CURE-MED: Curriculum-Informed Reinforcement Learning for Multilingual Medical Reasoning

Eric Onyame[1*] **Akash Ghosh**[2*] **Subhadip Baidya**[3] **Sriparna Saha**[2]
Xiuying Chen[4] **Chirag Agarwal**[1]

University of Virginia[1] Indian Institute of Technology Patna[2]
Indian Institute of Technology Kanpur[3] MBZUAI[4]

Abstract

While large language models (LLMs) have shown to perform well on monolingual mathematical and commonsense reasoning, they remain unreliable for multilingual medical reasoning applications, hindering their deployment in multilingual healthcare settings. We address this by first introducing CUREMED-BENCH, a high-quality multilingual medical reasoning dataset with open-ended reasoning queries with a single verifiable answer, spanning thirteen languages, including underrepresented languages such as Amharic, Yoruba, and Swahili. Building on this dataset, we propose CURE-MED, a curriculum-informed reinforcement learning framework that integrates code-switching-aware supervised fine-tuning and Group Relative Policy Optimization to jointly improve logical correctness and language stability. Across thirteen languages, our approach consistently outperforms strong baselines and scales effectively, achieving 85.21% language consistency and 54.35% logical correctness at 7B parameters, and 94.96% language consistency and 70.04% logical correctness at 32B parameters. These results support reliable and equitable multilingual medical reasoning in LLMs. The code and dataset are available at `cure_med`.

1 Introduction

Recent progress in large language models (LLMs) and reasoning-oriented systems has produced strong performance in mathematical reasoning and code generation (Li et al., 2022; Liu et al., 2024a; Wei et al., 2022; Huang and Chang, 2022). While these advances suggest LLMs can learn structured solution strategies beyond pattern completion, medical reasoning remains challenging (Magrabi et al., 2019; Stead, 2018; Ghosh et al., 2025a, 2026) because it requires domain knowledge, careful use of context, and reasoning that clinicians can inspect (Patel et al., 2005; Arocha et al., 2005).

Prior work shows promising medical QA and text generation tasks, yet reliable medical reasoning still depends on reasoning-centric data and evaluations that test reasoning behavior rather than answer plausibility (Liévin et al., 2024; Singhal et al., 2025; Nori et al., 2023; Ghosh et al., 2024b,c,a). Without such resources, models may generate fluent, credible-sounding outputs without dependable reasoning. The problem is amplified in multilingual settings: progress remains English-centered, leaving mid- and low-resource languages underrepresented and reliability uneven across communities(Ghosal et al., 2025; Ghosh et al., 2025a; Singh et al., 2025; Maji et al., 2025). Despite cross-lingual transfer, open-ended medical reasoning often exhibits two recurring failures: *reduced logical accuracy* and *unstable language behavior* (Cahyawijaya et al., 2024; Nguyen et al., 2023). For clinical use, these failures erode interpretability and trust, since clinicians and patients must understand not only what a system concludes, but how it arrives there (Amann et al., 2020).

While recent efforts attempt to strengthen medical capability through domain-specific supervision (Liu et al., 2024b; Shengyu et al., 2023), benchmarks primarily remain monolingual and rely on closed-form settings, providing limited visibility into multilingual reasoning quality and language fidelity (Qiu et al., 2024). As LLMs increasingly support clinical education and decision-making, systematic evaluation of multilingual reasoning and language consistency becomes essential for fairness, reliability, and generalization (Liévin et al., 2024; Cahyawijaya et al., 2024).

In this work, We study multilingual medical reasoning across 13 high-, mid-, and low-resource languages. We introduce CUREMED-BENCH, an open-ended benchmark where each query has a single verifiable answer, enabling independent eval-

*Equal contribution
†Corresponding author: reh6ed@virginia.edu

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30682–30703

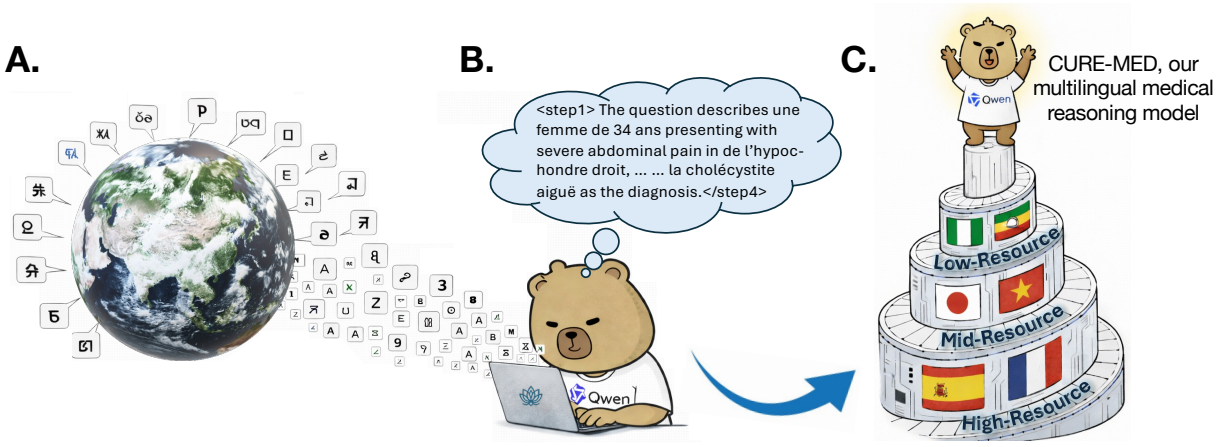


Figure 1: **The CURE-MED pipeline for multilingual medical reasoning.** The framework progresses through three stages: (A) curation of clinically validated multilingual data from sources like MedlinePlus to enable cross-lingual reasoning; (B) supervised fine-tuning of the Qwen2.5-Instruct backbone on code-switched reasoning traces; and (C) GRPO-guided curriculum reinforcement learning, progressively training from high- to mid- and low-resource languages to enhance logical correctness and language consistency.

uation of logical accuracy and language consistency and analysis of cross-lingual generalization under clinically grounded constraints.Next, we propose CUREMED, a two-stage training framework (see Figure 1) for multilingual medical reasoning. We apply code-switching-aware supervised fine-tuning (SFT) to stabilize language usage during intermediate reasoning steps and perform curriculum-informed GRPO to improve logical correctness and language fidelity. Our contributions are: **1)** We present a systematic evaluation of multilingual medical reasoning of LLMs using verifiable medical queries, enabling reliable measurement of logical accuracy and language consistency across languages; **2)** We introduce CUREMED-BENCH, a large-scale multilingual medical reasoning dataset spanning 13 languages across high-, mid-, and low-resource settings; **3)** We propose CURE-MED, a two-stage training framework for multilingual medical reasoning that combines code-switching-aware SFT with curriculum-informed reinforcement learning (RL) to jointly optimize logical correctness and linguistic fidelity; and **4)** Through extensive automatic and human evaluations, we show that CURE-MED achieve state-of-the-art performance on CUREMED-BENCH and demonstrate improved out-of-distribution generalization, including improved robustness in low-resource languages and stronger performance on unseen medical questions and languages.

2 Related Work

This work lies at the intersection of medical reasoning with LLMs and multilingual reasoning. We summarize key gaps in prior work and position CURE-MED as a unified response.

Large Medical Reasoning Models. LLMs have been widely studied for medical QA, clinical retrieval, and diagnostic tasks (Guo et al., 2022; Singhal et al., 2025; Liu et al., 2024b). Domain-specific pretraining and instruction tuning can improve factuality, yet benchmark gains often do not translate to reliable medical reasoning (Nori et al., 2023; Chen et al., 2025), with models producing fluent but clinically unsound explanations (Amann et al., 2020). A core issue is evaluation: many medical benchmarks are closed-form (e.g., multiple-choice), which hides intermediate reasoning and limits verification of logical validity (Chen et al., 2025; Qiu et al., 2024). Recent open-ended evaluations exist, but are largely monolingual or limited to a few high-resource languages, leaving multilingual medical reasoning underexplored (Qiu et al., 2024; Schmidgall et al., 2024).

We address these gaps by introducing open-ended medical queries with single verifiable answers across 13 diverse languages, enabling independent assessment of reasoning correctness.

Multilingual Reasoning and Language Fidelity. Prior work shows CoT prompting can enable cross-lingual inference transfer (Wei et al., 2022; Shi

et al., 2022; Kojima et al., 2022), but evaluations mostly target general-domain math/symbolic tasks and skew toward high-resource languages (Huang and Chang, 2022; Chen et al., 2023a; She et al., 2024; Nguyen et al., 2023; Cahyawijaya et al., 2024). In medical settings, models often exhibit degraded accuracy, language drift, and weak cross-lingual generalization (Qiu et al., 2024; Schmidgall et al., 2024). Methods such as language mixing and supervised reasoning distillation can improve fluency, but are typically studied in limited bilingual settings or overfit high-resource languages (Hämmerl et al., 2022; Yoo et al., 2024; Ge et al., 2023; Huang et al., 2024; Ye et al., 2025). RL has also been used to promote structured reasoning, but remains largely English-centric and general-domain (Ouyang et al., 2022; Achiam et al., 2023; Jaech et al., 2024; Guo et al., 2025; Luong et al., 2024).

CURE-MED *differs from prior work by optimizing language fidelity and reasoning correctness jointly. We evaluate across high-, mid-, and low-resource languages, and integrate code-switching-aware supervision with curriculum-informed RL for robust multilingual medical reasoning.*

3 Methodology

Here, we describe the construction of CUREMED-BENCH (Sec. 3.1), including dataset collection and human verification. Next, we present CURE-MED: cold-start initialization (Sec. 3.2), reward design (Sec. 3.3), and GRPO-guided curriculum reinforcement learning (Sec. 3.4).

3.1 Dataset Collection

We construct CUREMED-BENCH, a multilingual medical reasoning dataset of 15,774 open-ended QA instances across 13 languages spanning Africa, Asia, and Europe, enabling evaluation under diverse linguistic conditions (including African languages such as Hausa, Yoruba, and Swahili). A breakdown by language and language family is provided in Appendix C.

Source Material and Question Generation. CUREMED-BENCH is grounded in *MedlinePlus*, a clinically validated medical resource curated by U.S. federal health agencies. Following tool-assisted synthetic data generation (Parisi et al., 2022; Taori et al., 2023; Zhou et al., 2023; Wang et al., 2022; Schick et al., 2023; Ghosh et al., 2025b), we use GPT-4o to retrieve MedlinePlus content and draft closed-ended multiple-choice questions in each target language. Each item is anchored to the source, includes four options with exactly one correct answer, and provides clinically grounded supervision prior to conversion to open-ended prompts.

Filtering for Reasoning Difficulty. Following Chen et al. (2024), we apply multi-stage filtering to retain questions requiring substantive medical reasoning. We remove trivial items by discarding questions answered correctly by all three compact LLMs: Qwen2.5-3B/7B (Xu et al., 2025) and LLaMA-3.1-8B (Grattafiori et al., 2024). We further exclude under-specified or ambiguous questions, retaining samples with a single, unambiguous correct answer and consistent cross-lingual interpretation; GPT-4o is used to identify cases with multiple valid answers or cross-lingual inconsistency.

Conversion to Open-Ended Problems. We convert each remaining item into an open-ended prompt x using GPT-4o, and generate an explicit reasoning chain r with a free-form ground-truth answer y^*. This removes multiple-choice cues and yields open-ended instances with supervised reasoning, enabling direct evaluation of reasoning quality and answer correctness. We define the dataset as $\mathcal{D} = \{(x, r, y^*)\}$, where each instance has a single clinically grounded solution supported by an explicit reasoning trace. As summarized in Table 1, CUREMED-BENCH contains 15,774 instances across 13 languages, including low-resource languages, extending prior benchmarks that are largely multiple-choice and/or linguistically limited.

Human Verification and Ethical Review. All samples are verified by native speakers and medical experts (physicians, advanced medical students, and nursing PhD candidates). Reviewers assess clinical correctness, linguistic fidelity, and cultural appropriateness, revising culture-specific terminology and removing translation artifacts or medically inappropriate content. Across 13 languages, user studies report an average rating of **4.89/5**, supporting clinical validity (Appendix Table 5). All procedures were approved by an Institutional Review Board for social and behavioral sciences and followed established ethical research standards. Additional details are provided in Appendix D.

Dataset	Lang.	Size	Open-ended?	Reasoning Supervision	Low-resource?
MMedBench	6	8.5k	✗	✓	✗
MedQA	3	13k	✗	✗	✗
MedExpQA	4	2,488	✗	✓	✗
PubMedQA	1	211k	✗	✓	✗
MedQAUSMLE	1	11.4k	✗	✗	✗
MedMCQA	1	193k	✗	✓	✗
OphthaLingua	7	1,184	✗	✗	✓
MCMLE	1	270k	✗	✗	✗
XMedBench	4	8,280	✗	✗	✗
WorldMedQA	4	568	✗	✗	✗
HealthSearchQA	1	3,173	✓	✓	✗
CURE-MED-Bench	**13**	**15,774**	✓	✓	✓

Table 1: Comparison of medical domain benchmarks.

3.2 Cold-Start Initialization via Supervised Fine-Tuning (SFT)

We initialize multilingual reasoning with a cold-start SFT stage on *code-switched long CoT* trajectories. This stage stabilizes multi-step reasoning in the base model before we introduce stricter language-consistency constraints in later training. Given an input query x in the target language ℓ, we construct a multi-step reasoning trajectory that allows controlled code-switching in intermediate steps (see Figure 2 for a French subset example). Each trajectory contains reasoning steps $\mathbf{r} = \{r_1, \ldots, r_T\}$, where step r_t may be written in language $\ell_t \in \mathcal{L}$, followed by a final answer y^* written in the target language ℓ.

We fine-tune the model by maximizing the likelihood of the reasoning trajectory and final answer conditioned on the input: $\mathcal{L}_{\text{SFT}} = -\log p_\theta(\mathbf{r}, y^* \mid x)$, training the model to produce multi-step reasoning before generating the final response. Code-switching in $\mathbf{r}$ allows the LLM use the most effective language for intermediate inference while keeping the final answer in ℓ. The resulting language-adaptive reasoning behavior provides a strong initialization for RL stages that enforce language consistency without degrading logical accuracy.

3.3 Reward Design

We train CURE-MED with a weighted reward that promotes clinical correctness, language fidelity, and adherence to a structured output format. We use a closed-source multilingual reward model that performs competitively on RewardBench (Lambert et al., 2025). To mitigate same-model judge bias, we use a separate model for LLM-as-a-judge verification (Verga et al., 2024; Bansal et al., 2023).

Correctness Reward. Following Zheng et al. (2023), we use GPT-4.1 as a verifier to score semantic and clinical equivalence between the model output (y), and reference answer (y^*). The verifier returns a continuous score in $[0, 1]$:

$$R_{\text{acc}}(y \mid x, y^*) = v_{\text{acc}}(x, y, y^*) \in [0, 1]. \quad (1)$$

We use exact-match scoring for closed-ended questions. For open-ended questions, the verifier assigns partial credit when the response reaches the correct conclusion via clinically valid reasoning, even under paraphrase (Su et al., 2025), providing smoother learning signals.

Language Consistency Reward. We enforce strict output-language fidelity by scoring whether y is written entirely in the query language ℓ:

$$R_{\text{lang}}(y \mid \ell) = \begin{cases} 1 & \text{if the language of } y \text{ matches } \ell \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Format Reward. A parser checks compliance with the required structure (`<thinking>`, numbered `<step n>`, and `<answer>` tags):

$$R_{\text{fmt}}(y) = \begin{cases} 1 & \text{if the required format is followed} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The final composite reward is defined as:

$$R(y \mid x, y^*, \ell) = \lambda_{\text{acc}} R_{\text{acc}}(y \mid x, y^*) + \lambda_{\text{lang}} R_{\text{lang}}(y \mid \ell) + \lambda_{\text{fmt}} R_{\text{fmt}}(y) \quad (4)$$

3.4 GRPO-guided curriculum reinforcement learning

After SFT, we fine-tune the model with curriculum-guided GRPO (Shao et al., 2024; Guo et al., 2025) for optimizing the reasoning policy under the multilingual verifier-driven reward described in Sec. 3.3.

Curriculum Design. We design the curriculum around language resource availability rather than problem complexity. This is motivated by the observation that models achieve higher reasoning accuracy in high-resource languages, providing more stable reward signals early in reinforcement learning. We therefore treat languages as tasks of increasing difficulty and progress from high→medium→low-resource tiers. Based on baseline performance, we define three tiers: high- (French, Japanese, Spanish, Vietnamese), medium- (Korean, Thai, Turkish, Bengali), and low-resource (Amharic, Yoruba, Hausa, Hindi, Swahili). We start GRPO on the high-resource and progressively expand training to lower-resource tiers. To reduce

Example: Cold-Start Code-Switched Reasoning (French)

Question. Une femme de 34 ans se présente aux urgences avec des douleurs abdominales sévères localisées dans l'hypochondre droit, associées à des nausées et des vomissements. Étant donné ses antécédents de calculs biliaires, quel est le diagnostic le plus probable pour son état?

Reasoning (code-switched CoT). <thinking> <step1> The question describes une femme de 34 ans presenting with severe abdominal pain in the région de l'hypochondre droit, along with des nausées et des vomissements. These are classic upper abdominal symptoms. </step1> <step2> Ses symptômes suggèrent une atteinte possible de la vésicule biliaire, especially given her history of calculs biliaires. This could indicate a gallbladder-related condition. </step2> <step3> The presence of fièvre à 38,5°C and tenderness upon palpation of the quadrant supérieur droit indicates an inflammatory process — likely une infection. </step3> <step4> In this context, on peut fortement envisager la cholécystite aiguë (acute cholecystitis) as the most probable diagnosis given the presentation and risk factors. </step4> </thinking>

Answer. <answer> Le diagnostic le plus probable est la cholécystite aiguë </answer>

Language: French **Type:** Cold-start code-switched CoT sample

Figure 2: An example from the cold-start multilingual dataset showing CoT reasoning in French. The reasoning combines English-based clinical terms and local-language expressions, reflecting code-switching in medical contexts.

Example: Baseline vs. CURE-Med (Spanish)

Question. *Un paciente presenta congestión nasal y tos leve desde hace dos días. No tiene fiebre ni dificultad para respirar. ¿Cuál es la causa más probable?*

Baseline model (incorrect)

Reasoning (flawed). El cuadro parece un resfriado común, pero la ausencia de fiebre podría indicar que no es viral y la tos podría ser señal de algo más serio como una infección pulmonar temprana. La congestión nasal podría ser un síntoma inicial de una patología más grave.

Answer. Podría tratarse de una infección pulmonar temprana. ✗

CURE-Med (correct)

Reasoning (code-switched CoT). <step1> The symptoms are mild, lo que coincide con un resfriado leve. </step1> <step2> No fever, lo que reduce la probabilidad de neumonía. </step2> <step3> Lo más probable es un resfriado viral leve. </step3>

Answer. Lo más probable es un resfriado viral leve. ✓

Figure 3: Qualitative Spanish medical-reasoning example comparing a baseline Qwen2.5-7B-Instruct model and CURE-MED-7B. The baseline model produces fluent but clinically flawed reasoning (red) and an incorrect diagnosis, whereas CURE-MED generates a structured, code-switched CoT (blue) and arrives at the correct diagnosis (green).

catastrophic forgetting, we retain a fixed fraction of samples from the previous phase when introducing a new tier. Formally, curriculum phase C_i draws samples from languages in tier $L_i \in \{\text{high}, \text{medium}, \text{low}\}$.

Training Procedure. While following prior works (Shao et al., 2024; Guo et al., 2025; Hwang et al., 2025), we apply GRPO without modifying the optimization rule, the training was designed in curriculum phases. When reward improvements plateau within a tier, we expand sampling to include the next tier while mixing in data from the previous phase to preserve earlier capabilities. At phase i, we sample batches from: $\mathcal{D}_i = \alpha\,\mathcal{D}_{i-1} + (1-\alpha)\,\mathcal{D}_{L_i}$, where $\mathcal{D}_{L_i}$ denotes data from tier L_i, $\mathcal{D}_{i-1}$ is the retained data from phase $i-1$, and α=0.85 controls the retention ratio. This retention-aware curriculum supports incremental transfer to low-resource languages while maintaining performance.

4 Experiments

Next, we outline the experimental setup, baseline models, training and evaluation procedures used to address key research questions: **RQ1)** Does CURE-MED improve multilingual medical reasoning over instruction-tuned baselines and their vanilla variants? **RQ2)** What is the performance trade-off between language fidelity and medical reasoning accuracy? **RQ3)** How does curriculum-guided learning affect performance across model scales? **RQ4)** Does CURE-MED generalize to unseen medical questions and languages under out-of-distribution evaluation?

4.1 Experimental Setup

Dataset and Splits. All experiments are conducted on CUREMED-BENCH, where the dataset is partitioned into 80% train and 20% held-out test set. The train set is further divided into 80% for supervised fine-tuning and 20% for reinforcement fine-tuning. Dataset construction and filtering procedures are described in Sec. 3.

Baselines. We benchmark CURE-MED against 28 baseline models comprising i) general-purpose, including Qwen 2.5-Instruct (Yang et al., 2024), LLaMA (Dubey et al., 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), Apollo2 (Zheng et al., 2024), and Ministral (Team, 2024); and ii) medical-specific, including MedAlpaca (Han et al., 2025), Meditron (Chen et al., 2023b), UltraMedical (Zhang et al., 2024), HuatuoGPT (Zhang et al.), OpenBioLLM (Labs, 2024), BioMistral (Labrak et al., 2024), and MMed-LLaMA (Qiu et al., 2024). All models are evaluated in a zero-shot setting across three independent runs.

Model Training and Evaluation. We use Qwen-

Model	Consistency (↑)	Accuracy (↑)
Small Models ($\leq$ 3B)		
LLaMA-3.2-3B	23.69±0.36	10.41±0.38
Qwen2.5-Instruct-1.5B	3.84±0.25	6.20±0.24
Qwen2.5-Instruct-3B	8.39±0.42	10.83±0.60
CURE-MED-Qwen2.5-1.5B	57.60±0.65	28.32±0.35
CURE-MED-Qwen2.5-3B	**74.28**±0.60	**42.93**±0.60
Medium Models (7–9B)		
BioMistral-7B	7.10±0.90	4.80±0.95
Gemma-7B	0.37±0.25	1.23±0.80
MedAlpaca-7B	3.50±0.90	2.47±0.95
Meditron-7B	0.43±0.40	2.50±1.10
Mistral-7B	18.70±1.30	15.23±1.20
Apollo2-7B	25.63±1.35	15.93±1.35
Qwen2.5-Instruct-7B	25.44±0.36	29.56±0.42
LLaMA-3.1-Instruct-8B	36.56±0.31	18.91±0.18
HuatuoGPT-o1-8B	67.30±0.14	46.86±0.09
OpenBioLLM-Llama3-8B	1.47±0.45	36.62±0.72
MMed-Llama-3-8B	21.38±0.56	28.09±0.62
UltraMedical LLaMA-3-8B	47.03±1.03	35.29±1.10
Ministral-8B	46.93±0.45	42.87±0.21
LLaMA-3-8B	31.58±0.12	28.93±0.42
Gemma-9B	23.22±1.14	36.97±1.03
CURE-MED-Qwen2.5-7B	**85.21**±0.63	**54.35**±0.50
Large Models ($\geq$ 14B)		
MedAlpaca-13B	0.10±0.17	0.07±0.12
Qwen2.5-Instruct-14B	35.57±0.38	41.79±0.39
Qwen2.5-Instruct-32B	41.51±0.38	49.69±0.40
Qwen2.5-Instruct-72B	70.73±1.10	58.80±1.20
LLaMA-3.1-70B	75.68±1.01	54.65±0.31
LLaMA-3.3-Instruct-70B	79.66±0.32	60.80±0.72
HuatuoGPT-o1-70B	86.79±0.44	66.67±0.24
OpenBioLLM-Llama3-70B	70.30±0.43	51.22±0.41
Meditron-70B	0.21±0.55	4.54±0.59
MMed-LLaMA-3.1-70B	26.49±0.36	37.85±0.76
CURE-MED-Qwen2.5-14B	90.27±0.31	63.74±0.43
CURE-MED-Qwen2.5-32B	**94.96**±0.40	**70.04**±0.04

Table 2: Mean results across 13 languages on 28 baseline models and CURE-MED. We observe that CURE-MED models outperform models in each parameter scale. **Consistency** denotes language consistency and **Accuracy** denotes logical accuracy. Best overall results are **bold**, best baselines are underlined.

2.5-{1.5B,3B,7B,14B,32B} instruction-tuned models as backbones. Training is performed on eight NVIDIA A100 GPUs in two stages: i) SFT on the multi-step cold-switched dataset for three epochs and ii) language-resource-aware curriculum fine-tuning with GRPO. Reinforcement progresses from high- to low-resource languages, retaining 85% of data from earlier stages to mitigate catastrophic forgetting. See Appendix C.1 for additional details on our high-/low-resource language definition and the criteria used to assign languages to each group.

Following Chen et al. (2024), we evaluate on the held-out test set using an LLM-as-a-judge framework, with GPT-4o used to match each model output to the known ground-truth answer. We assess *logical accuracy* (LA), defined as the clinical accuracy of the final answer, and *language consistency* (LC), defined as whether the final answer is produced in the question's corresponding target language. Figure 3 provides a representative Spanish example, illustrating how curriculum-guided reinforcement improves accuracy while maintaining language consistency compared to a fluent but incorrect baseline. See Appendix B for Additional implementation details.

5 Results

Here, we report results that answer RQ1–RQ4 from Sec. 4. We compare CURE-MED to instruction-tuned baselines and analyze language-reasoning trade-offs, scaling under curriculum-guided reinforcement, and out-of-distribution generalization.

RQ1) CURE-MED outperforms baselines. Table 2 compares CURE-MED to three baseline families: general-purpose instruction-tuned LLMs, medical-domain instruction-tuned models, and medical-specialized LLMs. Across scales, CURE-MED improves both logical accuracy and target-language consistency. At $\leq$3B, baselines show low correctness and frequent language violations, while CURE-MED reaches 42.93% logical correctness and 74.28% consistency (3B). At 7–9B, CURE-MED improves over the best baseline in logical correctness (54.35% vs. 46.86%) while maintaining 85.21% consistency. At $\geq$14B, CURE-MED remains best, reaching 70.04% logical correctness and 94.96% consistency. Notably, our 32B model is competitive with closed-source systems and outperforms several proprietary models on CUREMED-BENCH (See Appendix E.3, E.4; Tables 8, 9, 10).

RQ2) CURE-MED achieves better language and reasoning trade-offs. Figure. 4 shows that while baselines exhibit a weak trade-off between language consistency and logical correctness, CURE-MED shifts this in the upper-right corner, highlighting that CURE-MED improves medical reasoning without sacrificing target-language fidelity, addressing a key failure mode of prior multilingual medical systems. We observe that CURE-MED-1.5B outperform several baselines ranging from 7B to 70B and our CURE-MED-32B model outperform all 28 baseline models.

RQ3) Scaling Trends of CURE-MED. Fig. 5 shows that CURE-MED smoothly scale language consistency (57.6%@1.5B $\rightarrow$ 95.0%@32B) and logical correctness (28.3%$\rightarrow$70.0%). By comparison, instruction-tuned baselines exhibit only modest gains in language consistency as scale increases, re-

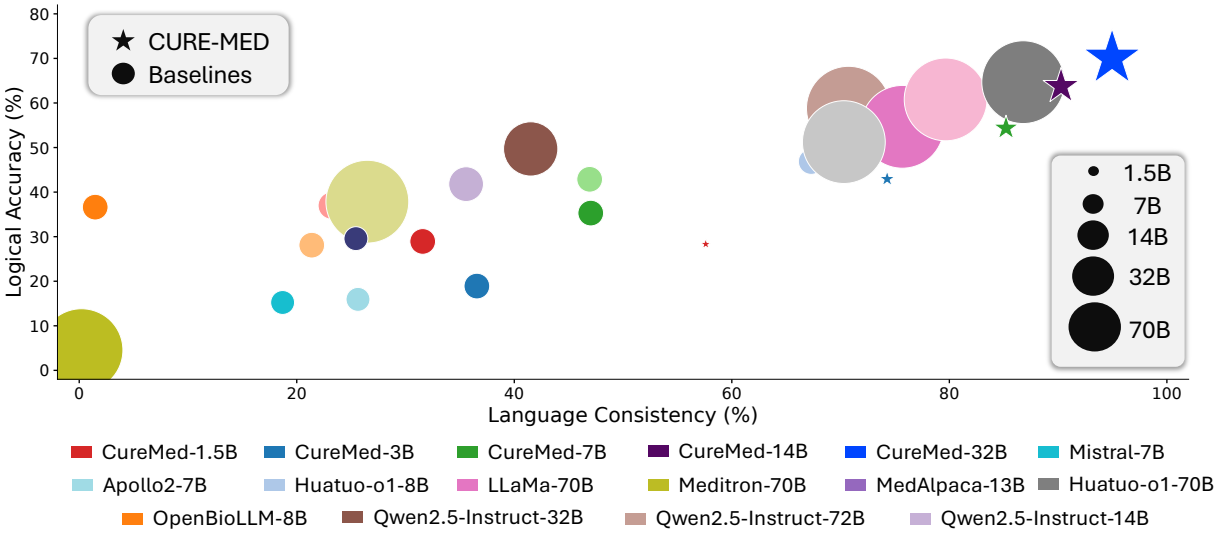


Figure 4: Trade-off performance between language consistency and logical accuracy of multilingual medical reasoning models, where each point represents a model instance with bubble size reflecting model scale. Baseline and CURE-MED models are shown as ● and ★, respectively. CURE-MED shifts performance toward the upper-right, indicating consistent gains in language consistency and logical accuracy.

maining unreliable even at larger scale. Tables 6-7 in App. E report per-language results, showing that CURE-MED consistently improves performance across languages and scales effectively. These trends indicate that curriculum-guided reinforcement fundamentally alters scaling behavior by coupling reasoning optimization with language fidelity.

RQ4) Out-of-distribution cross-lingual generalization. We evaluate transfer to held-out medical benchmarks: MMedBench (Qiu et al., 2024), MedExpQA (Alonso et al., 2024), and MedQA (Jin et al., 2021). Across all three benchmarks, CURE-MED improves accuracy over the Qwen2.5 backbones in the majority of language–scale settings, with the clearest gains for smaller models. On MMedBench (Table 4), the 1.5B backbone increases from 6.00→24.00 and from 20.00→57.50 on representative languages, demonstrating strong transfer under limited capacity. MedExpQA (Table 11) shows a similar large jump at 1.5B, rising from 1.40→44.80, while MedQA (Table 12) improves from 21.00→59.50 at 1.5B on Chinese variants. These gains remain at larger scales, indicating that curriculum-guided RL transfers beyond in-domain training to unseen questions and language variants.

6 Ablation Study

Here, we ablate CURE-MED's key components and measure their impact on logical accuracy. We also assess robustness by evaluating CURE-MED across multiple multilingual medical QA benchmarks and strong medical-domain LLM baselines.

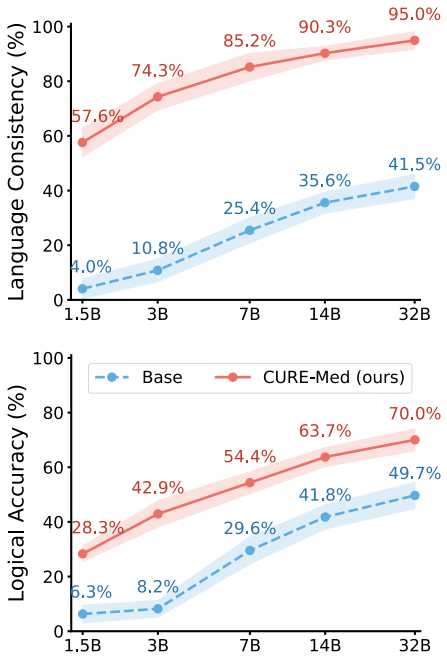


Figure 5: Scaling performance of CURE-MED vs. base across Qwen2.5-Instruct variants on language consistency **(top)** and logical accuracy **(bottom)**. Our method (solid red line) consistently outperforms the base model (dashed blue line), with performance gaps widening at larger model scales, highlighting the effectiveness of CURE-MED for multilingual medical reasoning.

Effect of Codeswitched Supervised Fine-Tuning. We isolate the effect of code-switched supervision during SFT by contrasting the base model, naïve SFT trained on multilingual long-CoT data,

Model size	Base	Naïve SFT	CURE-MED (w/o RL)	Naïve RFT	CURE-MED (w/ RL)
Qwen2.5-Instruct — Language consistency (↑)					
1.5B	3.84±0.25	8.60±1.23	53.67±0.38 (**+45.07**)	8.81±0.34	57.60±0.65 (**+48.79**)
3B	8.39±0.42	13.07±0.33	72.68±0.38 (**+59.61**)	13.28±0.57	74.28±0.60 (**+61.00**)
7B	25.44±0.36	37.11±0.44	83.46±0.36 (**+46.35**)	38.99±0.68	85.21±0.63 (**+46.22**)
14B	35.57±0.38	37.20±0.33	84.28±0.35 (**+47.08**)	39.10±1.05	90.27±0.31 (**+51.17**)
32B	35.57±0.38	43.00±0.27	90.29±0.21 (**+47.29**)	45.10±1.12	94.96±0.40 (**+49.86**)
Qwen2.5-Instruct — Logic accuracy (↑)					
1.5B	6.20±0.24	4.61±0.36	22.97±0.57 (**+18.36**)	8.80±0.47	28.32±0.35 (**+19.52**)
3B	10.83±0.60	9.50±0.38	39.13±0.53 (**+29.63**)	10.06±0.45	42.93±0.60 (**+32.87**)
7B	29.56±0.42	30.05±1.10	50.03±0.48 (**+19.98**)	38.50±0.38	54.35±0.50 (**+15.85**)
14B	41.79±0.39	43.10±0.13	61.91±0.45 (**+18.81**)	45.20±0.55	63.74±0.43 (**+18.54**)
32B	49.69±0.40	51.21±0.15	66.34±0.43 (**+15.13**)	53.40±0.49	70.04±0.04 (**+16.64**)

Table 3: Ablation study of CURE-MED. Results are averaged over three runs and reported as mean ± standard deviation and green columns denote CURE-MED variants.

Model	French	Japanese	Russian	Spanish
Qwen2.5-1.5B	6.00	11.06	20.00	20.00
↪ CURE-MED	**24.00**	**35.18**	**57.50**	**44.50**
Qwen2.5-3B	6.50	24.62	22.50	23.00
↪ CURE-MED	**42.00**	**37.69**	**60.50**	**56.00**
Qwen2.5-7B	42.00	**51.76**	53.50	63.00
↪ CURE-MED	**50.00**	46.73	**66.00**	**64.00**
Qwen2.5-14B	61.00	57.29	63.00	71.50
↪ CURE-MED	**64.00**	**65.83**	**75.50**	**78.00**
Qwen2.5-32B	69.50	67.84	72.00	29.50
↪ CURE-MED	**78.50**	**77.29**	**80.00**	**82.50**

Table 4: OOD accuracy on MMedBench. CURE-MED improves reasoning performance across all model sizes, showing strong cross-lingual generalization to unseen medical questions and languages. See Tables 11-12 for results on MedExpQA and MedQA datasets.

and CURE-MED SFT without reinforcement learning. Naïve SFT yields small and sometimes unstable improvements: language consistency rises from 8.39%→13.07% at 3B, yet logic accuracy decreases from 10.83%→9.50%, indicating that multilingual instruction tuning does not consistently strengthen medical reasoning as shown in Table 3. In contrast, code-switched SFT in CURE-MED produces large, consistent gains across model scales. At 1.5B, language consistency increases from 3.84%→53.67% and logic accuracy from 6.20%→22.97%. These improvements persist as scale increases, reaching 90.29% language consistency and 66.34% logic accuracy at 32B. In summary, the results show that structured code-switching during SFT drives the strongest gains, while naïve multilingual SFT remains insufficient for reliable multilingual medical reasoning.

Effect of GRPO-guided curriculum reinforcement learning. We assess whether RL adds value beyond SFT by comparing naïve single stage GRPO based RFT against the curriculum and language resource-aware RL used in CURE-MED, with results summarized in Table 3. Naïve RFT yields limited and uneven gains, especially at smaller scales, suggesting that uniform reinforcement signals do not consistently shape multilingual behavior. In contrast, CURE-MED applies RL after code switched SFT and delivers reliable improvements in both language consistency and logical accuracy across all model sizes. These results show that curriculum and resource-aware RL stabilizes optimization and strengthens multilingual medical reasoning beyond naïve GRPO.

CURE-MED vs. Medical LLM baselines across Benchmarks. We evaluate CURE-MED against strong medical-domain LLM baselines across four multilingual medical benchmarks (see Fig. 6). CURE-MED remains consistent, with CURE-MED-32B achieving the best performance on CUREMED-BENCH (70.04%) and MMed-Bench (79.57%), and remains competitive on MedQA and MedExpQA, where HuatuoGPT-70B leads narrowly. CURE-MED-14B also provides strong results across all benchmarks, while other medical baselines lag behind more substantially, highlighting CURE-MED's robustness across diverse evaluation settings.

7 Conclusion

We introduce CUREMED-BENCH, a multilingual medical reasoning benchmark of open-ended questions with explicit reasoning traces and a single verifiable answer across 13 languages, including low-resource settings. Using CUREMED-BENCH, we propose CURE-MED, which combines cold-

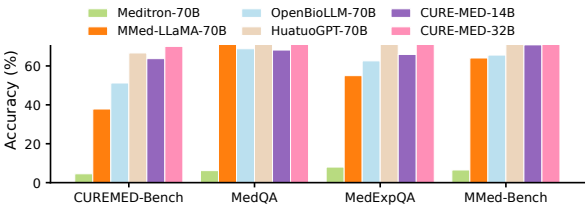


Figure 6: CURE-MED vs. medical LLM baselines across four multilingual medical QA benchmarks. Results show logi cal accuracy, highlighting CURE-MED's consistent across diverse evaluation settings.

start code-switched initialization, structured supervised fine-tuning, and language-resource-aware curriculum-RL to improve reasoning while preserving target-language fidelity. Across languages, datasets, and model scales, CURE-MED improves logical correctness and language consistency over strong baselines; ablations show supervised and RL stages provide complementary gains for stable multilingual reasoning.

8 Limitations

CUREMED-BENCH is constrained by the availability of clinically reliable source material across languages, which limits coverage and can create uneven difficulty between high- and low-resource settings. Our benchmark targets open-ended questions with a single verifiable answer and thus does not capture longitudinal care trajectories, multi-visit decision-making, or multimodal clinical evidence. In addition, parts of our pipeline rely on API-based models (e.g., for generation and/or verification), which can be costly and may hinder reproducibility for some researchers; a practical direction is to replace these components with smaller open-source models trained for the same roles and to release prompts, code, and verifier alternatives to reduce dependence on paid APIs. Future work will expand language coverage, broaden clinical settings and modalities, and further reduce reliance on proprietary APIs.

9 Ethical Considerations

This work supports the evaluation and training of multilingual medical reasoning systems by measuring reasoning correctness and target-language fidelity across diverse languages. CUREMED-BENCH is derived from publicly available, clinically curated sources and contains no patient records or personally identifiable information. Native speakers and medical experts reviewed all samples for clinical correctness, linguistic fidelity, and cultural appropriateness under IRB-approved procedures, and we report per-language results to surface reliability differences across resource levels.

10 Acknowledgment

We would like to thank all the anonymous reviewers of ACL for their valuable feedback. C.A. is supported, in part, by grants from Capital One, LaCross Institute for Ethical AI in Business, the UVA Environmental Institute, OpenAI Researcher Program, Thinking Machine's Tinker Research Grant, and Cohere. The views expressed are those of the authors and do not reflect the official policy or the position of the funding agencies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):310.

Jose F Arocha, Dongwen Wang, and Vimla L Patel. 2005. Identifying reasoning strategies in medical decision making: a methodological guide. *Journal of biomedical informatics*, 38(2):154–171.

Hritik Bansal, John Dang, and Aditya Grover. 2023. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. 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, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2023a. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yubin Ge, Devamanyu Hazarika, Yang Liu, and Mahdi Namazifar. 2023. Supervised fine-tuning of large language models on human demonstrations through the lens of memorization.

Soumya Suvra Ghosal, Vaibhav Singh, Akash Ghosh, Soumyabrata Pal, Subhadip Baidya, Sriparna Saha, and Dinesh Manocha. 2025. Relic: Enhancing reward model generalization for low-resource indic languages with few-shot examples. *arXiv preprint arXiv:2506.16502*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024b. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024c. Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560.

Akash Ghosh, Tajamul Ashraf, Rishu Kumar Singh, Numan Saeed, Sriparna Saha, Xiuying Chen, and Salman Khan. 2026. Carepilot: A multi-agent framework for long-horizon computer task automation in healthcare. *arXiv preprint arXiv:2603.24157*.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025a. The multilingual mind: A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*.

Akash Ghosh, Srivarshinee Sridhar, Raghav Kaushik Ravi, Muhsin Muhsin, Sriparna Saha, and Chirag Agarwal. 2025b. Clinic: Evaluating multilingual trustworthiness in language models for healthcare. *arXiv preprint arXiv:2512.11437*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Quan Guo, Shuai Cao, and Zhang Yi. 2022. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11):8548–8564.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2025. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*.

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *arXiv preprint arXiv:2507.05418*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Saama AI Labs. 2024. Openbiollm: Llama3-based biomedical large language model. `https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B`. Model card. Paper in preparation.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, and 1 others. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024b. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.

Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, and 1 others. 2019. Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications. *Yearbook of medical informatics*, 28(01):128–134.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, and 1 others. 2025. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*.

Vimla L Patel, José F Arocha, and Jiajie Zhang. 2005. Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning*, 14:727–750.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*.

Zhang Shengyu, Dong Linfeng, Li Xiaoya, Zhang Sen, Sun Xiaofei, Wang Shuhe, Li Jiwei, Runyi Hu, Zhang Tianwei, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Punit Kumar Singh, Nishant Kumar, Akash Ghosh, Kunal Pasad, Khushi Soni, Manisha Jaishwal, Sriparna Saha, Syukron Abu Ishaq Alfarozi, Asres Temam Abagissa, Kitsuchart Pasupa, and 1 others. 2025. Let's play across cultures: A large multilingual, multicultural benchmark for assessing language models' understanding of sports. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15205–15252.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

William W Stead. 2018. Clinical implications and challenges of artificial intelligence and deep learning. *Jama*, 320(11):1107–1108.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Mistral AI Team. 2024. Un ministral, des ministraux. Accessed: 2025-12-24.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, and 1 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2024. Code-switching curriculum learning for multilingual transfer in llms. *arXiv preprint arXiv:2411.02460*.

H Zhang, J Chen, F Jiang, F Yu, Z Chen, J Li, G Chen, X Wu, Z Zhang, Q Xiao, and 1 others. Huatuogpt, towards taming language model to be a doctor. arxiv (2023). *arXiv preprint arXiv:2305.15075*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2024. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. *Preprint*, arXiv:2410.10626.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and

chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Part I
Appendix

Table of Contents

A LLM-as-a-Judge Verification Protocol

Inspired by (Chen et al., 2024), We employ an LLM-as-a-judge framework to automatically evaluate the correctness of model-generated responses. In this setup, GPT-4o acts as a verifier that compares a model's prediction against a reference answer and determines whether the response is logically correct and linguistically valid. The verifier outputs a binary decision, returning `True` when the response aligns with the reference and `False` otherwise. Fig. 7 shows the exact prompt used for verification.

Prompt for the LLM-as-a-Judge Evaluator

```
<Model Response>
{Model Response}
</Model Response>

<Reference Answer>
{Ground-truth Answer}
</Reference Answer>
```

You are given a model-generated response and a reference answer. Determine whether the model response is correct with respect to the reference. Output `"True"` if the response is correct and `"False"` otherwise.

Figure 7: Prompt used for LLM-as-a-judge verification.

B Training and Verification Protocols

This section documents the prompts, reward verification procedures, and training hyperparameters used for supervised and reinforcement fine-tuning. Together, these components define the optimization signals and structured supervision underlying the proposed framework.

B.1 Reward Verification and Weighting

We design a composite reward that jointly enforces clinical correctness, language fidelity, and output format compliance. The final reward is defined as

$$R = 0.65 \times R_{\text{accuracy}} + 0.30 \times R_{\text{language}} + 0.05 \times R_{\text{format}}.$$

This weighting prioritizes medical correctness while explicitly penalizing language drift and format violations.

B.2 Verifier Models and Prompts

Both correctness and language rewards are scored using **gpt-4.1** with `temperature=0.0` and `max_tokens=10`. For each prompt, we generate 16 candidate responses to estimate stable reward signals.

B.3 Accuracy Verifier.

You are an expert multilingual medical evaluator. Score the generated response for correctness and medical validity on a continuous scale from 0.0 to 1.0. Give 1.0 if the reasoning is clinically sound and semantically correct, even if phrased differently from the reference. Focus on factual and clinical accuracy rather than wording.
Question: {question}
Ground truth answer: {ground_truth}
Generated response: {generated}
Output only a float between 0.0 and 1.0.

B.3.1 Language Consistency Verifier Prompt

You are an expert multilingual medical evaluator. Determine whether the model response is written entirely in the same language as the question.
Question language: {language}
Generated response: {generated}
Output 1.0 if the language matches exactly; otherwise output 0.0.

B.3.2 Format Reward

We apply a deterministic rule-based check requiring exactly one `<thinking>` block and one `<answer>` block, implemented using regular expressions with `re.DOTALL`. This constraint ensures consistent structure during reinforcement learning.

B.4 Training Hyperparameters

Compute and reproducibility configuration. All experiments were run on a single node equipped with $8\times$ NVIDIA A100 GPUs (80GB each). We used distributed training via DeepSpeed ZeRO-3 across 8 GPUs. The software stack was PyTorch 2.1, Transformers 4.40, DeepSpeed 0.14, and CUDA 12.1. For reproducibility, we report results over 3 independent runs and fix the random seed to 42.

Compute budget. Wall-clock training time was approximately 6–12 hours for SFT and 10–20 hours for GRPO-based reinforcement fine-tuning, corresponding to roughly 130–250 total GPU-hours.

B.4.1 Supervised Fine-Tuning (SFT).

- Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- Learning rate: 1×10^{-5} (cosine scheduler, warmup ratio 0.1)
- Epochs: 3
- Effective batch size: 32
- Max sequence length: 4096
- Precision: bf16
- Distributed optimization: DeepSpeed ZeRO-3 (single-node, 8 GPUs) with gradient checkpointing

B.5 SFT and RFT Prompts.

As shown in Figure 8, we adopt the instruction and formatting template from Hwang et al. (2025) with minor modifications for our multilingual medical reasoning setting. We use the same prompt structure for both supervised fine-tuning (SFT) and reinforcement fine-tuning (RFT); the only difference lies in the training objective, not in the prompt text.

System message 1. You are an expert multilingual medical doctor. When answering a medical question, follow these steps:

1. First, search your internal knowledge base thoroughly for relevant background information about the topic.
2. Understand and reason the question fully in English first.
3. Reason mainly in English, but code-switch naturally into the target language whenever useful for clarity or domain accuracy.
4. Consider multiple perspectives and potential answers before settling on your final response.
5. Evaluate the confidence in your answer based on the information available to you.
6. Provide the final answer clearly in the target language, making sure it's well-supported by your reasoning.
7. If there are significant uncertainties or gaps in your knowledge, acknowledge them transparently.

Your goal is to provide accurate, well-reasoned responses that demonstrate depth of understanding, not just surface-level answers.

System message 2. You are an expert multilingual medical doctor. When answering a medical question, think and reason mainly in English with natural code-switching to the target language. Use multi-step reasoning wrapped in `<step>` tags inside `<thinking>`.

User message. The question is in {language}. {question} Please think carefully with English-guided reasoning and code-switching, return your reasoning inside `<thinking>` `</thinking>` tags, and the final answer inside `<answer>` `</answer>` tags. Final answer ONLY in {language}.

Figure 8: SFT and RFT prompt template used for multilingual medical reasoning.

B.5.1 Reinforcement Fine-Tuning (RFT).

- Algorithm: GRPO
- Learning rate: 1×10^{-6} (cosine scheduler, warmup ratio 0.1)
- Weight decay: 0.1
- Effective batch size: 16
- Generations per prompt: 16
- Max training steps: 500
- Max prompt / completion length: 1024 / 1024
- Distributed optimization: DeepSpeed ZeRO-3 (single-node, 8 GPUs)

C Dataset Details

This appendix characterizes the linguistic composition of CUREMED-BENCH. Figure 9 shows the per-language instance distribution, with French contributing the largest share (13.5%) and Bengali the smallest (2.9%), and most languages occupying a mid-range band of roughly 7–10% of the data. The figure also groups the 13 languages into eight language families, spanning Afroasiatic and Niger–Congo as well as Indo–European, Turkic, Austroasiatic, Tai–Kadai, Japonic, and Koreanic. Together, these statistics highlight both the dataset's uneven language coverage and its broad typological diversity.

C.1 Language-based Curriculum Tiers

We construct our curriculum by defining difficulty along the linguistic axis rather than by question complexity. To operationalize this design, we use Qwen2.5-14B-Instruct as a reference model and estimate baseline reasoning accuracy separately for each language. The model performs best on high-resource languages and degrades as linguistic resources and model familiarity decrease, so we treat high-resource languages as easier tasks and progressively introduce more challenging languages during training. This curriculum aims to transfer reasoning competence learned in high-resource settings to underrepresented languages while maintaining language fidelity.

Based on the baseline accuracy ranking, we partition languages into three tiers. The high-resource tier includes French, Japanese, Spanish, and Vietnamese. The medium-resource tier includes Korean, Thai, Turkish, and Bengali. The low-resource tier includes Amharic, Yoruba, Hausa, Hindi, and Swahili. This tiering reflects the reference model's initial proficiency distribution and provides a structured progression from easier to harder multilingual reasoning conditions.

D Data Curation

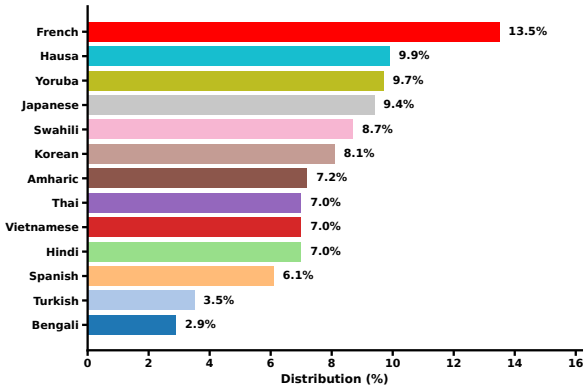
The following prompt was used to generate the initial pool of medically grounded multiple-choice questions across 13 languages. Inspired by the approach of Hwang et al. (2025) and Zhang et al., we adapted their template and instructed GPT-4o to query MedlinePlus directly and independently construct questions in each target language rather than translating from a shared source. This ensures linguistic naturalness, cultural appropriateness, and strong domain grounding across all languages.

D.1 Why an MCQ-to-Open-Ended Pipeline?

Following prior medical LLM data curation workflows, we adopt a two-step MCQ-to-open-ended pipeline: (i) generate and filter source-grounded multiple-choice questions (MCQs), then (ii) convert the retained items into open-ended queries (Hwang et al., 2025; Zhang et al.). We prefer this over direct open-ended generation for three reasons. First, MCQs are easier to verify at scale because the correct answer is discrete, enabling efficient automated checks and filtering. Second, MCQ structure provides an explicit anchor that helps reduce ambiguity and the risk of ungrounded content (hallucinations) during generation. Third, this approach improves practicality in multilingual settings by strengthening quality control before producing open-ended questions that better reflect real use while remaining strictly evaluable (Hwang et al., 2025; Zhang et al.).

D.2 Construction of the SFT Dataset

For each (*question*, *target language*, *gold answer*) triplet in our curated set, we use a strong teacher model (GPT-4o) to produce supervised fine-tuning targets. Specifically, the teacher is prompted to generate: (i) a stepwise reasoning trace that may mix English clinical terminology with the target language for precision and readability, and (ii) a final answer that must be written entirely in the target language. To standardize training targets across languages, the teacher output is constrained to a fixed schema consisting of a `<thinking>` block containing explicit `<step>` tags, followed by an `<answer>` block containing only the final answer.



Language Family	Languages
Afroasiatic	Amharic (Am), Hausa (Ha)
Niger–Congo	Swahili (Sw), Yoruba (Yo)
Indo-European	Bengali (Bn), Hindi (Hi), French (Fr), Spanish (Es)
Turkic	Turkish (Tr)
Austroasiatic	Vietnamese (Vi)
Tai–Kadai	Thai (Th)
Japonic	Japanese (Ja)
Koreanic	Korean (Ko)

Figure 9: **Language and family composition of CUREMED-BENCH. Left:** Number of dataset instances per language across the 13 languages. **Right:** Assignment of languages to eight language families with standard abbreviations.

After generation, we apply human verification to ensure (1) the final answer is clinically consistent with the gold reference answer, and (2) the output satisfies formatting and language constraints, most importantly, that the content inside `<answer>` is strictly in the target language. Samples that pass verification are retained as SFT training instances. The resulting SFT dataset is then split according to the protocol described in Section 4.1 (§4.1).

D.3 Human Verification Protocol and Rater Instructions

This section documents the human verification procedures used to validate the quality of our synthetic data. Participation in the study was completely voluntary, as participants received no payment for participating in the study. We provide the exact instructions used by medical professionals who assessed the clinical correctness of question-answer pairs and by native speakers who evaluated the language's correctness and fidelity in the target language, as shown in Figures 11 and 12. These materials specify the task setup, scoring rubric, and optional comment guidelines used throughout our verification pipeline.

D.4 Human Verification Scores by Language

We report per-language human verification scores from two rater groups. Medical professionals score clinical correctness of each question–answer pair, while native speakers score target-language quality and fidelity. Table 5 summarizes both scores on a 1–5 scale, where higher values indicate better quality. Scores are generally high across languages, reflecting the rigorous curation process applied during dataset construction. Notably, lower-resource languages such as Amharic, Thai, and Turkish receive slightly lower scores, suggesting that these languages present greater challenges for both clinical accuracy and fluent translation. Overall, the consistently high ratings across both dimensions confirm that the dataset meets the quality bar required for reliable multilingual medical reasoning evaluation.

Language	Medical correctness	Language quality
Amharic	4.45	4.45
Bengali	4.92	4.96
French	5.00	5.00
Hausa	4.96	5.00
Hindi	5.00	4.92
Japanese	4.96	4.96
Korean	5.00	5.00
Spanish	5.00	5.00
Swahili	5.00	4.96
Thai	4.70	4.70
Turkish	4.60	4.60
Vietnamese	4.95	4.95
Yoruba	5.00	5.00

Table 5: Per-language human verification scores (1–5) from medical professionals (clinical correctness) and native speakers (language quality). Higher is better.

E Per-Language Model Performance

This section provides a fine-grained analysis of multilingual medical reasoning performance broken down by language. We compare CURE-MED with instruction-tuned baselines across all 13 languages in CUREMED-BENCH, enabling a detailed examination of logical correctness and language consistency under diverse linguistic and resource conditions. This per-language view complements aggregate results by revealing where gains are most pronounced and where challenges remain.

Prompt for Generating Multilingual Medical Multiple-Choice Questions

Task: You are an expert medical content generator. Generate {num_questions} high-quality, medically accurate multiple-choice questions (MCQs) based strictly on content from MedlinePlus by searching and curating from the website.

You must independently compose each question in **ALL** of the following languages: Amharic, Bengali, French, Hausa, Hindi, Japanese, Korean, Spanish, Swahili, Thai, Turkish, Vietnamese, Yoruba.

Requirements:

1. **Medical Grounding:** All information must be sourced from MedlinePlus, covering symptoms, causes, risk factors, diagnostics, treatments, or prevention strategies.
2. **Independent Composition:** Each language version must be originally written (not translated) using natural phrasing and medically appropriate terminology for that language.
3. **Clinical Reasoning Depth:** Questions must require genuine clinical reasoning beyond trivial fact recall. Each question should have exactly one unambiguous correct answer.
4. **Format:** 4-option MCQ (A/B/C/D) with one correct answer.

Output Format: Return valid JSON array:

```
[
\{"question_id": "<id>", "source_concept": "<MedlinePlus_topic>",
"mcq_items": [\{"language_code": "<lang>", "question": "<text>",
"option_A": "<text>", "option_B": "<text>", "option_C": "<text>", "option_D": "<text>",
"correct_answer": "<A|B|C|D>"\}, ...]\}
]
```

IMPORTANT: Return ONLY valid JSON without explanations, formatting, or additional text. Ensure all special characters are properly escaped.

Figure 10: Prompt for Stage 1 multilingual MCQ generation. Here, `{num_questions}` specifies the number of questions to generate, and GPT-4o queries MedlinePlus directly to construct clinically grounded questions independently in each of the 13 target languages.

Participant Instructions: Verification Task

Task Overview
You will review synthetically generated medical question–answer pairs based on public sources such as MedlinePlus. These pairs are generated synthetically and do not involve real patient data. Your role is to assess medical correctness and accuracy.

What You Will Do
For each question–answer pair:

- Read the question and the provided answer.
- Check for medical correctness: ensure the information is accurate, logically sound, and aligned with standard medical knowledge.
- Assign a score from 1 to 5:
  - **1**: Completely inaccurate or misleading.
  - **2**: Mostly inaccurate with major errors.
  - **3**: Partially accurate but with notable issues.
  - **4**: Mostly accurate with minor issues.
  - **5**: Fully accurate and reliable.
- (Optional) Provide a brief comment if necessary (e.g., explain errors, suggest corrections, or note cultural/language specifics). Comments are optional but helpful.

You will receive batches of 50–100 pairs via an online survey. The task takes approximately 1–2 hours and can be completed remotely at your convenience. You may skip any pair or stop at any time.

Figure 11: Instructions provided to medical professional annotators for verifying clinical correctness of synthetic question–answer pairs.

E.1 Per-Language Results for Qwen2.5-7B

Table 6 reports per-language performance for the Qwen2.5-7B-Instruct baseline and its CURE-MED variant. Across all 13 languages, CURE-MED substantially improves both logical accuracy and language consistency. Gains are especially large in low-resource languages such as Amharic, Hausa, Swahili, and Yoruba, where the baseline frequently

Participant Instructions: Language Verification Task

Task Overview
You will review synthetically generated medical question–answer pairs written in one of the following target languages: Amharic, Bengali, French, Hausa, Hindi, Japanese, Korean, Spanish, Swahili, Thai, Turkish, Vietnamese, and Yoruba. These pairs are generated synthetically and do not include real patient data. Your role is to verify whether the question and answer are written correctly and naturally in the target language.

What You Will Do
For each question–answer pair:

- Read the question and the provided answer.
- Verify language correctness and fidelity:
  - The text is in the requested target language (no switching to another language).
  - The wording is grammatical and understandable for a native speaker.
  - The phrasing is natural and appropriate for medical communication.
  - Medical terms are expressed in an acceptable way for the target language (including common loanwords, when appropriate).
- Assign a score from 1 to 5:
  - **1**: Not in the target language or largely unintelligible.
  - **2**: Major language errors; difficult to understand.
  - **3**: Understandable but with noticeable errors or unnatural phrasing.
  - **4**: Mostly correct and natural with minor issues.
  - **5**: Fully correct, natural, and clearly in the target language.
- (Optional) Provide a brief comment to note issues (e.g., incorrect language, grammar problems, unnatural phrasing, or better word choices).

You will receive batches of 50–100 pairs via an online survey. The task takes approximately 1–2 hours and can be completed remotely at your convenience. You may skip any pair or stop at any time.

Figure 12: Instructions provided to native-speaker annotators for verifying language correctness and target-language fidelity of synthetic question–answer pairs.

fails to produce correct or language-faithful responses. In higher-resource languages such as French, Japanese, and Spanish, CURE-MED yields more moderate but consistent improvements, indicating that GRPO-guided curriculum RL enhances reasoning robustness without degrading performance in well-resourced settings. Overall, these results show that CURE-MED improves multilingual medical reasoning uniformly while significantly narrowing performance disparities across languages.

E.2 Per-Language Results for Qwen2.5-3B

Table 7 shows that CURE-MED consistently improves the 3B model across all evaluated languages in both logical correctness and language accuracy. The baseline 3B model exhibits extremely low performance for several languages, including Amharic, Hausa, Swahili, and Turkish, whereas the CURE-MED variant achieves large absolute gains, often exceeding 40–80 percentage points. Even in languages where the base model is already relatively stronger, such as French, Japanese, Spanish, and Vietnamese, CURE-MED delivers clear and reliable improvements. These results demonstrate that curriculum-guided reinforcement is particularly effective for small models, enabling robust multilingual medical reasoning despite limited model capacity.

E.3 Proprietary Model Performance on CUREMED-BENCH

Table 8 summarizes inference-only performance of frontier models on CUREMED-BENCH, reporting language consistency and logical accuracy averaged over 13 languages. While some models maintain strong target-language adherence (*e.g.,* Claude 3 Haiku), results reveal substantial brittleness: GPT-5-nano exhibits notably weaker language consistency, and the Gemini 2.5 family degrades sharply in both language control and reasoning quality (with Gemini 2.5 Pro nearly collapsing). These averages also conceal larger failures in low-resource languages, where models more frequently drift from the target language and show steeper drops in logical accuracy (see Appendix E.4). Overall, CUREMED-BENCH exposes a reliability gap for proprietary LLMs: strong performance in some

Language	Logic (Base)	Logic (CURE-MED)	Δ	Lang. (Base)	Lang. (CURE-MED)	Δ
Amharic	0.95	17.14	**+16.19**	0.00	64.76	**+64.76**
Bengali	10.00	60.00	**+50.00**	2.14	91.43	**+89.29**
French	67.86	77.86	**+10.00**	71.43	96.43	**+25.00**
Hausa	5.06	43.04	**+37.98**	0.00	77.22	**+77.22**
Hindi	4.48	48.51	**+44.03**	5.97	90.30	**+84.33**
Japanese	68.57	77.14	**+8.57**	60.00	94.29	**+34.29**
Korean	41.33	52.00	**+10.67**	26.67	84.00	**+57.33**
Spanish	62.86	72.38	**+9.52**	60.95	96.19	**+35.24**
Swahili	0.00	35.71	**+35.71**	0.00	67.14	**+67.14**
Thai	51.02	59.18	**+8.16**	37.76	86.73	**+48.97**
Turkish	12.50	43.75	**+31.25**	3.57	75.89	**+72.32**
Vietnamese	66.67	70.48	**+3.81**	61.90	94.29	**+32.39**
Yoruba	0.00	40.86	**+40.86**	0.00	77.42	**+77.42**

Table 6: Per-language performance of Qwen2.5-7B-Instruct (Base) and the CURE-MED 7B variant on CUREMED-BENCH. We report logical correctness and language accuracy, along with absolute gains Δ (CURE-MED −Base).

Language	Logic (Base)	Logic (CURE-MED)	Δ	Lang. (Base)	Lang. (CURE-MED)	Δ
Amharic	0.95	14.29	**+13.34**	0.00	40.95	**+40.95**
Bengali	2.86	55.71	**+52.85**	0.00	85.00	**+85.00**
French	12.14	70.71	**+58.57**	22.14	95.71	**+73.57**
Hausa	2.53	27.85	**+25.32**	0.00	64.56	**+64.56**
Hindi	5.97	28.36	**+22.39**	0.00	83.58	**+83.58**
Japanese	23.81	62.86	**+39.05**	26.67	89.52	**+62.85**
Korean	8.00	36.00	**+28.00**	2.67	76.00	**+73.33**
Spanish	17.14	62.86	**+45.72**	23.81	94.29	**+70.48**
Swahili	0.00	17.86	**+17.86**	0.00	51.43	**+51.43**
Thai	10.20	58.16	**+47.96**	0.00	73.47	**+73.47**
Turkish	1.79	28.57	**+26.78**	0.00	53.57	**+53.57**
Vietnamese	44.76	69.52	**+24.76**	79.05	80.00	**+0.95**
Yoruba	6.45	17.20	**+10.75**	0.00	69.89	**+69.89**

Table 7: Per-language performance of the 3B Base model and its CURE-MED variant on CUREMED-BENCH. We report logical correctness and language accuracy, along with absolute gains Δ (CURE-MED −Base).

Model	Lang. Consistency (↑)	Logical Acc. (↑)
GPT-5-nano	69.11	73.24
GPT-5-mini	75.33	80.57
Gemini 2.5 Flash	48.01	54.79
Gemini 2.5 Pro	4.33	10.62
Claude 3 Haiku	93.43	73.31

Table 8: Inference-only performance of proprietary models on CUREMED-BENCH (averaged across 13 languages).

settings does not ensure robust multilingual reasoning or consistent target-language adherence.

E.4 Per-language Performance of Closed-source Models

We report per language results for proprietary models on CUREMED-BENCH using logical accuracy (Table 9) and language consistency (Table 10). Although aggregate scores are strong for several systems, per language analysis reveals clear cross lingual brittleness. Higher resource languages such as French and Spanish show consistently high logical accuracy and strong target language adherence, with similarly stable behavior in Japanese, Korean, Thai, Turkish, and Vietnamese. In contrast, low resource languages expose systematic failures. Amharic shows frequent target language breakdown for several models, where language consistency drops sharply even when accuracy remains non trivial for some settings. Hausa exhibits a different pattern in which models drift from the target language despite moderate to high logical accuracy, indicating that medical reasoning does not guarantee language control under inference only prompting. Yoruba is the most challenging overall, with both language consistency and logical accuracy decreasing across models. Overall, these results motivate evaluating multilingual medical reasoning using joint measures of correctness and target language fidelity.

F Failure Mode Analysis

To move beyond aggregate leaderboard metrics, we analyze the failure patterns of baseline models and ablation variants on CUREMED-BENCH. Our framework evaluates logical accuracy and language consistency independently on open ended questions with a single verifiable answer, revealing three recurring failure modes that motivate CURE-MED. These failure modes are not mutually exclusive; a single model may exhibit several simultaneously,

and their co-occurrence often masks the true source of performance degradation under standard evaluation. Understanding each failure mode in isolation allows us to trace specific architectural and training choices to their downstream effects, and to design targeted interventions that address the root causes rather than surface level symptoms.

Failure Mode I: Clinically Plausible but Logically Incorrect Reasoning

The most common failure occurs when models produce fluent and well structured responses that contain incorrect clinical reasoning. Figure 3 illustrates this pattern. The Qwen2.5-7B-Instruct baseline generates a coherent Spanish response but predicts an incorrect diagnosis, interpreting a mild viral presentation as early pulmonary infection despite the absence of fever and respiratory distress. Here, the model matches superficial symptom patterns instead of ruling out alternatives through differential reasoning, a behavior also noted in prior studies of medically unsound but persuasive model explanations (Amann et al., 2020; Nori et al., 2023).

CUREMED-BENCH exposes this failure more clearly than multiple choice benchmarks because it requires free form reasoning and a verifiable final answer. We observe severe reasoning errors even in domain specific models. In Table 2, MEDALPACA-7B reaches only 2.47% logical accuracy and 3.50% language consistency, while MEDITRON-7B reaches 2.50% logical accuracy and 0.43% language consistency. These results show that domain specific pretraining alone does not ensure reliable clinical reasoning under open ended evaluation.

Failure Mode II: Language Drift and Target Language Infidelity

A second failure mode appears when models do not answer in the language of the query. They often switch to English or produce mixed language outputs. This tendency is consistent with the English dominant behavior reported in multilingual language models (Nguyen et al., 2023; Cahyawijaya et al., 2024), especially in low resource languages.

Because CUREMED-BENCH evaluates logical accuracy and language consistency separately, it distinguishes reasoning ability from language fidelity and reveals a clear mismatch between them. For example, OPENBIOLLM-LLAMA3-8B attains 36.62% logical accuracy but only 1.47% language consistency in Table 2. The model often reaches the correct medical conclusion but fails to express it in the required target language. Table 10 shows that this problem persists even in strong proprietary systems. Gemini 2.5 Pro achieves only 1.90% language consistency in Amharic and 2.15% in Yoruba, despite showing nontrivial logical accuracy in those languages in Table 9.

These results show that reasoning correctness and target language fidelity do not emerge together by default. We therefore optimize both through the language consistency reward R_{lang} and the code switching aware supervised fine tuning stage described in Sections 3.2 and 3.3.

Failure Mode III: Systematic Low Resource Language Degradation

The third failure mode is the sharp decline in performance for low resource languages. Table 6 shows that Qwen2.5-7B-Instruct achieves 0.00% language consistency and near zero logical accuracy in Amharic, Hausa, Swahili, and Yoruba, while performing much better in French with 71.43% language consistency and 67.86% logical accuracy, and in Japanese with 60.00% language consistency and 68.57% logical accuracy. This pattern points to weak cross lingual generalization rather than a general inability to perform medical reasoning.

Our ablations confirm that training design drives this failure. In Table 3, naïve multilingual supervised fine tuning produces limited and sometimes negative gains, with logical accuracy dropping from 10.83% to 9.50% at the 3B scale. Table 14 provides stronger evidence: when training proceeds from low resource to high resource languages, performance degrades consistently across model sizes, and at the 7B scale, language consistency falls from 85.21% to 64.21%. These findings show that curriculum order strongly affects transfer to underrepresented languages. We address this failure with a high to low resource curriculum and retention aware data mixing ($\alpha = 0.85$), which establish stable early learning signals and support later transfer to lower resource settings.

G Additional baselines

To strengthen our comparisons, we include additional modern baselines that cover general purpose distilled models (DeepSeek Qwen), recent multilingual language models (Qwen3), and medical reasoning focused models (MedReason, MedS3,

Language	GPT-5-nano	GPT-5-mini	Gemini 2.5 Flash	Gemini 2.5 Pro	Claude 3 Haiku
Amharic	5.71	41.90	24.76	0.95	70.48
Bengali	65.00	73.57	38.57	9.29	62.86
French	89.29	93.57	75.71	23.57	90.71
Hausa	78.48	89.87	43.04	3.80	55.70
Hindi	78.36	78.36	62.69	14.93	76.12
Japanese	84.76	84.76	69.52	13.33	87.62
Korean	78.67	80.00	62.67	6.67	73.33
Spanish	89.52	94.29	74.29	18.10	88.57
Swahili	84.29	86.43	48.57	7.14	77.14
Thai	85.71	90.82	64.29	6.12	75.51
Turkish	79.46	84.82	53.57	6.25	70.54
Vietnamese	88.57	88.57	63.81	18.10	84.76
Yoruba	35.48	56.99	25.81	2.15	25.81

Table 9: Logical accuracy (%) of proprietary models on CUREMED-BENCH across 13 languages under inference-only prompting. We report accuracy against the single ground-truth answer.

Language	GPT-5-nano	GPT-5-mini	Gemini 2.5 Flash	Gemini 2.5 Pro	Claude 3 Haiku
Amharic	1.90	24.76	12.38	1.90	92.38
Bengali	39.29	65.71	34.29	1.43	95.71
French	92.86	98.57	67.86	5.00	98.57
Hausa	56.96	43.04	35.44	1.27	73.42
Hindi	53.73	73.88	58.96	8.21	97.76
Japanese	80.00	88.57	56.19	5.71	96.19
Korean	92.00	88.00	56.00	4.00	97.33
Spanish	97.14	98.10	71.43	5.71	91.43
Swahili	82.86	77.86	32.86	2.86	98.57
Thai	94.90	88.78	59.18	9.18	97.96
Turkish	90.18	93.75	52.68	7.14	96.43
Vietnamese	89.52	91.43	60.00	0.95	99.05
Yoruba	27.96	32.26	23.66	2.15	67.74

Table 10: Language consistency (%) of proprietary models on CUREMED-BENCH across 13 languages under inference-only prompting. We report the fraction of outputs that adhere to the requested target language.

Model	English	French	Italian	Spanish
Qwen2.5-1.5B	1.40	6.40	4.80	6.80
↪ CURE-MED	**44.80**	**47.20**	**24.00**	**32.80**
Qwen2.5-3B	24.8	12.00	13.60	13.60
↪ CURE-MED	**48.00**	**50.60**	**36.80**	**48.80**
Qwen2.5-7B	**54.40**	44.00	34.40	48.00
↪ CURE-MED	53.60	**56.80**	**47.20**	**57.60**
Qwen2.5-14B	61.60	54.40	46.40	60.00
↪ CURE-MED	**66.40**	**64.40**	**64.80**	**68.00**
Qwen2.5-32B	**72.80**	**73.60**	64.80	70.40
↪ CURE-MED	72.20	73.00	**72.60**	**76.20**

Table 11: OOD accuracy on MedExpQA across four languages. CURE-MED improves reasoning performance across model sizes, showing cross-lingual generalization to unseen medical questions and languages.

Model	English	Simplified Chinese	Traditional Chinese
Qwen2.5-1.5B	18.50	21.00	16.00
↪ CURE-MED	**37.80**	**59.50**	**47.50**
Qwen2.5-3B	32.50	55.00	36.00
↪ CURE-MED	**41.00**	**68.00**	**54.00**
Qwen2.5-7B	50.50	**73.00**	**60.00**
↪ CURE-MED	**51.50**	70.00	57.00
Qwen2.5-14B	56.00	**80.50**	69.50
↪ CURE-MED	**59.50**	75.00	**70.00**
Qwen2.5-32B	63.00	**84.00**	71.00
↪ CURE-MED	**64.00**	81.00	**76.00**

Table 12: OOD accuracy on MedQA across English and Chinese. CURE-MED improves reasoning performance across model sizes, demonstrating robustness across unseen languages.

and M1 variants). We report both *Language Consistency* and *Logical Accuracy* using the same evaluation protocol.

Table 13 shows that performance generally increases with model scale within each baseline family. Across all comparisons, CURE MED achieves the strongest overall results, with particularly large gains in language consistency and consistent improvements in logical accuracy. The best performance is obtained by CURE MED Qwen2.5 32B, which reaches 94.96 language consistency and 70.04 logical accuracy, substantially exceeding the other baselines.

Model	Consistency (↑)	Accuracy (↑)
DeepSeek-Qwen 7B	1.12±0.31	28.65±0.84
DeepSeek-Qwen 14B	6.74±0.42	33.18±0.79
DeepSeek-Qwen 32B	11.83±0.58	40.27±0.74
Qwen3-4B	24.18±0.95	22.76±0.88
Qwen3-8B	33.96±0.90	31.24±0.82
Qwen3-14B	43.71±0.86	38.62±0.77
Qwen3-32B	55.83±0.80	48.57±0.71
MedReason (8B)	26.74±0.89	34.12±0.81
MedS3 (8B-PRM)	19.28±0.88	36.41±0.80
m1-7B-23K	17.61±0.56	42.90±0.73
m1-32B-1K	48.92±0.83	54.63±0.68
CURE-MED		
CURE-MED-Qwen2.5-1.5B	57.60±0.65	28.32±0.35
CURE-MED-Qwen2.5-3B	74.28±0.60	42.93±0.60
CURE-MED-Qwen2.5-7B	85.21±0.63	54.35±0.50
CURE-MED-Qwen2.5-14B	90.27±0.31	63.74±0.43
CURE-MED-Qwen2.5-32B	**94.96±0.40**	**70.04±0.04**

Table 13: Additional baseline results under the same evaluation protocol. We report mean±std over three runs for language consistency and logical accuracy.

H Curriculum Ordering Ablation

We evaluate whether the ordering of our curriculum affects training outcomes. Our default schedule trains from high resource languages to medium resource languages and then to low resource languages. As a counterfactual, we reverse the schedule and train from low resource languages to medium resource languages and then to high resource languages. Due to resource constraints, we run this ablation for the 1.5B, 3B, and 7B settings only. Table 14 reports language consistency and logical accuracy under the same evaluation protocol.

The results show that curriculum ordering is important. Reversing the order yields a consistent degradation in both metrics across all model scales, indicating that the gains are not an artifact of a particular model size or random variation. At the same time, even under the reverse ordering, CURE MED remains stronger than representative size matched instruction tuned baselines, which suggests that the method is beneficial beyond the choice of schedule. Overall, these findings support the use of the proposed high resource to low resource ordering as it provides the most reliable improvements in both language consistency and logical accuracy.

Model	Consistency (↑)	Accuracy (↑)
Curriculum: High → Medium → Low		
CURE-MED-Qwen2.5-1.5B	57.60 ± 0.65	28.32 ± 0.35
CURE-MED-Qwen2.5-3B	74.28 ± 0.60	42.93 ± 0.60
CURE-MED-Qwen2.5-7B	85.21 ± 0.63	54.35 ± 0.50
Reverse curriculum: Low → Medium → High		
CURE-MED-Qwen2.5-1.5B	33.48 ± 0.86	15.92 ± 0.71
CURE-MED-Qwen2.5-3B	47.76 ± 0.79	23.64 ± 0.66
CURE-MED-Qwen2.5-7B	64.21 ± 0.73	36.58 ± 0.60
Representative size matched baselines		
Qwen2.5-Instruct-1.5B	3.84 ± 0.25	6.20 ± 0.24
Qwen2.5-Instruct-3B	8.39 ± 0.42	10.83 ± 0.60
Qwen2.5-Instruct-7B	25.44 ± 0.36	29.56 ± 0.42

Table 14: Curriculum ordering ablation. We report mean ± standard deviation over three runs for language consistency and logical accuracy.