llama. Prior methods show less improvement in rare diseases than common tasks, while ours significantly breaks the rare disease improvement upper bound of +3%.

ing capabilities of the policy model, followed by RL training for further improvement. While effective, these approaches rely on expensive API calls to proprietary models to distill long reasoning chains (Huang et al., 2025). More importantly, they yield limited gains on rare diseases. As illustrated in Figure 2, we find that all existing methods exhibit significantly lower gains on rare diseases compared to common medical tasks. Even when employing the strong baseline of fully supervised GRPO, performance improvements on rare disease data still fail to surpass a 3% ceiling.

To efficiently enhance medical reasoning without neglecting data-scarce subdomains, synthetic data generation is a promising alternative. Recent studies have explored using synthetic data for LLM post-training (Le et al., 2022; Xu et al., 2024). However, generated data often suffer from low fidelity and factual errors, which are especially unacceptable in medical scenarios (Qin et al., 2025; Das et al., 2025). Therefore, data synthesis in this domain must incorporate structured medical knowledge to prevent factual errors and ensure clinical rationality. Furthermore, how to effectively utilize such synthetic data without introducing additional costs or instability remains a challenging problem.

In this paper, we propose MedSSR to efficiently enhance the medical reasoning capabilities of LLMs through the generation and utilization of synthetic data. Our framework comprises two synergistic components: (1) a medical knowledge-enhanced data synthesis pipeline, and (2) a semi-supervised RL training strategy that jointly leverages synthetic and real data. MedSSR synthesizes complex medical reasoning questions based on real-world seed data. Through an optional, on-demand retrieval of rare disease knowledge, it al-

lows explicit control over the proportion of rare disease data. This enables significant improvement in rare diseases without sacrificing general effectiveness (see Figure 2). To avoid reward hacking, we employ the base model itself to assign pseudo-labels to the synthetic questions via offline majority voting, ensuring the data is well-matched to the model’s learning trajectory. Finally, our training strategy follows an intrinsic-to-extrinsic learning curriculum: we first perform self-supervised RL on pseudo-labeled synthetic data, followed by supervised RL on human-annotated real data. In summary, our contributions are threefold:

- We identify that existing methods for medical reasoning suffer from suboptimal improvement on rare diseases. We propose a more effective and efficient paradigm via data synthesis and semi-supervised RL.
- We propose MedSSR, a comprehensive framework for leveraging synthetic data, comprising (a) medical knowledge-enhanced data synthesis and (b) a semi-supervised RL training strategy. This enables controllable data generation and efficient training scale-up.
- Extensive experiments on both Qwen3-8B and Llama-3.1-8B-Instruct demonstrate that our approach substantially outperforms baselines and existing medical LLMs, with gains up to **5.93%** on rare disease and **3.91%** in general.

2 Related Work

2.1 Medical Reasoning Models

Tremendous efforts have been made to develop medical reasoning models (Huang et al., 2025). HuatuoGPT-O1 (Chen et al., 2025b) distills reasoning traces from GPT-4o via supervised fine-tuning, and then applies a second-stage RL train-

ing. MedReason employs knowledge graphs during CoT generation to improve factual accuracy (Wu et al., 2025). Another line of work focuses on training powerful process reward models (PRM) to supervise the policy model. MedS³ (Jiang et al., 2025) and MedPRM (Yun et al., 2025) first generate reasoning paths for PRM training, then use the PRM to supervise the policy model. While effective, these methods fail to provide considerable gains for rare diseases. As illustrated in Figure 2, MedSSR employs proprietary models only to generate questions instead of long reasoning traces, and trains the policy model via pure RL. This pipeline bypasses the cost of PRM deployment and trace distillation, achieving higher efficiency while maintaining superior performance on rare diseases.

2.2 Data Generation via LLMs

Generating synthetic data has emerged as a powerful strategy to mitigate data scarcity, particularly in domains like healthcare, where collecting real-world data is challenging (Yu et al., 2025; Guo et al., 2025b). The advent of LLMs has significantly accelerated this paradigm. Early works like Self-Instruct (Wang et al., 2023) and Alpaca (Taori et al., 2023) leveraged LLMs to generate instruction-following data, while later efforts like WizardMath (Luo et al., 2025) focused on synthesizing complex math problems. Currently, synthetic data has become an indispensable component for scaling up LLM training (Chen et al., 2025c). Compared to prior approaches, our method tailors data generation to the medical domain by injecting rare disease knowledge, enhancing the factual accuracy and clinical relevance.

3 Method

In this section, we first present the problem formulation in § 3.1. Then, we introduce the details of our method in two parts. § 3.2 describes the data synthesis pipeline. § 3.3 presents the semi-supervised training strategy. The overall framework of MedSSR is illustrated in Figure 3.

3.1 Preliminary

Let $\mathcal{D} = \{(x, y)\}$ denote the training set, where x is a question and y is the ground-truth answer. We aim to optimize an LLM π_θ , parameterized by θ , which takes a question x as input and generates a response y' autoregressively. Typically, the response is a reasoning trace with a final answer $\text{ans}(y')$, where $\text{ans}(\cdot)$ is the function for answer extraction.

The objective is to maximize the verification reward $r(y, y')$, which evaluates the correctness of the answer against the ground truth y :

$$r(y, y') = \mathbb{I}[\text{ans}(y') = y],$$

where $\mathbb{I}[\cdot]$ is the indicator function. The optimization problem can be formulated as:

$$\max_{\pi_\theta} \mathbb{E}_{(x, y) \in \mathcal{D}, y' \sim \pi_\theta(x)} \left[\hat{A}(x, y') - \beta \cdot \mathbb{D}_{\text{KL}}[\pi_\theta(y'|x) \parallel \pi_{\text{ref}}(y'|x)] \right], \quad (1)$$

where $\hat{A}(x, y')$ denotes the advantage estimated on the reward $r(y, y')$ and the coefficient β controls the Kullback-Leibler (KL) (Kullback and Leibler, 1951) divergence between the policy model π_θ and the initial policy π_{ref} .

3.2 Knowledge-enhanced Data Synthesis

To expand the knowledge coverage of medical LLMs, especially on rare diseases, we introduce a distribution-controllable data synthesis pipeline.

Data Synthesis Given two seed questions $\{x_1^s, x_2^s\} \subset \mathcal{D}$, we generate new synthetic questions using a powerful proprietary LLM (e.g., GPT-4.1 (Hurst et al., 2024)), denoted as LLM_{gen} . To ground the generation in factual medical knowledge, we augment the question generation pipeline with a comprehensive medical knowledge corpus \mathcal{P} , which integrates multiple authoritative sources. To precisely control the proportion of rare disease content in the synthetic dataset, we introduce a threshold $\alpha \in [0, 1]$. For each generation step, we sample a random value $\rho \sim \text{Uniform}(0, 1)$. If $\rho < \alpha$, we randomly select a rare disease entity e from a predefined rare disease list E_{rare} as the query to retrieve the top- k most relevant medical documents from \mathcal{P} :

$$\mathcal{C}(e) = \underset{p \in \mathcal{P}}{\text{Top-}k \text{ Sim}(e, p)},$$

where $\text{Sim}(\cdot, \cdot)$ denotes the similarity score, which is computed by the dot product of the embeddings of e and p . These embeddings are encoded by the widely-adopted medical retriever MedCPT (Jin et al., 2023). The retrieved rare disease context $\mathcal{C}(e)$ is then concatenated with the seed questions (x_1^s, x_2^s) to form the input prompt for LLM_{gen} .

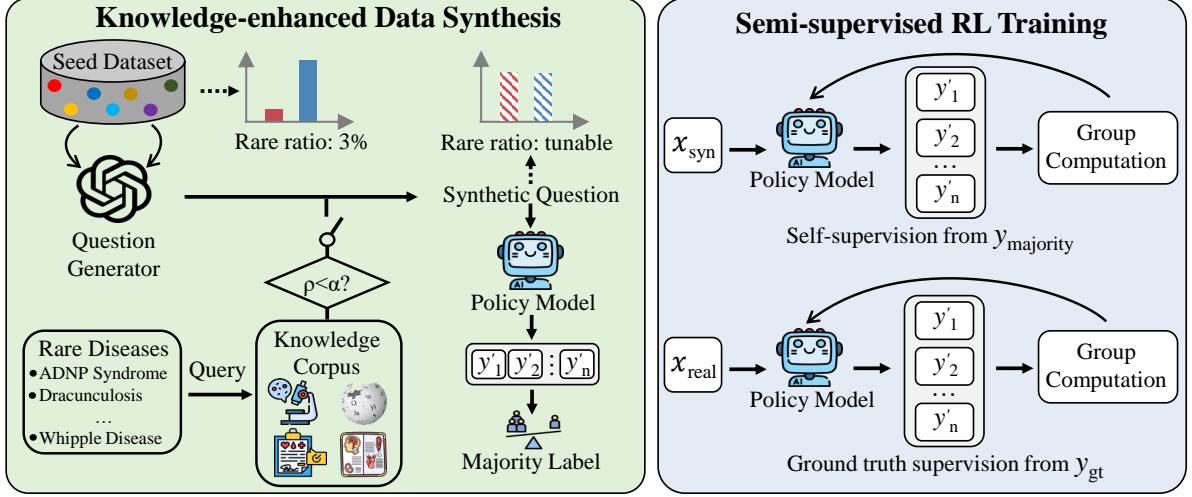


Figure 3: Overview of the proposed method. We first synthesize questions with tunable distribution, then train the policy model via a semi-supervised reinforcement learning mechanism. Rare disease knowledge injection is triggered when the random value ρ is less than the threshold α .

If the random value $\rho \geq \alpha$, no additional disease-specific knowledge is injected. The synthesis process can then be formulated as:

$$x_{\text{syn}} = \begin{cases} \text{LLM}_{\text{gen}}(x_1^s, x_2^s, \mathcal{C}(e)), & \text{if } \rho < \alpha, \\ \text{LLM}_{\text{gen}}(x_1^s, x_2^s), & \text{if } \rho \geq \alpha, \end{cases} \quad (2)$$

where x_{syn} is the generated question. Following this procedure, we can obtain a scalable synthetic dataset \mathcal{D}_{syn} with a tunable focus on rare diseases.

For more details of the rare disease list E_{rare} , knowledge corpus \mathcal{P} , and the prompt used for generation, please refer to Appendix B.

3.3 Semi-supervised RL training

To utilize the synthetic dataset, we perform self-labeling to avoid labor-intensive human labeling.

Pseudo-label Generation For high-quality data tailored for policy model training, we leverage π_θ itself to generate pseudo-labels. Given a synthetic question $x_{\text{syn}} \in \mathcal{D}_{\text{syn}}$, we sample G independent responses from the base policy model:

$$\{y_{\text{syn}}^1, y_{\text{syn}}^2, \dots, y_{\text{syn}}^G\} \sim \pi_\theta(\cdot | x_{\text{syn}}).$$

From each response y_{syn}^i , we extract a final answer a_{syn}^i . The pseudo-label y_{majority} is then determined by majority voting over these G answers:

$$y_{\text{majority}} = \arg \max_a \sum_{i=1}^G \mathbb{I}[a_{\text{syn}}^i = a]. \quad (3)$$

For samples within the knowledge scope of the base model, which constitute the vast majority of

the questions, the majority-voted answers are reliable. Even when y_{majority} is incorrect, the reward assignment remains partially valid since other wrong answers still receive zero reward, just as they would under correct supervision.

The resulting labeled dataset $\mathcal{D}_{\text{syn}} = \{(x_{\text{syn}}, y_{\text{majority}})\}$ is then used for self-supervised RL, providing model-aligned intrinsic reward signals. Crucially, this offline labeling strategy mitigates the reward hacking problem commonly observed in online self-supervised methods, where the policy model collapses to output identical answers across generations to exploit the reward. In § 4.4 and Appendix D.2, we provide empirical analysis that this offline voting strategy enables long-term stable training for more than 1000 steps.

Self-supervised Training We employ the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the algorithm for Eq. (1), which provides a stable and efficient computation for advantages. The training proceeds in two sequential phases: self-supervised training on synthetic data, followed by supervised training on real data.

The policy model is first optimized on \mathcal{D}_{syn} to incentivize its intrinsic reasoning ability. Given a sample $(x_{\text{syn}}, y_{\text{majority}})$, the reward signal is:

$$r(y_{\text{majority}}, y_{\text{syn}}) = \mathbb{I}[\text{ans}(y_{\text{syn}}) = y_{\text{majority}}].$$

For each question x_{syn} , we sample a group of G outputs from the current policy model. The group-wise advantage for each output y_{syn}^i is computed by

Model	Symptoms	Causes	Diagnosis	Others	RDs	Treatment	Affects	Avg
Large Language Models								
Qwen3-8B	68.63	72.01	57.95	66.44	70.39	52.18	68.28	65.13
R1-Distill-Llama-8B	53.58	63.04	45.49	60.19	55.52	46.57	63.28	55.38
Medical Large Language Models								
UltraMedical-3.1-8B	73.53	73.01	57.28	65.28	77.47	50.83	73.79	67.31
HuatuoGPT-o1-8B	74.90	78.47	61.64	70.10	79.53	52.89	78.68	70.89
m1-7B-23K	74.51	74.64	57.82	69.44	68.59	52.37	72.76	67.16
MedPRM-8B	73.14	76.00	61.79	66.20	78.45	52.37	78.97	69.56
MedReason-8B	73.53	74.39	62.85	70.43	76.43	57.10	75.33	70.01
Qwen3-8B-Base	67.16	70.74	54.92	68.75	59.05	49.30	67.41	62.48
+ Fully-Supervised	73.24	75.41	60.65	65.97	70.39	54.41	72.24	67.47
+ MedSSR	79.71	80.89	68.19	72.22	76.64	57.51	78.62	73.40
↑	6.47	5.48	7.54	6.25	6.25	3.10	6.38	5.93
Llama-3.1-8B-Instruct	71.67	74.28	60.85	64.81	79.28	52.97	76.55	68.63
+ Fully-Supervised	76.27	77.72	60.18	67.13	81.09	54.55	83.45	71.48
+ MedSSR	80.00	80.30	64.96	71.99	82.89	59.79	86.38	75.19
↑	3.73	2.58	4.78	4.86	1.80	5.24	2.93	3.70

Table 1: Performance on rare diseases. “RDs” is short for Related Disorders. The best results are in **Bold**. ↑ denotes the relative improvement of our method compared to the fully-supervised baseline. Ours is highlighted in gray.

normalizing its reward against the group statistics:

$$\hat{A}_{\text{syn}}^i = \frac{r(y_{\text{majority}}, y_{\text{syn}}^i) - \mu_{\text{syn}}}{\sigma_{\text{syn}}}, \quad (4)$$

where μ_{syn} and σ_{syn} are the mean and standard deviation of rewards within the group. The optimization objective is then formulated as:

$$\begin{aligned} \mathcal{J}_{\text{self}}(\theta) = & \mathbb{E}_{(x_{\text{syn}}, y_{\text{majority}}) \in \mathcal{D}_{\text{syn}}, \{y_{\text{syn}}^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | x_{\text{syn}})} \\ & \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_{\text{syn}}^i|} \sum_{t=1}^{|y_{\text{syn}}^i|} \left(\min \left[c_{i,t} \hat{A}_{\text{syn}}^i, \text{clip}(c_{i,t}, \right. \right. \right. \\ & \left. \left. \left. 1 - \epsilon, 1 + \epsilon) \hat{A}_{\text{syn}}^i \right] - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}) \right) \right], \quad (5) \end{aligned}$$

where

$$c_{i,t} = \frac{\pi_{\theta}(y_{\text{syn},t}^i | x_{\text{syn}}, y_{\text{syn},<t}^i)}{\pi_{\theta_{\text{old}}}(y_{\text{syn},t}^i | x_{\text{syn}}, y_{\text{syn},<t}^i)}$$

and ϵ is a clipping hyperparameter. This phase leverages self-supervised signals to fully exploit the reasoning capacity within the ability boundary of the policy model.

Fully-supervised Training To further extend the model’s capability beyond its intrinsic boundary, we subsequently employ external supervision signals from real-world data. The training procedure

is analogous to the self-supervised phase described in Eq. (5), but with the reward computed using ground-truth answers. Specifically, for a real sample $(x, y) \in \mathcal{D}$, the reward is defined as:

$$r(y, y^i) = \mathbb{I}[\text{ans}(y^i) = y],$$

and the group-wise advantage is:

$$\hat{A}^i = \frac{r(y, y^i) - \mu}{\sigma}, \quad (6)$$

where μ and σ are the mean and standard deviation of rewards over the G sampled responses for x .

The supervised phase further expands the model’s reasoning ability based on verified real data. Together, our intrinsic-to-extrinsic training paradigm maximizes the utility of both synthetic and real data, enabling efficient and effective scaling of medical reasoning models.

4 Experiments

4.1 Experimental Setup

Datasets We employ 11 widely adopted medical datasets. Among them, five datasets containing training splits are employed for both training and testing, including: MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2020), BioASQ (Tsatsaronis et al., 2015), HeadQA (Vilares and Gómez-Rodríguez, 2019), and PubMedQA (Jin et al.,

Model	BioASQ	MedMCQA	MedQA	MedXpertqa	MMLU	PubMedQA	NEJM	Lancet	Medbullets	Avg
Large Language Models										
Qwen3-8B	75.32	62.14	66.22	15.91	83.20	48.00	61.48	63.77	48.86	58.32
R1-Distill-Llama-8B	69.82	46.42	46.60	6.28	65.24	47.30	42.08	48.36	31.66	44.86
Medical Large Language Models										
UltraMedical-3.1-8B	81.42	60.75	64.61	13.46	75.16	58.08	64.84	60.14	51.06	58.84
HuatuoGPT-o1-8B	80.97	64.77	77.79	16.68	79.57	58.35	65.92	62.68	54.30	62.34
m1-7B-23K	78.55	62.23	71.84	17.10	78.94	49.90	61.28	60.26	53.82	59.32
MedPRM-8B	79.83	58.54	63.24	15.66	76.10	58.80	63.81	60.31	48.78	58.34
MedReason-8B	77.34	61.56	72.00	19.14	78.38	59.60	64.51	57.34	53.90	60.42
Qwen3-8B-Base	76.92	58.06	62.12	13.80	79.89	49.35	60.74	60.20	46.43	56.39
+ Fully-Supervised	81.91	64.25	74.98	17.40	83.88	54.15	68.74	67.66	55.11	63.12
+ MedSSR	86.00	67.79	81.14	20.66	86.64	54.70	71.97	70.94	63.47	67.03
↑	4.09	3.54	6.16	3.26	2.76	0.55	3.23	3.28	8.36	3.91
Llama-3.1-8B-Instruct	79.80	56.72	62.55	14.29	75.94	59.10	61.29	61.09	45.94	57.41
+ Fully-Supervised	83.41	62.50	72.25	15.46	78.67	58.65	63.85	60.86	54.38	61.11
+ MedSSR	84.91	64.73	74.96	17.04	80.65	58.75	66.62	65.71	57.06	63.38
↑	1.50	2.23	2.71	1.58	1.98	0.10	2.77	4.85	2.68	2.27

Table 2: Performance on general medical tasks. All results are the average of four runs.

2019). The remaining benchmarks with only test splits are used exclusively for evaluation, including: MMLU-Med (Hendrycks et al., 2021), MedXpertqa (Zuo et al., 2025a), Medbullets (Chen et al., 2025a), NEJM, and Lancet (Thapa et al., 2025). For detailed evaluation, we adopt the rare disease focusing benchmark ReDis-QA (Wang et al., 2024) and curate an additional 951 rare-disease questions from the above benchmarks, forming the RareDis-Sub set. We categorize its questions into seven types: Symptoms, Causes, Affects, Diagnosis, Related Disorders, Treatment, and Others. Dataset statistics are detailed in Appendix C.1.

Models and Baselines We choose Qwen3-8B-Base (Yang et al., 2025) and Llama-3.1-8B-Instruct (Dubey et al., 2024) as the backbone, and compare our method against the fully-supervised GRPO baseline. To compare with existing models, we further include baselines in the following categories: (1) General LLMs, including the Qwen3 series, and R1-Distill-Llama-8B (Guo et al., 2025a); (2) Medical LLMs, including HuatuoGPT-o1-8B (Chen et al., 2025b), m1-7B-23K (Huang et al., 2025), MedReason-8B (Wu et al., 2025), MedPRM-8B (Yun et al., 2025), and UltraMedical-3.1-8B (Zhang et al., 2024).

Training and Evaluation Details For RL training, we set the number of rollouts to 8, temperature to 1.0, and response length to 4K. For the fully-supervised baseline, we carefully selected all the reasoning-heavy samples from existing datasets, resulting in a 43K training set. For our method, we adopt 43K generated data for self-supervised train-

ing and use the same 43K real data for supervised training, resulting in a 1:1 ratio of synthetic data vs. real data. During evaluation, we set the recommended decoding hyperparameters for each model to ensure optimal performance. And we report the average accuracy over four runs. More details are provided in Appendix C.2.

4.2 Main Results

By controlling α , we adjust the proportion of rare disease content to 25%, which yields an optimal balance between rare disease and general performance. As presented in Table 1 & 2, we have the following key findings:

Superiority on Data-scarce Domains: MedSSR exhibits significant performance gains on all rare disease tasks. On Qwen, our method achieves an average improvement of 5.93% over the supervised baseline, while on Llama, an average gain of 3.70% is observed, demonstrating the strong generalization across model architectures. Notably, both models outperform all existing (medical) LLMs. The Llama-based model exhibits particularly strong initial performance on rare diseases. MedSSR can further boost it with 6.56%, while other Llama-based baselines show marginal improvement.

Effectiveness on General Data: As shown in Table 2, our method delivers substantial performance improvements not only on rare diseases but also across a broad spectrum of general medical benchmarks. For both models, MedSSR consistently surpasses the GRPO baseline across all nine datasets. Particularly on Qwen, where the supervised baseline already provides a significant gain over 6%,

Stage	BioASQ	MedMCQA	MedQA	MedXpertqa	MMLU	PubMedQA	NEJM	Lancet	Medbullets	Avg
One-stage	82.16	67.10	78.06	17.77	86.04	54.70	69.49	68.32	62.26	65.10
Reverse-stage-I	81.91	64.25	74.98	17.40	83.88	54.15	68.74	67.66	55.11	63.12
Reverse-stage-II	82.83	67.48	78.67	17.27	85.56	53.30	69.36	69.66	60.55	64.96
Two-stage-I	82.51	65.44	75.61	17.09	86.25	55.05	66.91	68.87	55.92	63.74
Two-stage-II	86.00	67.79	81.14	20.66	86.64	54.70	71.97	70.94	63.47	67.03

Table 3: Ablation on training strategy. “One-stage” denotes mixed training without separating synthetic/real data.

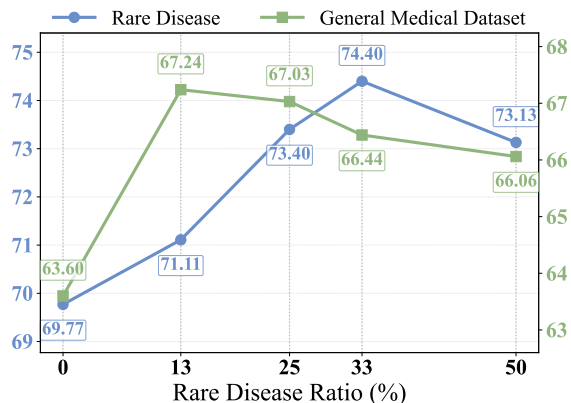


Figure 4: Average performance on rare disease and general datasets across different rare disease ratios.

our method further enhances the average performance by 3.91%, demonstrating that the benefits of MedSSR extend well beyond data-scarce domains. By providing richer and more diverse training data, our method can efficiently facilitate training scale-up. Detailed standard deviations of these results are provided in Appendix D.1.

4.3 Ablation Study

Impact of Rare Disease Ratio To investigate the impact of the proportion of rare disease in the training data, we control α to generate synthetic datasets with varying rare disease ratios (0%, 13%, 25%, 33%, and 50%). The 13% ratio represents the natural distribution obtained when α is always set to 0 (i.e., no rare disease knowledge injection during data synthesis). The experiments are conducted on Qwen3-8B-Base.

Figure 4 illustrates the average performance on both rare disease and general medical tasks. As the ratio increases, performance on both sides first increases and then declines. Notably, general performance reaches its maximum at 13% and degrades as the ratio increases. For rare diseases, peak performance occurs at 33%. Further increasing the ratio to 50% declines the abilities on both sides. The optimal trade-off between rare disease and general performance is achieved at 25%. This ratio balances specialization and generalization, maxi-

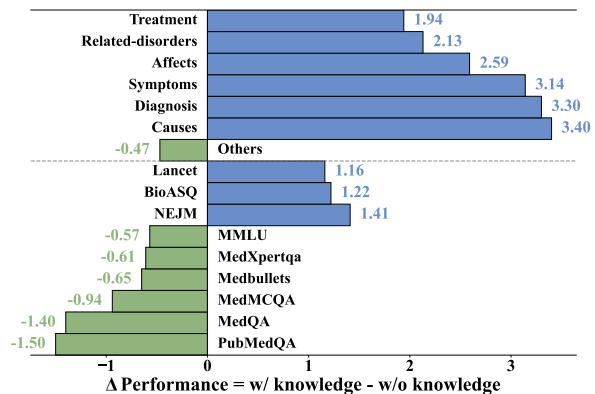


Figure 5: Impact of rare disease knowledge injection. The upper/lower part denotes rare/general tasks.

mizing overall utility for real-world applications.

Impact of Knowledge Injection The difference between the 25% and 13% ratios directly quantifies the impact of rare disease knowledge injection. As detailed in Figure 5, knowledge injection yields substantial gains across six rare disease categories. On general benchmarks, performance improves on three datasets while slightly drops on others. Overall, the large gains on rare diseases (+2.29 in average) outweigh the minimal losses in general (-0.21). MedSSR enables targeted enhancement in data-scarce domains without compromising overall generalizability. The ability to precisely control this trade-off makes our method adaptable for various scenarios. More detailed results of Figure 4 and 5 on each dataset can be found in Appendix D.3.

Ablation on the Training Strategy In this section, we provide an ablation study of our two-stage training strategy. We compare our method with a single-stage training strategy that does not separate synthetic data and real data. And a reversed two-stage strategy that first performs supervised training, then self-supervised training. As presented in Table 3, our two-stage strategy consistently outperforms the single-stage and the reversed strategy on all benchmarks, validating the necessity of the proposed training paradigm: self-supervised learning first elicits the model’s intrinsic capability,

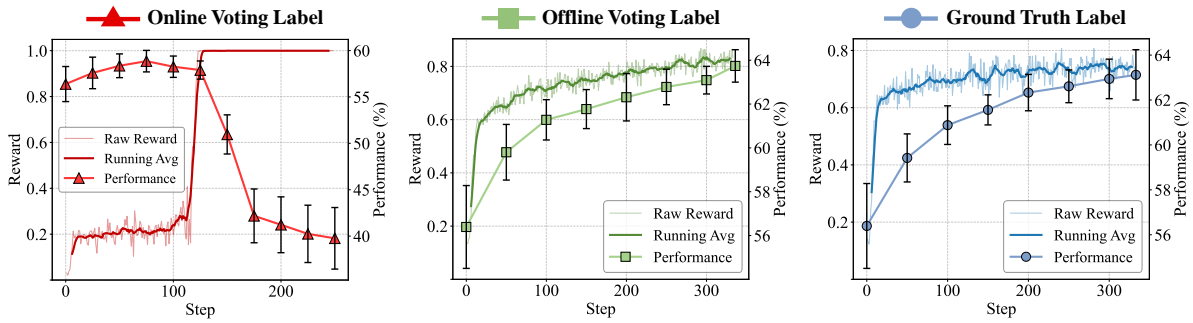


Figure 6: Rewards (left) and performance (right) curves of different labeling strategies across training steps.

Group	Incorrect	Harmful	Plausibility
Senior	1/200	1/200	4.86/5
Junior	0/200	0/200	4.74/5
Overall	0.50%	0.50%	4.80/5

Table 4: Physician evaluation of synthetic samples.

followed by supervised training to enhance it with external signals. Notably, the stage-I of our strategy matches or even slightly surpasses supervised training on the same amount of real data, proving the high quality of our generated samples.

4.4 Further Discussion

Human Evaluation of Synthetic Data To validate the quality of our synthetic data, we invited eight physicians (four senior, four junior) to independently evaluate a subset of 200 samples (labeled by Qwen3-8B-Base). Each sample was rated for correctness, harmfulness, and plausibility (1-5 scale). As summarised in Table 4, only one sample (0.50%) was marked incorrect and harmful by senior doctors. The average plausibility score reached 4.80/5.0. Both senior and junior groups gave consistently high ratings, confirming the clinical reliability of our generated data.

Offline Labeling Prevents Model Collapse In self-supervised RL, the most straightforward approach is to *online* vote pseudo-labels at each training step. However, this strategy is prone to reward hacking. In this section, we empirically analyze why *offline* labeling is critical for stable training.

Figure 6 presents the training dynamics of: (a) online voting, (b) offline voting, and (c) ground-truth supervision. Under online voting, both reward and performance initially rise but soon diverge sharply: the reward rapidly surges to 1.0, while the performance drops substantially, indicating similar outputs across all rollouts. Similar phenomena also occur in other self-supervised

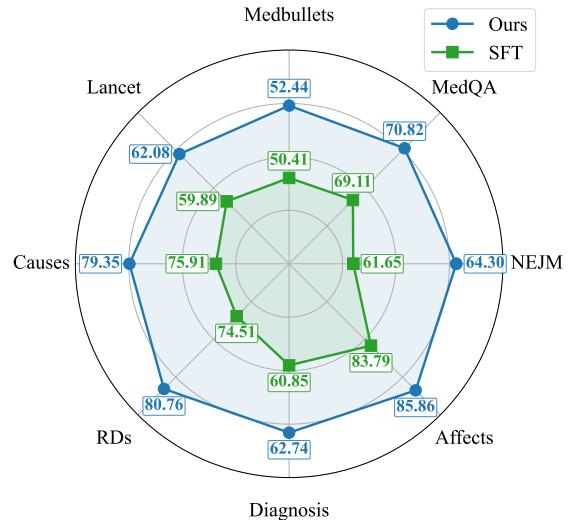


Figure 7: Comparison of our question-only synthesis and long reasoning chain synthesis (SFT).

methods based on entropy or other metrics.

In contrast, offline voting yields stable training curves that closely resemble ground-truth labels: both reward and performance steadily increase. Decoupling label generation from the training loop can effectively prevent reward hacking.

Question-only vs. Reasoning Chain To validate that our question-synthesis approach outperforms the prior paradigm of synthesizing entire reasoning chains, we directly compare them on Llama-3.1-8B-Instruct. Specifically, “Ours” denotes training solely on synthetic questions (43K) via self-supervised RL, and “SFT” represents training on the HuatuoGPT-o1 dataset (20K) with 3 epochs via supervised fine-tuning. As shown in Figure 7, our question-only approach consistently surpasses the SFT method across all tasks. These results demonstrate that our method not only reduces generation costs but also delivers superior performance, establishing a more efficient and effective pathway for scaling medical reasoning models.

5 Conclusion

In this work, we introduce MedSSR to address the data scarcity and limited improvement in rare diseases when developing medical reasoning models. Our method includes a medical knowledge-enhanced data synthesis pipeline for generating distribution-controllable questions and a semi-supervised RL training strategy, which first elicits the model’s intrinsic ability on synthetic data and then enhances it with ground-truth supervision. Evaluations across diverse benchmarks and architectures demonstrate the effectiveness of MedSSR.

Limitations

Although we have taken concrete steps to mitigate its impact, some limitations remain in our work. First, due to computational constraints, we did not include experiments on larger LLMs. Currently, we have tried our best to conduct experiments up to 14B parameter scales. However, our framework enables efficient scaling. We believe the core methodology, knowledge-enhanced data synthesis coupled with semi-supervised RL, would generalize effectively to larger models, potentially yielding even greater performance gains as model capacity increases.

Another limitation is that the synthetic data of MedSSR has not been fully validated by human experts. To mitigate this concern, we have invited eight physicians to assess the factual correctness and clinical plausibility of a subset. Detailed results and analysis in Appendix B.4 indicate that the majority of generated samples were rated as clinically valid. This provides preliminary evidence for the reliability of our synthetic data. Nevertheless, full-scale expert verification remains desirable for real-world deployment.

Ethical Consideration

The deployment of large language models in healthcare raises ethical considerations regarding data privacy and real-world applicability. All datasets used in our work are publicly available de-identified medical datasets under licenses that permit research use, and their respective terms are respected in our study. We ensure that no personally identifiable information is leaked to the question generator (GPT-4.1) or other models that are evaluated.

An advantage of our method is that its enhanced reasoning ability offers more transparent outputs. By outputting the detailed reasoning process, the

model allows medical experts to better understand the response and assess its reliability in clinical contexts. While our framework demonstrates strong performance on established benchmarks, it is primarily intended for research purposes to advance medical reasoning capabilities. Any potential real-world deployment would require rigorous clinical validation and expert oversight to ensure safety.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62306178) and STCSM (No. 22DZ2229005), 111 plan (No. BP0719010), and Shanghai Artificial Intelligence Laboratory.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. 2015. Omim.org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025b. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573. Association for Computational Linguistics.
- Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025c. Self-questioning language models. *arXiv preprint arXiv:2508.03682*.

- Zhe Chen, Yusheng Liao, Shuyang Jiang, Zhiyuan Zhu, Haolin Li, Yanfeng Wang, and Yu Wang. 2025d. Heterorag: A heterogeneous retrieval-augmented generation framework for medical vision language tasks. *arXiv preprint arXiv:2508.12778*.
- Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhonghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.
- Anindya Bijoy Das, Shahnewaz Karim Sakib, and Shibir Ahmed. 2025. Trustworthy medical imaging with large language models: A study of hallucinations across modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1265–1272.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. *The llama 3 herd of models*. *CoRR*, abs/2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Yiduo Guo, Zhen Guo, Chuanwei Huang, Zi-Ang Wang, Zekai Zhang, Haofei Yu, Huishuai Zhang, and Yikang Shen. 2025b. Synthetic data rl: Task definition is all you need. *arXiv preprint arXiv:2505.17063*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. *Measuring massive multitask language understanding*. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Dulhan Jayalath, Shashwat Goel, Thomas Foster, Parag Jain, Suchin Gururangan, Cheng Zhang, Anirudh Goyal, and Alan Schelten. 2025. Compute as teacher: Turning inference compute into reference-free supervision. In *NeurIPS 2025 Workshop on Efficient Reasoning*.
- Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds3: Towards medical slow thinking with self-evolved soft dual-sided process supervision. *arXiv preprint arXiv:2501.12051*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. *What disease does this patient have? A large-scale open domain question answering dataset from medical exams*. *CoRR*, abs/2009.13081.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. *Pubmedqa: A dataset for biomedical research question answering*. *CoRR*, abs/1909.06146.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. *Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval*. *Bioinform.*, 39(10).
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

- Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328.
- Haolin Li, Tianjie Dai, Zhe Chen, Siyuan Du, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2025. [RAD: Towards trustworthy retrieval-augmented multi-modal clinical diagnosis](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. 2025. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. *arXiv preprint arXiv:2508.14029*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, et al. 2025. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Ankit Pal, Logesh Kumar Umaphathi, and Malaikanan Sankarasubbu. 2022. [Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering](#). *CoRR*, abs/2203.14371.
- Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.
- Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R Fung, et al. 2025. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. 2025. Can large reasoning models self-train? *arXiv preprint arXiv:2505.21444*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Shuaijie She, Yu Bao, Yu Lu, Lu Xu, Tao Li, Wenhao Zhu, Shujian Huang, Shanbo Cheng, Lu Lu, and Yuxuan Wang. 2025. Dupo: Enabling reliable llm self-verification via dual preference optimization. *arXiv preprint arXiv:2508.14460*.
- 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*, pages 1279–1297.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, et al. 2025. Disentangling reasoning and knowledge in medical large language models. *arXiv preprint arXiv:2505.11462*.
- George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. [An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition](#). *BMC Bioinform.*, 16:138:1–138:28.
- David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966.
- Guanchu Wang, Junhao Ran, Ruixiang Tang, Chiyuan Chang, Yu-Neng Chuang, Zirui Liu, Vladimir Braverman, Zhandong Liu, and Xia Hu. 2024. Assessing and enhancing large language models in rare disease question-answering. *arXiv preprint arXiv:2408.08422*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

- et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Steffanie S Weinreich, R Mangon, JJ Sikkens, ME En Teeuw, and MC Cornel. 2008. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.
- Ridong Wu, Shuhong Chen, Xiangbiao Su, Yuankai Zhu, Yifei Liao, and Jianming Wu. 2024. A multi-source retrieval question answering framework based on RAG. *CoRR*, abs/2405.19207.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Ping Yu, Jack Lanchantin, Tianlu Wang, Weizhe Yuan, Olga Golovneva, Ilia Kulikov, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2025. Cot-self-instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks. *arXiv preprint arXiv:2507.23751*.
- Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, et al. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. Huatuoqpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885.
- Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.
- Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.
- Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. On the evolution of language models without labels: Majority drives selection, novelty promotes variation. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.
- Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025a. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*.
- Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. 2025b. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

Appendix

A More Related Work	13
A.1 Reinforcement Learning for LRMs	13
B Method Details and Verification	13
B.1 Knowledge Corpus	13
B.2 Rare Disease Sources	13
B.3 Prompt for Data Synthesis	13
B.4 Expert verification of synthetic data	13
C Experimental Details	14
C.1 Dataset Details	14
C.2 Implementation Details	15
D Additional Experiments	16
D.1 Standard Deviation of Main Results	16
D.2 Offline Voting Enables Long-term Stable Training	16
D.3 Impact of Rare Disease Ratio	17
D.4 Comparison under Similar Cost	17
D.5 Generalizability across Different Scales	17
E Case Studies	18
E.1 Synthetic Examples	18
E.2 Failure Mode Analysis	18
F Prompt List	19
A More Related Work	

A.1 Reinforcement Learning for LRMs

Reinforcement learning has become a dominant approach for post-training of large reasoning models (LRMs). Early approaches employ the PPO algorithm to align models with human preferences using reward models (Schulman et al., 2017; Ouyang et al., 2022). Subsequent RLVR methods simplify the training pipeline by computing rule-based rewards directly from ground-truth answers (Jaech et al., 2024; Lambert et al., 2024; Guo et al., 2025a). GRPO further enhances reasoning capabilities by optimizing group-wise relative policies (Shao et al., 2024). Another line of research explores self-supervised RL methods, which derive intrinsic rewards from online majority voting answers or response entropy (Shafayat et al., 2025; Zhao et al., 2025). However, these approaches often suffer from reward hacking (Prabhudesai et al., 2025; Zhang et al., 2025). In contrast, MedSSR combines synthetic and real data to form a more stable

semi-supervised RL framework, avoiding reward hacking through offline voting on generated data.

B Method Details and Verification

B.1 Knowledge Corpus

Following prior works (Xiong et al., 2024; Wu et al., 2024), we construct the medical knowledge corpus \mathcal{P} from four authoritative sources: (1) PubMed¹ contains the latest biomedical research articles from the PubMed database. (2) Wikipedia includes general medical concepts and descriptions. (3) StatPearls consists of clinical practice summaries and certification exams. (4) Textbooks provide medical textbooks covering diverse domain-specific knowledge. This multi-source corpus incorporates real-world medical knowledge into the data synthesis process, improving factual accuracy.

B.2 Rare Disease Sources

To select samples related to rare diseases in the datasets, we extracted the rare disease list E_{rare} from four large-scale, certified rare disease databases. The database sources are Orphanet (Weinreich et al., 2008), Online Mendelian Inheritance in Man (Amberger et al., 2015), National Organization for Rare Disorders of the United States, and Compendium of China’s Rare Diseases. After deduplication, the extracted list E_{rare} contains a total of 12,445 rare diseases. E_{rare} is also employed as the retrieval query in § 3.2.

B.3 Prompt for Data Synthesis

The input prompt for the question generator LLM_{gen} is presented in § F. In practice, we adopt GPT-4.1 (Hurst et al., 2024) to synthesize questions.

B.4 Expert verification of synthetic data

The MedSSR synthetic dataset comprises questions generated by the question generator and pseudo-labels assigned by the policy model itself. To systematically evaluate the quality of this data against gold standards, we conducted a human evaluation with medical experts in the main content § 4.4. Here, we provide a more detailed breakdown of the evaluation results.

We randomly sampled 200 synthetic question–answer pairs (labeled by Qwen3-8B-Base) and invited eight physicians to assess their clinical validity. All participating physicians were recruited

¹<https://ftp.ncbi.nlm.nih.gov/pubmed/baseline>

Group	# Sample	# Incorrect	# Harmful	# Rating	# SD
Senior	200	1	1	4.86	0.42
1	50	1	1	4.92	-
2	50	0	0	4.78	-
3	48	0	0	4.78	-
4	52	0	0	4.95	-
Junior	200	0	0	4.74	0.54
1	50	0	0	4.98	-
2	51	0	0	4.90	-
3	50	0	0	4.78	-
4	49	0	0	4.30	-

Table 5: Physician evaluation results of synthetic data. “SD” denotes the standard deviation of the plausibility rating within the group.

Number	Incorrect	Harmful	Plausibility
200	0.50%	0.50%	4.80/5
300	1.00%	0.67%	4.88/5
500	4/500 (0.80%)	3/500 (0.60%)	4.85/5

Table 6: Expanded Physician evaluation.

through established clinical-academic collaborations at our institution. Each physician received a standard academic honorarium commensurate with their time commitment and expertise level, following our institution’s guidelines for professional consultation in research projects. The doctors were divided into two groups: four junior doctors (3-5 years of clinical experience; 3 male, 1 female) and four senior doctors (8-10 years of clinical experience; 2 male, 2 female). Each group needs to evaluate all samples, with each doctor assessing approximately 50 samples. They were asked to judge (1) whether the question and its answer were correct, (2) for any incorrect sample, whether it was harmful, and (3) the clinical plausibility of the question on a 1-5 Likert scale (1 = highly implausible, 5 = highly plausible).

The results are summarised in Table 5. Overall, the synthetic data demonstrates high factual accuracy and clinical soundness. Across all 200 questions, only one (0.50%) was marked as incorrect, and it was flagged as potentially harmful. The average plausibility score reached 4.80/5.0, indicating that the vast majority of generated questions are considered clinically realistic. Between the two groups, senior doctors rated the questions slightly higher with lower score variance. Junior doctors marked zero errors or harmful cases. These findings suggest that our knowledge-enhanced synthesis and self-labeling pipeline produces data that is scalable and trustworthy for training medical rea-

soning models.

Additionally, we expanded the physician evaluation to 500 synthetic samples. Three physicians assessed the additional 300 samples under the same criteria described above. Results are summarized in Table 6. The error rate (0.80%) and harmful rate (0.60%) remain extremely low, and the average plausibility score improved to 4.85/5.0. These results confirm the robustness of our findings and further validate the high quality of the MedSSR synthetic data.

C Experimental Details

Dataset	# All	# Train	# Test	# Rare
BioASQ	1482	700	782	144
MedQA	11451	10178	1273	154
MedMCQA	184022	182822	1200	45
PubMedQA	1000	500	500	63
HeadQA	296	296	-	-
MedXpertqa	2450	-	2450	301
MMLU-Med	1089	-	1089	57
NEJM	603	-	603	65
Lancet	412	-	412	70
Medbullets	308	-	308	52
ReDis-QA	1360	-	1360	1360
Total	204473	194496	9977	2311

Table 7: The statistics of the datasets.

C.1 Dataset Details

The detailed statistics of the datasets adopted in our work are shown in Table 7. “# Rare” represents the selected test samples related to rare disease that constitute RareDis-Sub. Detailed information on each dataset is as follows:

- **BioASQ** (Tsatsaronis et al., 2015) is a benchmark from an annual biomedical question-answering challenge². We use the official test set of 782 questions from Task B, which require binary yes/no answers. The remaining 700 samples are used in the training set.
- **MedQA** (Jin et al., 2020) consists of questions from the United States Medical Licensing Examination (USMLE), covering a wide range of complex clinical scenarios. The dataset includes both four-option and five-option versions; we use the four-option version and evaluate on the official test set of 1,273 questions.

²<https://www.bioasq.org/>

- **MedMCQA** (Pal et al., 2022) is a large-scale dataset sourced from Indian medical entrance exams, spanning over 2,400 healthcare topics. Due to the absence of labels in the official test set, we select 1,200 questions from the labeled development split as our test set.
- **MedXpertqa** (Zuo et al., 2025a) is a challenging, expert-level medical benchmark designed for complex clinical reasoning. It includes both text-based QA and multimodal VQA tasks; we use the text-only subset, which contains 2,480 test samples.
- **MMLU-Med** (Hendrycks et al., 2021) is the medical subset of the popular MMLU benchmark, covering six biomedical topics: college biology, college medicine, anatomy, clinical knowledge, human genetics, and professional medicine. The test set includes 1089 samples.
- **PubMedQA** (Jin et al., 2019) is a biomedical question-answering dataset based on PubMed abstracts, originally offering both context-provided and context-free settings. We adopt the context-free version and use the official split.
- **HEAD-QA** (Vilares and Gómez-Rodríguez, 2019) is a challenging multiple-choice QA benchmark derived from Spanish healthcare specialty exams. We use the English subset processed by prior works (Chen et al., 2025b) for training.
- **NEJM** and **Lancet** (Thapa et al., 2025) are real-world medical QA datasets derived from clinical research articles published in medical journals. Both contain only test samples, and we use the preprocessed versions provided by Thapa et al. (2025).
- **MedBullets** (Chen et al., 2025a) originates from the same source as MedQA and includes both four-option and five-option multiple-choice questions. We use the five-option version, evaluating on its official test set of 308 questions.
- **ReDis-QA** (Wang et al., 2024) is a rare disease-focused benchmark containing 1,360 questions covering 205 distinct rare conditions. Since it is constructed by extracting samples from existing datasets. ReDis-QA

Param	Qwen	Llama
Data Configuration		
Global batch size	128	128
PPO mini batch size	128	128
PPO micro batch size	8	8
Max prompt length	1024	1024
Max response length	4096	4096
Model Configuration		
Algorithm	grpo	grpo
Learning rate	2e-6	3e-7
KL loss enabled	True	False
KL coefficient	1e-3	0
Rollout Configuration		
Number of rollouts (n)	8	8
Temperature	1.0	1.0
Top-p	1.0	1.0
Top-k	-1	-1

Table 8: Detailed configuration for RL training.

includes some samples used in the training of our baselines. Thus, we perform deduplication to ensure no data leakage when testing.

To construct the RareDis-Sub benchmark, we first deduplicated the ReDis-QA dataset by removing 189 samples that overlapped with the training sets of other benchmarks, preventing potential data leakage. Following the rare disease list E_{rare} derived in §B.2, we further screened the remaining nine general medical benchmarks and identified 951 additional samples about rare diseases. Combining these, we formed the final RareDis-Sub set comprising 2,122 samples. The samples are categorized into seven categories based on the question focus: Symptoms (255), Causes (552), Affects (145), Related-disorders (152), Diagnosis (371), Treatment (539), and Others (108). This categorization enables fine-grained analysis of model performance across diverse aspects of rare disease.

C.2 Implementation Details

For knowledge retrieval, the top- k is set to 4. For the reinforcement learning training, our codes are based on the VeRL framework (Sheng et al., 2025). Since RL training on Llama is prone to collapse, we carefully tuned the hyperparameters for Llama,

Model	BioASQ	MedMCQA	MedQA	MedXpertqa	MMLU	PubMedQA	NEJM	Lancet	Medbullets	RareDis
Large Language Models										
Qwen3-8B	1.90	0.23	1.33	0.36	1.08	1.19	0.68	0.54	1.31	0.41
R1-Distill-Llama-8B	2.08	0.88	1.31	0.64	1.61	1.63	1.59	0.88	2.51	0.40
Medical Large Language Models										
UltraMedical-3.1-8B	1.13	0.97	1.15	0.70	1.31	1.15	0.35	1.64	1.06	0.64
HuatuoGPT-o1-8B	1.42	0.21	0.18	0.58	1.00	0.92	1.05	1.12	0.62	0.31
m1-7B-23K	1.96	1.52	1.51	0.76	0.93	1.30	1.37	1.82	1.53	0.32
MedPRM-8B	0.88	0.56	1.01	0.38	0.98	0.81	0.63	0.86	1.66	1.07
MedReason-8B	1.27	0.94	0.91	0.73	1.21	0.70	1.19	0.79	2.20	0.77
Qwen3-8B-Base										
+ Fully-Supervised	0.59	0.37	0.92	0.70	0.76	2.13	1.32	1.12	2.14	0.74
+ MedSSR	0.62	0.66	0.33	0.25	0.68	0.88	0.95	1.12	1.86	0.66
Llama-3.1-8B-Instruct										
+ Fully-Supervised	0.43	1.00	0.28	0.71	0.78	0.77	1.24	1.96	0.84	0.55
+ MedSSR	0.45	0.65	0.38	0.47	0.77	0.54	1.16	0.75	1.35	1.01
	0.72	0.70	0.34	0.23	0.93	0.80	0.85	0.73	1.29	0.36

Table 9: Detailed standard deviations across all models and benchmarks.

such as the learning rate and KL loss. The specific training configurations for the two models are shown in the Table 8. During offline label voting, we set the same configurations as training.

For evaluation, we test all models using the vLLM framework. We set the recommended decoding hyperparameters for each model to ensure optimal performance. To ensure answer validity, we apply logit bias in a second decoding pass to avoid invalid outputs.

Additionally, the ‘‘Avg Token/Sample’’ values shown in Figure 2 are computed using the official tokenizers of the corresponding proprietary models. For our method, we used tiktoken³ to count tokens in generated questions. For the SFT baseline, we analyzed the publicly available 23K CoT dataset released by the m1 model⁴. Token counting for DeepSeek-R1 followed its official tokenizer⁵.

D Additional Experiments

D.1 Standard Deviation of Main Results

Due to page limitations, we have not presented the standard deviations in our main results in § 4.2. Here, we provide a detailed breakdown of standard deviations across all models and benchmarks in Table 9.

Notably, our method exhibits consistently lower standard deviations compared to the fully supervised baselines and the base model on Qwen. In summary, our method can achieve higher average

³<https://github.com/openai/tiktoken>

⁴<https://huggingface.co/datasets/UCSC-VLAA/m23k-tokenized>

⁵https://cdn.deepseek.com/api-docs/deepseek_v3_tokenizer.zip

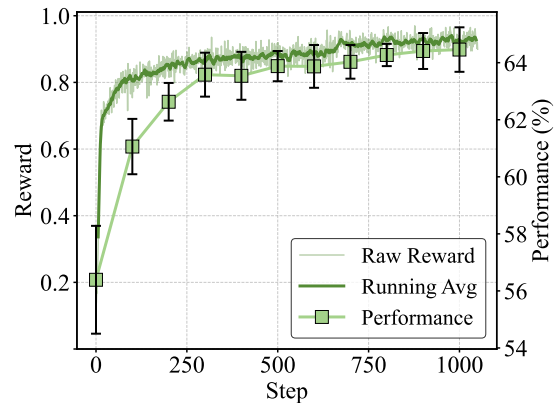


Figure 8: Rewards and performance curves of offline labeling with more than 1K steps. We report the average performance on test sets with error bars.

performance with superior training stability, making it better suited for deployment in safety-critical domains.

D.2 Offline Voting Enables Long-term Stable Training

In self-supervised reinforcement learning for LLM post-training, the online strategy of computing internal rewards typically rely on majority voting (Zuo et al., 2025b) or metrics calculated based on the distribution of output (She et al., 2025). However, these methods are often prone to reward hacking, leading to model collapse in the early stages of training. This phenomenon has been discussed in previous works (Jayalath et al., 2025; Liang et al., 2025; Zhou et al., 2025). In this work, we adopt a simple offline voting strategy to prevent this.

To further substantiate that offline voting enables sustained and stable training, we extend the

Ratio	Symptoms	Causes	Diagnosis	Others	RDs	Treatment	Affects	Avg
0%	73.73	77.76	63.07	70.37	71.22	55.84	76.38	69.77
13%	76.57	77.49	64.89	72.69	74.51	55.57	76.03	71.11
25%	79.71	80.89	68.19	72.22	76.64	57.51	78.62	73.40
33%	78.14	78.49	68.33	75.93	79.77	59.60	80.52	74.40
50%	79.12	78.62	65.84	72.45	78.78	58.63	78.45	73.13

Table 10: Performance on RareDis-Sub across different rare disease ratios. “RDs” is short for Related Disorders.

Ratio	BioASQ	MedMCQA	MedQA	MedXpertqa	MMLU	PubMedQA	NEJM	Lancet	Medbullets	Avg
0%	83.47	65.89	75.69	16.40	84.53	55.20	67.50	67.59	56.09	63.60
13%	84.78	68.73	82.54	21.27	87.21	56.20	70.56	69.78	64.12	67.24
25%	86.00	67.79	81.14	20.66	86.64	54.70	71.97	70.94	63.47	67.03
33%	83.38	68.56	81.38	20.97	86.13	53.55	71.39	67.72	64.86	66.44
50%	84.14	67.50	80.20	19.42	86.36	56.15	69.32	66.99	64.45	66.06

Table 11: Performance on general medical datasets across different rare disease ratios.

self-supervised training curve beyond 1,000 steps. As shown in Figure 8, both the raw reward and its running average remain steadily increasing and correlated with the actual model performance. There is no sign of the sharp reward increase or performance collapse in online voting methods (see § 4.4). This long-term stability demonstrates the scalability of our approach for efficiently scaling up RL training, especially for larger models.

D.3 Impact of Rare Disease Ratio

In this subsection, we provide the full results of the ablation on the rare disease ratio discussed in § 4.3. The detailed results on RareDis-Sub and general medical benchmarks are presented in Table 10 and Table 11, respectively.

As discussed in the main text, performance on both rare disease and general tasks follows a consistent trend: it initially improves with increasing rare disease ratio, peaks at an optimal point, and then gradually declines. The detailed results confirm that a ratio of 25% strikes the best balance, delivering strong gains on rare disease tasks (e.g., +2.29% over the 13% ratio natural distribution) while maintaining competitive performance on general benchmarks. This quantitative breakdown supports our analysis of the trade-off between specialization and generalization, and justifies the selection of 25% as the default ratio in our main experiments.

D.4 Comparison under Similar Cost

In the main experiments, we employed a 1:1 ratio of synthetic to real data, resulting in our

semi-supervised method using twice the data volume of the fully-supervised baseline. To ensure a fair comparison under equivalent training cost, we extend the fully-supervised baseline to the same data scale. Specifically, we add an additional 43K real samples from the rest of the collected training data (195K in Table 7), resulting in an 86K fully-supervised dataset.

The results are presented in Table 12. Even when doubling the real training data, the fully-supervised baseline achieves only a marginal average improvement (+1.12% over the 43K baseline). The additional 43K real samples contribute little to reasoning performance, likely because they are less suitable for complex reasoning tasks than the curated 43K reasoning samples used in previous experiments. In contrast, MedSSR with the same total data budget (43K synthetic + 43K real) delivers a substantially larger gain of +3.91% over the 43K fully-supervised baseline. This gap underscores that our synthetic data is not merely “more data” but higher-quality, reasoning-targeted data, effectively compensating for the scarcity of reasoning samples in existing medical corpora. In general, our method enables efficient scaling without requiring additional human annotation or costly data collection.

D.5 Generalizability across Different Scales

To evaluate the scalability and generalizability of our method, we conduct experiments on three scales of the Qwen3 model series (1.7B, 4B, 8B, and 14B). Results on the rare diseases and on the

Data Scale	Method	BioASQ	MedMCQA	MedQA	MedXpertqa	MMLU	PubMedQA	NEJM	Lancet	Medbullets	Avg
×1	Fully-supervised	81.91	64.25	74.98	17.40	83.88	54.15	68.74	67.66	55.11	63.12
	Self-supervised	82.51	65.44	75.61	17.09	86.25	55.05	66.91	68.87	55.92	63.74
×2	Fully-supervised	81.40	66.79	77.36	16.97	84.78	55.05	69.03	68.14	58.60	64.24
	MedSSR	86.00	67.79	81.14	20.66	86.64	54.70	71.97	70.94	63.47	67.03

Table 12: Performance across nine datasets under similar training cost. We define “×1 scale” as the 43K data volume used in prior experiments. The results of both “×1 scale” and “×2 scale” experiments are reported for comparison.

Model	Symptoms	Causes	Diagnosis	Others	RDs	Treatment	Affects	Avg
Qwen3-1.7B-Base	37.35	46.65	34.03	40.05	37.17	33.77	35.86	37.84
+ MedSSR	49.71	55.43	44.81	48.84	49.84	40.35	47.24	48.03
↑	12.36	8.78	10.78	8.79	12.67	6.58	11.38	10.19
Qwen3-4B-Base	60.29	66.80	48.99	58.80	53.62	45.73	63.97	56.89
+ MedSSR	70.69	75.14	58.69	68.29	70.23	53.76	72.41	67.03
↑	10.40	8.34	9.70	9.49	16.61	8.03	8.44	10.14
Qwen3-8B-Base	67.16	70.74	54.92	68.75	59.05	49.30	67.41	62.48
+ MedSSR	79.71	80.89	68.19	72.22	76.64	57.51	78.62	73.40
↑	12.55	10.15	13.27	3.47	17.59	8.21	11.21	10.92
Qwen3-14B-Base	70.67	72.05	58.69	69.83	67.89	53.87	70.03	66.15
+ MedSSR	80.39	82.11	67.52	75.93	83.55	62.29	81.38	76.17
↑	9.72	10.06	8.83	6.10	15.66	8.42	11.35	10.02

Table 13: Performance on RareDis-Sub across different LLM scales. “RDs” is short for Related Disorders.

general benchmarks are summarized in Table 13 and Table 14, respectively.

Our framework consistently delivers substantial gains across all model sizes. On rare-disease tasks, the average improvement exceeds 10% at every scale. This demonstrates that our method is particularly effective at amplifying performance in data-scarce domains, regardless of the base model’s initial capacity. On general medical benchmarks, the absolute improvement grows noticeably as the model scale increases: from +8.5% at 1.7B to +10.6% at 8B. This trend indicates that larger models may benefit more from our synthetic data and semi-supervised training, likely because they possess a stronger latent knowledge base that can be elicited through post-training. Overall, these results confirm that MedSSR is not tailored to a specific model size; rather, it provides a scalable and model-agnostic framework for enhancing medical reasoning, with particularly strong performance in data-scarce subdomains.

E Case Studies

E.1 Synthetic Examples

To showcase the generated question and the label voting process, we present case studies with

and without rare disease knowledge injection in Table 15 and Table 16, respectively. We present an example of the majority answer and that of a minority answer.

The comparison between Table 15 and Table 16 shows the impact of disease-specific knowledge injection. When using the rare disease Worster-Drought Syndrome to retrieve, we obtain the relevant clinical descriptions. This informed context steers the generation toward a targeted, clinically-grounded diagnostic question specific to Worster-Drought Syndrome. The case demonstrates how our knowledge-enhanced synthesis pipeline effectively tailors question content to underrepresented medical domains, enhancing both the relevance and factual grounding of the generated data.

E.2 Failure Mode Analysis

To examine whether systematic errors in the base model could be amplified by majority voting in our self-labeling pipeline, we analyzed an incorrect case flagged in the physician evaluation to characterize the nature of such failures.

As shown in Table 17, the base model’s majority vote (8/8 responses) selected answer B (Cer-

Model	BioASQ	MedMCQA	MedQA	MedXpertqa	MMLU	PubMedQA	NEJM	Lancet	Medbullets	Avg
Qwen3-1.7B-Base	56.10	41.36	39.20	11.58	53.81	45.35	37.94	32.16	30.20	38.63
+ MedSSR	71.99	48.17	46.86	12.54	61.85	49.05	48.84	46.66	38.48	47.16
↑	15.89	6.81	7.66	0.96	8.04	3.70	10.90	14.50	8.28	8.53
Qwen3-4B-Base	76.53	52.44	55.58	11.87	74.03	50.20	54.39	52.49	41.56	52.12
+ MedSSR	82.61	63.86	72.25	16.54	81.96	54.05	65.30	64.14	57.47	62.02
↑	6.08	11.42	16.67	4.67	7.93	3.85	10.91	11.65	15.91	9.90
Qwen3-8B-Base	76.92	58.06	62.12	13.80	79.89	49.35	60.74	60.20	46.43	56.39
+ MedSSR	86.00	67.79	81.14	20.66	86.64	54.70	71.97	70.94	63.47	67.03
↑	9.08	9.73	19.02	6.86	6.75	5.35	11.23	10.74	17.04	10.64
Qwen3-14B-Base	75.58	62.98	72.45	18.32	82.30	47.80	63.06	60.07	55.19	59.75
+ MedSSR	86.01	72.92	84.98	24.00	88.98	55.45	75.91	71.60	72.40	70.25
↑	10.43	9.94	12.53	5.68	6.68	7.65	12.85	11.53	17.21	10.50

Table 14: Performance on general medical datasets across different LLM scales.

vical radiculopathy). Examining the reasoning traces revealed that the model incorrectly output that *...Carpal tunnel syndrome typically presents with numbness and tingling in the thumb, index, and middle fingers, but usually not with thenar muscle atrophy or neck pain....* However, thenar atrophy is exactly a hallmark of advanced/severe carpal tunnel syndrome. This erroneous internal knowledge led the model to confidently exclude the correct answer and settle on a plausible but incorrect alternative.

Crucially, the model’s reasoning in this case was internally consistent. It did not exhibit random guessing or logical contradictions, but rather a coherent chain based on wrong knowledge. This type of failure represents a systematic knowledge error.

While systematic errors can propagate through majority voting, our analysis shows that they do not undermine the overall training effectiveness for two reasons. First, our physician evaluation confirms that such failures are extremely rare. Second, even when incorrect pseudo-labels are used in the self-supervised stage, the subsequent supervised training stage introduces accurate external supervision from real data, which can help correct such errors and refine the model’s knowledge boundaries.

F Prompt List

The prompt used for data synthesis **without** medical knowledge, data synthesis **with** medical knowledge, reasoning data selection, and evaluation are present in Prompt F.1, F.2, F.3, F.4, respectively.

Case Study w/ rare disease	
Seed Question 1	An army recruit, smoker and 6 months into training staid complaining of pain at postero medial aspect of both legs. There was acute point tenderness and the pain was aggravated on physical activity. The most likely diagnosis is A. Bearer's disease B. Gout C. Lumbar canal stenosis D. Stress fracture
Seed Question 2	A patient presenting with haemoptysis and renal failure with antibasement membrane antibodies has - A. Good pasture's B. Wegener's C. Churg Strauss D. Henoch-schonlein purpura
Rare Disease	Worster-Drought syndrome/Suprabulbar paresis, congenital
Retrieved Knowledge	Worster-Drought syndrome is a form of congenital suprabulbar paresis that occurs in some children with cerebral palsy. It is caused by inadequate development of the corticobulbar tracts and causes problems with the mouth and tongue including impaired swallowing. A similar syndrome in adults is called anterior opercular syndrome. A 1986 study of a family in which multiple members had Worster-Drought syndrome suggested it might be hereditary. A 2000 review of cases classified Worster-Drought Syndrome as a form of cerebral palsy, caused by early damage to the brain, but identified no obvious causes during gestation or birth and found some families with a history of the condition. The syndrome was named after Cecil Charles Worster-Drought , the doctor who described it in 1956.
Synthetic Question	A 5-year-old child presents with significant speech delay, difficulty swallowing since infancy, and drooling. Neurological examination reveals impaired movement of the tongue and soft palate, with no limb weakness. The child's mother reports similar symptoms in her own childhood. MRI shows underdevelopment of the corticobulbar tracts. Which of the following is the most likely diagnosis? A. Moebius syndrome B. Spinal muscular atrophy (bulbar form) C. Worster-Drought syndrome D. Anterior opercular syndrome (acquired suprabulbar palsy)
Majority Answer Num: 5/8	To determine the most likely diagnosis for this 5-year-old child, let's go through the information step-by-step. ### Step-by-Step Analysis: 1. Clinical Presentation : - Significant speech delay - Difficulty swallowing since infancy - Drooling - Impaired movement of the tongue and soft palate - No limb weakness. 2. Family History : - Similar symptoms in the mother's childhood. 3. Neurological Examination : - Impaired movement of the tongue and soft palate - No limb weakness. 4. Imaging Findings : - MRI shows underdevelopment of the corticobulbar tracts. ### Differential Diagnosis: 1. Moebius Syndrome (A) : - Characterized by congenital facial paralysis (including inability to move the eyes laterally and the mouth) and sometimes other cranial nerve abnormalities. - Often associated with feeding difficulties, facial asymmetry, and small eyes. - Does not typically involve limb weakness or underdevelopment of the corticobulbar tracts. 2. Spinal Muscular Atrophy (Bulbar Form) (B) : - Typically presents with muscle weakness and wasting, particularly in the bulbar muscles (speech, swallowing, and respiratory muscles). - Often has a progressive course. - Usually involves limb weakness due to lower motor neuron involvement. 3. Worster-Drought Syndrome (C) : - Also known as congenital central hypoventilation syndrome. - Characterized by underdevelopment of the brainstem, particularly the pontine respiratory centers. - May present with respiratory issues, but not typically with speech delay or swallowing difficulties unless it affects the cranial nerves. - Usually involves other neurological symptoms like poor sucking and feeding. 4. Anterior Opercular Syndrome (D) : - Also known as acquired suprabulbar palsy. - Caused by lesions affecting the corticobulbar tracts, leading to weakness or paralysis of the muscles supplied by the lower motor neurons (e.g., facial, tongue, and pharyngeal muscles). - Typically presents in adults, not children. - Associated with spasticity and hyperreflexia due to upper motor neuron involvement. ### Final Considerations: - The combination of significant speech delay, difficulty swallowing since infancy, drooling, impaired movement of the tongue and soft palate, and underdevelopment of the corticobulbar tracts suggests a congenital condition affecting the brainstem and cranial nerves. - The presence of similar symptoms in the mother's childhood suggests a familial pattern, which is characteristic of Worster-Drought syndrome or Moebius syndrome. Given the specific combination of symptoms and imaging findings, Worster-Drought syndrome (also known as congenital central hypoventilation syndrome with associated craniofacial anomalies) is the most likely diagnosis. This condition is characterized by underdevelopment of the brainstem, which would explain the underdevelopment of the corticobulbar tracts and the associated cranial nerve abnormalities. ### Final Answer: The final answer is C.
Minority Answer Num: 3/8	# Step-by-step reasoning: To determine the most likely diagnosis for the child, let's break down the information step-by-step. ### Step 1: Clinical Presentation 1. Speech delay : Significant speech delay. 2. Difficulty swallowing : Since infancy. 3. Drooling : Present. 4. Neurological examination : Impaired movement of the tongue and soft palate, no limb weakness. 5. Family history : Mother reports similar symptoms in her childhood. ### Step 2: MRI Findings - Underdevelopment of the corticobulbar tracts : This suggests involvement of the motor pathways controlling the muscles of the face, tongue, and pharynx. ### Step 3: Differential Diagnosis 1. Moebius syndrome (A) : - Characterized by facial paralysis (including inability to move the eyes laterally) and impairment of tongue movement. - Often associated with feeding difficulties and speech problems. - Typically congenital, and there may be other anomalies like clubfoot or heart defects. - No limb weakness. 2. Spinal muscular atrophy (bulbar form) (B) : - Characterized by muscle weakness and wasting, primarily affecting the bulbar muscles (those involved in speech, swallowing, and breathing). - Often presents with respiratory distress in infancy. - Progressive muscle weakness, including limbs, is a key feature. - No specific mention of congenital facial paralysis or family history in this case. 3. Worster-Drought syndrome (C) : - Also known as congenital central hypoventilation syndrome (CCHS). - Characterized by congenital abnormalities of the autonomic nervous system, leading to hypoventilation during sleep. - Often associated with Hirschsprung disease and Hirschsprung-associated intestinal ganglioneuropathy. - May have facial paralysis and tongue movement issues, but not typically speech delay or drooling alone. - MRI findings are not specific for this condition. 4. Anterior opercular syndrome (acquired suprabulbar palsy) (D) : - Characterized by contralateral upper motor neuron (UMN) signs of the face (e.g., central facial palsy), contralateral limb weakness, and contralateral sensory loss. - Typically due to stroke or other acquired causes. - Not congenital, and there is usually limb involvement. - No family history of similar symptoms. ### Step 4: Family History - The mother reporting similar symptoms suggests a genetic or congenital condition rather than an acquired one. ### Step 5: MRI Findings - Underdevelopment of the corticobulbar tracts is consistent with congenital conditions affecting these pathways, such as Moebius syndrome or other congenital disorders. ### Step 6: Final Diagnosis Given the clinical presentation of speech delay, difficulty swallowing since infancy, drooling, impaired tongue and soft palate movement, no limb weakness, and a family history of similar symptoms, the most likely diagnosis is Moebius syndrome. ### Final Answer The final answer is A.

Table 15: Case study with rare disease knowledge injection in data synthesis.

Case Study w/o rare disease	
Seed Question 1	A 35wks POG, USG parameters more or less corresponding to POG. Doppler showing absent end diastolic flow Management: A. Monitor till it becomes reverse diastolic flow or 37 weeks B. Monitor till 37 weeks C. Give steroids and wait for 48 hours D. Plan for immediate termination
Seed Question 2	A drug 'X' belongs to the anticholinergic drug group. It is primarily used in pre anesthetic medication and also during surgery. Which of the following can be 'X' ? A. Glycopyrrolate B. Pipenzolate methyl bromide C. Isopropamide D. Dicyclomine
Synthetic Question	A 32-year-old woman at 33 weeks gestation presents with reduced fetal movements. Ultrasound reveals an estimated fetal weight below the 10th percentile for gestational age. Umbilical artery Doppler shows an elevated systolic/diastolic ratio but preserved end-diastolic flow. Non-stress test is non-reactive. What is the most appropriate next step in management? A. Repeat non-stress test in 24 hours and continue monitoring B. Administer corticosteroids and plan delivery after 48 hours C. Immediate delivery D. Start magnesium sulfate and continue pregnancy until 37 weeks
Majority Answer Num: 6/8	<p># Step-by-step reasoning:</p> <ol style="list-style-type: none"> Clinical Presentation: - The patient is a 32-year-old woman at 33 weeks gestation. - She presents with reduced fetal movements, which is concerning for fetal compromise. - Ultrasound reveals an estimated fetal weight below the 10th percentile for gestational age, indicating fetal growth restriction (FGR). - Umbilical artery Doppler shows an elevated systolic/diastolic ratio but preserved end-diastolic flow, suggesting chronic placental insufficiency but not acute compromise. - The non-stress test is non-reactive, which is abnormal and indicates reduced fetal reactivity. Interpretation of Findings: - Reduced fetal movements and estimated fetal weight below the 10th percentile are strong indicators of fetal growth restriction (FGR). - Elevated systolic/diastolic ratio on umbilical artery Doppler suggests chronic placental insufficiency. - Preserved end-diastolic flow indicates that there is still some placental function, but it is compromised. - A non-reactive non-stress test is abnormal and suggests reduced fetal reactivity, which is concerning for fetal distress. Management of FGR with Abnormal Non-Stress Test: - In cases of FGR with abnormal non-stress tests (e.g., non-reactive), the next step is to assess for acute fetal compromise. - If there is evidence of acute fetal distress (e.g., abnormal biophysical profile, abnormal Doppler findings), immediate delivery may be indicated. - In the absence of acute distress, the decision to deliver depends on the gestational age, fetal condition, and maternal factors. Gestational Age and Decision Making: - At 33 weeks gestation, the fetus is considered preterm. - If there is no evidence of acute fetal distress, delivery is generally not indicated unless there is clear evidence of worsening fetal condition. - Corticosteroids may be administered to promote fetal lung maturity if delivery is likely within 7 days. Most Appropriate Next Step: - Given the reduced fetal movements, FGR, and abnormal non-stress test, the next step should be to assess for acute fetal distress. - If there is no evidence of acute distress, corticosteroids should be administered to promote fetal lung maturity, and the patient should be monitored closely for further fetal deterioration. - Delivery is typically not indicated at 33 weeks unless there is clear evidence of acute fetal distress or worsening fetal condition. Final Decision: - Administer corticosteroids to promote fetal lung maturity. - Plan delivery after 48 hours if there is no improvement in fetal condition or if there is evidence of acute fetal distress. <p># Final Answer: The final answer is B.</p>
Minority Answer Num: 2/8	<p>To determine the most appropriate next step in the management of a 32-year-old woman at 33 weeks gestation with reduced fetal movements, let's break this down step-by-step.</p> <p>### Step-by-Step Reasoning:</p> <ol style="list-style-type: none"> Clinical Presentation: - The patient is at 33 weeks gestation. - She presents with reduced fetal movements. - Ultrasound reveals an estimated fetal weight below the 10th percentile for gestational age (indicating intrauterine growth restriction, IUGR). - Umbilical artery Doppler shows an elevated systolic/diastolic ratio but preserved end-diastolic flow (suggesting chronic placental insufficiency). - Non-stress test is non-reactive (indicating poor fetal well-being). Interpretation of Findings: - Reduced fetal movements + IUGR + abnormal Doppler + non-reactive NST suggest fetal compromise. - Elevated systolic/diastolic ratio indicates chronic placental insufficiency, which is often associated with poor outcomes if not addressed promptly. Management Considerations: - Intrauterine growth restriction (IUGR) with abnormal Doppler findings and non-reactive NST is concerning for fetal compromise. - Immediate intervention may be necessary to prevent adverse outcomes such as stillbirth or neonatal morbidity. Options Analysis: - A: Repeat non-stress test in 24 hours and continue monitoring - This is not appropriate given the presence of non-reactive NST and other concerning findings. - B: Administer corticosteroids and plan delivery after 48 hours - Corticosteroids can be given to promote fetal lung maturity, but immediate delivery might be warranted based on the findings. - C: Immediate delivery - Given the combination of reduced fetal movements, IUGR, abnormal Doppler, and non-reactive NST, immediate delivery is often recommended to prevent further fetal compromise. - D: Start magnesium sulfate and continue pregnancy until 37 weeks - Magnesium sulfate is typically used for neuroprotection in preterm labor, not for managing IUGR with compromised fetal status. Final Decision: - The presence of non-reactive NST and IUGR with abnormal Doppler findings suggests significant fetal compromise. - Immediate delivery is generally recommended in such cases to prevent adverse outcomes. <p>### Final Answer: Given the clinical presentation and findings, the most appropriate next step in management is immediate delivery. Therefore, the final answer is C.</p>

Table 16: Case study without rare disease knowledge injection in data synthesis.

Failure Case	
Question	A 55-year-old woman presents with progressive weakness in her right hand and difficulty manipulating small objects over the past 3 months. She also reports mild pain radiating from her neck to her right shoulder. On examination, there is atrophy of the right thenar muscles and decreased sensation over the lateral aspect of the right hand. What is the most likely diagnosis?
Options	A. Carpal tunnel syndrome, B. Cervical radiculopathy (C6-C7), C. Ulnar nerve entrapment at the elbow, D. Thoracic outlet syndrome.
Ground Truth	A. Carpal tunnel syndrome
Voting Results	"A": 0, "B": 8, "C": 0, "D": 0
Response Example	<p>To determine the most likely diagnosis for this patient, let's break it down step-by-step:</p> <p>### Step-by-Step Reasoning:</p> <ol style="list-style-type: none"> Clinical Presentation: - Age: 55 years old - Symptoms: Progressive weakness in the right hand, difficulty manipulating small objects, mild pain radiating from the neck to the right shoulder - Physical Exam: Atrophy of the right thenar muscles, decreased sensation over the lateral aspect of the right hand Key Findings: - Atrophy of the thenar muscles: This is characteristic of median nerve involvement. - Decreased sensation over the lateral aspect of the right hand: The lateral aspect of the hand is innervated by the radial nerve, but the thenar muscles are primarily innervated by the median nerve. - Radiating pain from the neck to the shoulder: This suggests a cervical spine issue rather than a peripheral nerve problem. Differential Diagnosis: - A: Carpal Tunnel Syndrome: Typically presents with numbness and tingling in the thumb, index, middle, and half of the ring finger, but usually not with thenar muscle atrophy or neck pain. - B: Cervical Radiculopathy (C6-C7): Commonly presents with weakness in the wrist extensors (extensor carpi radialis brevis, extensor carpi radialis longus, and extensor carpi ulnaris), which can cause difficulty with small object manipulation. It also often involves pain radiating from the neck to the shoulder, and sensory changes in the thumb, index, and middle fingers (C6) or the middle and ring fingers (C7). Thenar muscle atrophy is more commonly associated with C7 radiculopathy. - C: Ulnar Nerve Entrapment at the Elbow: Presents with weakness in the hypothenar muscles, adductor pollicis, and interossei, along with decreased sensation in the little finger and half of the ring finger. There is no significant atrophy of the thenar muscles. - D: Thoracic Outlet Syndrome: Can present with symptoms similar to those of cervical radiculopathy, including pain, numbness, and weakness in the upper extremity, but typically involves both arms and may be associated with vascular signs like cyanosis or pallor of the hand. Specificity of Symptoms: - Thenar muscle atrophy is strongly suggestive of median nerve involvement, which is often due to compression at the wrist (carpal tunnel syndrome) or cervical spine issues (e.g., C6 or C7 radiculopathy). - Radiating pain from the neck to the shoulder points towards a cervical spine origin. - Lateral aspect of the hand is not typically affected by median nerve issues but could be related to radial nerve issues, which are less likely given the presence of thenar muscle atrophy. Conclusion: - The combination of thenar muscle atrophy, decreased sensation over the lateral aspect of the hand, and radiating pain from the neck to the shoulder is most consistent with C6-C7 cervical radiculopathy. - The lateral aspect of the hand is innervated by the radial nerve, which can be affected by C6 radiculopathy, but the presence of thenar muscle atrophy points more strongly to C7 involvement. <p>### Final Answer: Based on the clinical presentation and physical exam findings, the most likely diagnosis is B: Cervical radiculopathy (C6-C7).</p> <p>The final answer is B"</p>

Table 17: Failure case analysis of majority-voted labels. The incorrect parts are highlighted in red.

Prompt F.1: Prompt for Data Synthesis without Medical Knowledge

The following reasoning-heavy seed questions are provided:

Seed Question 1: {seed_question1}

Seed Question 2: {seed_question2}

Your task is to generate a medical reasoning question following these requirements:

- (a) Using the seed question as inspiration, generate a novel and clinically accurate question of comparable difficulty.
- (b) Ensure all medical content is factually correct and avoid counterfactual content (e.g., male patients with ovarian cancer)
- (c) Questions must be self-contained and require multi-step reasoning, not simple fact recall
- (d) Include realistic multiple-choice options that are plausible but only one is correct

Prompt F.2: Prompt for Data Synthesis with Medical Knowledge

The following reasoning-heavy seed questions are provided:

Seed Question 1: {seed_question1}

Seed Question 2: {seed_question2}

MEDICAL KNOWLEDGE CONTEXT:

The following medical knowledge about {rare_disease} can be used:

{medical_knowledge}

Your task is to generate a medical reasoning question following these requirements:

- (a) Using the seed question as inspiration, generate a novel and clinically accurate question of comparable difficulty.
- (b) Incorporate the **MEDICAL KNOWLEDGE** about rare disease if it is useful
- (c) Ensure all medical content is factually correct and avoid counterfactual content (e.g., male patients with ovarian cancer)
- (d) Questions must be self-contained and require multi-step reasoning, not simple fact recall
- (e) Include realistic multiple-choice options that are plausible but only one is correct

Prompt F.3: Prompt for Reasoning Data Selection

Determine if this medical question requires:

1. MEMORIZATION (can be answered by recalling a single fact)
2. REASONING (requires multi-step analysis of information)

Key characteristics of reasoning questions:

- Require comparing multiple factors
- Involve interpreting clinical scenarios
- Need synthesis of information
- Ask "most likely" or "best next step"

Respond EXACTLY with:

<Memorization> - for fact-recall questions

<Reasoning> - for analysis questions

Examples:

Memorization:

"What is the diagnostic criteria for X?"

"What is the first-line treatment for Y?"

Reasoning:

"Given these symptoms A, B and C, what is the most likely diagnosis?"

"How would you manage this patient with X and Y conditions?"

Prompt F.4: Prompt for Evaluation

You are a professional medical expert to answer the # Question. Please first think step-by-step and then answer the question (only one option can be chosen). Your responses will be used for research purposes only, so please have a definite answer and keep your response complete. Please format the final answer at the end of your response as: The answer is [LETTER]. Example: "The answer is A".

Question: {question}