HiMed: Incentivizing Hindi Reasoning in Medical LLMs

Anonymous ACL submission

Abstract

Medical large language models hold promise for reducing healthcare disparities, yet Hindi remains severely underrepresented. While medical LLMs excel in high-resource languages, their performance degrades sharply in Hindi, particularly on Indian systems of medicine. We argue that robust cross-lingual medical transfer requires Hindi reasoning. To this end, we introduce **HiMed**, a comprehensive Hindi reasoning medical corpus and benchmark suite covering both Western and Indian medicine. We further propose **HiMed-8B**, a Hindi medical reasoning LLM based on LLaMA-3.1-8B-Instruct, through the design of decaying scaffolding reward. Extensive experiments have been conducted, demonstrating consistent improvements in Hindi medical reasoning performance and substantial reduction in the English–Hindi accuracy gap. Ablation studies further validate the contribution of each training stage and reward component. All data and code are available at an anonymous GitHub repository: https://anonymous.4open.science/r/anon-repo-54EC/README.md.

1 Introduction

Large language models (LLMs) have revolutionized medical applications, demonstrating strong performance in high-resource languages (Wang et al., 2025a; Lawrence et al., 2024). Medical decision-making is inherently reasoning-centric: life-critical judgments require causal analysis and factual consistency (Xu et al., 2024). Recent studies further show that strengthening reasoning capabilities of medical LLMs can substantially improve decision accuracy (Chen et al., 2025; Wu et al., 2025a), underscoring the importance of reasoning.

Beyond accuracy, medical reasoning LLMs also have the potential to reduce healthcare disparities by providing scalable clinical expertise to underserved populations across languages (Chinta et al., 2024; d'Elia et al., 2022). However, crossing linguistic boundaries introduces an additional challenge: models must not only transfer reasoning competence, but also align with the linguistic structures and cultural paradigms of the target language (Veselovsky et al., 2025; Sakai et al., 2025).

This challenge is especially pronounced in Hindi-speaking regions, where translation-based deployment exposes hallucination risks. Meanwhile, India's medical ecosystem requires both Western and Indian systems of medicine (Chandra and Patwardhan, 2018). Many core concepts in Indian systems of medicine lack English equivalents. Together, these factors indicate that Hindi medical reasoning benefits from operating directly within Hindi linguistic representations, motivating **Hindi reasoning** as a necessary objective.

In this work, we address Hindi medical reasoning at both the method and data levels. We propose **Decaying Scaffolding Reward Reinforcement Learning**, which progressively shifts optimization from reasoning behavior guidance to task-optimal objectives. To support this design and enable systematic evaluation, we introduce **HiMed**, a comprehensive Hindi medical corpus and benchmark suite spanning both Western and Indian systems of medicine. Experiments based on LLaMA-3.1-8B-Instruct (Team, 2024) produce the **HiMed-8B** model, which shows consistent improvements across medical benchmarks, with ablations validating the contribution of each training stage and reward component. An overview of HiMed and the training framework is shown in Figure 1.

Overall, this work identifies Hindi reasoning as a first-class objective and provides both the data foundation and training methodology. By jointly addressing a series of challenges, our study offers a systematic and extensible framework for adapting general-purpose LLMs to linguistically and culturally specialized medical domains.

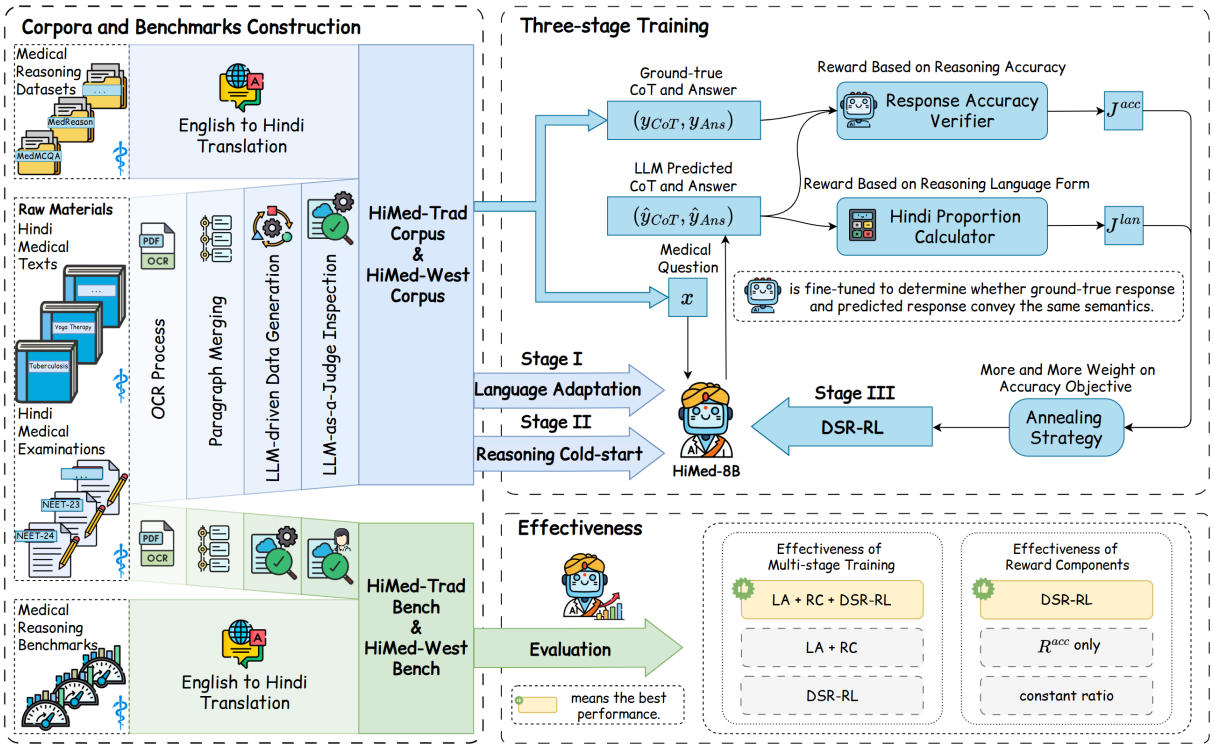

Figure 1: Overview of HiMed and the training framework.

2 Background and Motivation

2.1 Hindi Medical Reasoning Gaps

Evaluation Setup. Prior work has reported substantial cross-lingual degradation and poor alignment with Indian medical knowledge in LLMs (Jin et al., 2024; Qiu et al., 2024; Zhou et al., 2025; Wang et al., 2025b; Devane et al., 2025). To empirically characterize these gaps, we conduct a pilot evaluation on two representative medical benchmarks reflecting distinct medical paradigms: (i) the medical subsets of MMLU-Pro representing Western medicine, and (ii) RET, an Indian medical examination focusing on Indian systems of medicine. We evaluate GPT-4o (Hurst et al., 2024), HuatuoGPT-o1-8B (Chen et al., 2025), and LLaMA3-8B-UltraMedical (Zhang et al., 2024) on both benchmarks in English and Hindi. Δ denotes the accuracy drop from English to Hindi. Full evaluation details are provided in Appendix A.

Benchmark	GPT-4o	HuatuoGPT-o1	UltraMed
MMLU-Pro-En	60.7%	64.7%	61.3%
MMLU-Pro-Hi	50.7%	41.3%	34.0%
MMLU-Pro-Δ	**10.0%**	**23.4%**	**27.3%**
RET-En	88.0%	38.0%	52.0%
RET-Hi	70.0%	26.0%	28.0%
RET-Δ	**18.0%**	**12.0%**	**24.0%**

Table 1: Pilot evaluation of cross-lingual performance.

Results. As shown in Table 1, all evaluated models exhibit substantial English–Hindi accuracy drops, exposing a clear cross-lingual reasoning gap. The gap remains severe for Indian systems of medicine: on RET, open-source models fall to near-random performance in Hindi despite relatively strong English results. These results suggest that medical reasoning acquired predominantly through English-centric training does not consistently generalize to Hindi, especially in settings where effective reasoning depends on culturally grounded medical concepts and practices.

Findings 1. *LLMs exhibit a clear English–Hindi performance gap in both Western medicine and Indian systems of medicine.*

Implication. Taken together, this Hindi performance gap points to a deeper limitation in cross-lingual reasoning transfer: models may produce fluent Hindi outputs without actually reasoning in Hindi, revealing a disconnect between language form and reasoning process. Prior work further shows that cross-lingual reasoning degrades without native-language reasoning supervision (Park et al., 2025; Barua et al., 2025). These observations collectively motivate treating **Hindi reasoning** as a first-class objective.

Resource	Reference	Type	Language	Size	Native Hindi	Western Med.	Indian Med.	CoT
MedMCQA	(Pal et al., 2022)	Corpus/Benchmark	English	193K	No	Yes	No	Yes
BioASQ-QA	(Krithara et al., 2023)	Corpus/Benchmark	English	5K	No	Yes	No	No
PubMedQA	(Jin et al., 2019)	Corpus/Benchmark	English	273K	No	Yes	No	No
MedNLI	(Romanov and Shivade, 2018)	Corpus/Benchmark	English	14K	No	Yes	No	No
ReasonMed	(Sun et al., 2025)	Corpus	English	1.11M	No	Yes	No	Yes
MedReason	(Wu et al., 2025a)	Corpus	English	32K	No	Yes	No	Yes
IndicInstruct	(Gala et al., 2024)	Corpus	Hindi + English	404K	No	No	No	No
UltraMedical	(Zhang et al., 2024)	Corpus	English	410K	No	Yes	No	No
AI4Bharat-IndicQA	(Kunchukuttan et al., 2020)	Corpus	Hindi + Indic	18K	Yes	No	No	No
XLingHealth	(Lab, 2024)	Corpus	Hindi + English	15K	No	Yes	No	No
MILU-cleaned	(Murthyrudra, 2025)	Corpus	Hindi + Indic	74K	Yes	Yes	No	No
AyurGenixAI	(kagglekirti123, 2025)	Corpus	Primarily English	15K	Yes	No	Yes	No
Multilingual Healthcare	(Bagga, 2025)	Corpus	Hindi + English	4K	No	Yes	No	No
MedQA-USMLE	(Jin et al., 2020)	Benchmark	English	1K	No	Yes	No	–
MMLU-Pro-Med	(Wang et al., 2024b)	Benchmark	English	1K	No	Yes	No	–
HLE	(Phan et al., 2025)	Benchmark	English	2K	No	Yes	No	–
HealthBench	(Arora et al., 2025)	Benchmark	English	5K	No	Yes	No	–
GPQA-med	(Rein et al., 2023)	Benchmark	English	0.2K	No	Yes	No	–
MedXpertQA	(Zuo et al., 2025)	Benchmark	English	2K	No	Yes	No	–
BhashaBench-Ayur	(Devane et al., 2025)	Benchmark	Hindi + English	14K	Yes	No	Yes	–
HiMed-West Corpus	–	Corpus	Hindi	116K	No	Yes	No	Yes
HiMed-Trad Corpus	–	Corpus	Hindi	286K	Yes	No	Yes	Yes
HiMed-West Bench	–	Benchmark	Hindi	2K	No	Yes	No	–
HiMed-Trad Bench	–	Benchmark	Hindi	6K	Yes	No	Yes	–

Table 2: Comparison of existing medical corpora and benchmarks.

2.2 Why Translation Is Insufficient for Hindi Medical Reasoning

2.2.1 Reason I: Translation Hallucination

Evaluation Setup. A possible route to Hindi medical LLMs is translation-based deployment, which translates Hindi inputs into English for reasoning and back into Hindi for delivery. However, such methods are unreliable for terminology-dense Hindi medical content. To quantify this risk, we ask GPT-4o to translate 50 English medical sentences into Hindi and have expert translators conduct professional post-editing.

Case I: Translation Hallucination

English Source.
A 15-year-old **African-American male** with a BMI of 22 is brought to his physician by his mother to address concerns about a change in his dietary habits.

GPT Translation.
एक 15 वर्षीय **आफ्रिकान द्रीपानोसोमा संक्रमणज रोग पुरुष** जिसका BMI 22 है, उसकी माता उसे उसके पथ्यसम्बन्धी आदतों में परिवर्तन के बारे में चिंता व्यक्त करने के लिए उसके काया चिकित्सक के पास लाती हैं।

Corresponding English.
(A 15-year-old **African trypanosomal infectious disease male** with a BMI of 22 is brought by his mother to his body physician in order to express concern about a change in his dietary habits.)

Findings 2. *LLM-based translation can introduce semantic hallucinations in terminology-dense Hindi medical content.*

Results and implication. We find that 19 out of 50 translated sentences contain translation-induced semantic hallucinations, including mistranslated medical terms and hallucinated medical information absent from the source. The above example shows how translation-induced hallucinations can silently corrupt the original semantics. Consequently, Hindi medical LLMs must support **Hindi reasoning**, allowing valid semantic structures to be formed directly in the target language.

2.2.2 Reason II: Cultural Grounding

Domain Context. Beyond translation limitations, Hindi reasoning reflects the reality of medical practice in India, where Western medicine and Indian systems of medicine coexist as complementary paradigms (World Health Organization, 2010b; Woodyard, 2011). Medical reasoning is shaped by language and culture, and prior work shows that language-concordant care improves safety and diagnostic effectiveness (Molina and Kasper, 2019; Reaume et al., 2025). Moreover, many core concepts in Indian systems of medicine are linguistically grounded and lack faithful English equivalents.

Findings 3. *Medical reasoning in Indian systems relies on linguistic and cultural grounding that is not preserved under English-centric reasoning.*

Limitation. Existing medical LLMs are predominantly shaped by Western biomedical assumptions and English-centric training data (Pfohl et al., 2024). This limits their ability to reason over Indian systems of medicine in a conceptually faithful manner. Hindi reasoning is not a linguistic preference, but a prerequisite for faithful medical reasoning in the Indian context.

2.3 Why Is Hindi Reasoning Challenging?

Developing effective Hindi medical reasoning models involves two fundamental bottlenecks: a data bottleneck and an optimization bottleneck.

Data bottleneck. As summarized in Table 2, existing medical corpora and benchmarks provide limited support for Hindi medical reasoning. Most resources are either English-only or lack coverage of Indian systems of medicine. Consequently, models lack sufficient training signals to learn Hindi medical reasoning behaviors.

Optimization bottleneck. Beyond data scarcity, Hindi medical reasoning faces a fundamental optimization challenge. Under Hindi prompts, task-level accuracy provides no supervision on how reasoning should be expressed, allowing multiple reasoning styles to be equally optimal. In the following example, both responses select the same correct option, yet differ substantially in how medical reasoning is organized and expressed.

Case II: Reasoning Styles

Question.
एक रोगी में तेज़ बुखार, गर्दन में अकड़न और प्रकाश से डर पाया गया है। सबसे संभावित निदान क्या है?
(A patient presents with high fever, neck stiffness, and fear of light. What is the most probable diagnosis?)
(A) मस्तिष्क ट्यूमर (Brain tumor) (B) मेनिन्जाइटिस (Meningitis) (C) मिर्गी (Epilepsy) (D) माइग्रेन (Migraine)

Reasoning A: English answer-driven.
The correct answer is (B), as these symptoms are consistent with meningitis.

Reasoning B: Hindi symptom-to-mechanism.
तेज़ बुखार और गर्दन में अकड़न मस्तिष्कावरण की सूजन का संकेत देते हैं, जबकि प्रकाश से डर इस सूजन के कारण तंत्रिकीय संवेदनशीलता बढ़ने से होता है। इन लक्षणों का संयोजन विकल्प (B) मेनिन्जाइटिस से मेल खाता है।
(High fever and stiffness in the neck indicate inflammation of the meninges, while fear of light occurs due to increased neurological sensitivity caused by this inflammation. The combination of these symptoms matches option (B), meningitis.)

English translations are provided solely for reader convenience and are not part of the inputs or outputs, except for the option labels (A–D).

Findings 4. *Accuracy-only objectives fail to elicit Hindi reasoning form.*

Implication. As both reasoning A and reasoning B are equally correct, accuracy-based objectives do not provide a gradient signal to discourage answer-driven or English rationale. During continuous optimization, models may therefore converge to objective-equivalent but linguistically misaligned reasoning patterns. Collating findings 2, 3, and 4 together, Hindi medical reasoning requires both additional data and training objectives that explicitly shape reasoning behavior. This motivates the method introduced in the following section.

3 Method

3.1 Hindi Reasoning Objective

We define **Hindi reasoning** as a behavioral objective: under Hindi prompts, the model should organize medical rationales primarily in Hindi, independent of the final answer correctness. Rather than claiming to directly measure reasoning quality, we introduce a lightweight language-form scaffolding that discourages non-target language outputs by measuring the proportion of Hindi tokens in the generated sequence:

$$R^{\text{lan}} = \frac{1}{|y|} \sum_{t=1}^{|y|} \mathbb{I}[y_t \text{ is a Hindi token}], \tag{1}$$

where y denotes the generated sequence. This signal serves only as auxiliary guidance: maximizing Hindi token ratio alone does not ensure correct or well-structured reasoning. We therefore further assess Hindi reasoning through downstream accuracy and controlled human evaluations.

3.2 Decaying Scaffolding Reward Reinforcement Learning (DSR-RL)

Overview. We propose Decaying Scaffolding Reward Reinforcement Learning (DSR-RL), which optimizes reasoning tasks under two complementary rewards: a task-optimal accuracy reward and an auxiliary language-form scaffolding reward that encourages Hindi reasoning. A time-dependent schedule controls the relative influence of the two rewards, enabling early behavioral guidance without constraining the final task-optimal policy.

Notations. We treat the language model as an auto-regressive policy and optimize it accordingly. Let $\mathcal{X}$ denote the space of prompts and $\mathcal{Y}$ the space of output sequences. The policy $\pi_\theta(y \mid x)$ is parameterized by θ, where $x \in \mathcal{X}$ and $y = (y_1, \ldots, y_T) \in \mathcal{Y}$. For each prompt x, we sample a group of K candidate responses $y_i \sim \pi_\theta(\cdot \mid x)$, indexed by $i \in \mathcal{G}(x)$. The overall reward consists of a task-optimal accuracy reward R^{acc} and an auxiliary scaffolding language-form reward R^{lan}. Their relative influence is controlled by a time-dependent coefficient $\lambda(\tau)$.

Task-optimal reward: reasoning accuracy. Each sampled response y_i receives a discrete correctness reward

$$R_i^{\text{acc}} \in \{0, 0.1, 1\}, \tag{2}$$

where 0 denotes responses without reasoning, 0.1 denotes responses with reasoning but incorrect answers, and 1 denotes responses with both reasoning and correct answers, following prior work (Trung et al., 2024; Chen et al., 2025).

Auxiliary scaffolding reward: reasoning language form. To encourage Hindi reasoning, we define a language-form reward to the whole sequence as

$$R_i^{\text{lan}} = \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \mathbb{I}[y_{i,t} \text{ is a Hindi token}] \in (0, 1]. \tag{3}$$

Group-relative advantages. For each reward type $k \in \{\text{acc}, \text{lan}\}$, we compute a group-relative normalized advantage within $\mathcal{G}(x)$:

$$\widetilde{R}_i^k = R_i^k - \text{mean}_{j \in \mathcal{G}(x)} \left(R_j^k \right). \tag{4}$$

$$\widehat{R}_i^k = \frac{\widetilde{R}_i^k}{\text{std}_{j \in \mathcal{G}(x)} \left(\widetilde{R}_j^k \right) + \varepsilon_{\text{norm}}}. \tag{5}$$

The advantage is broadcast to the token level:

$$\widehat{A}_{i,t}^k = \widehat{R}_i^k, \qquad \forall t,\ k \in \{\text{acc}, \text{lan}\}. \tag{6}$$

Optimization objectives. At each token position (i, t), we compute the likelihood ratio

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}. \tag{7}$$

Using the clipped operator

$$\mathcal{C}(r, A) = \min \Big(rA,\ \text{clip}(r, 1 - \varepsilon_{\text{clip}}, 1 + \varepsilon_{\text{clip}})\, A \Big), \tag{8}$$

we define the objective for each reward type as

$$\mathscr{J}^{\text{acc}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|y_i|} \sum_t \mathcal{C} \left(r_{i,t}(\theta), \widehat{A}_{i,t}^{\text{acc}} \right). \tag{9}$$

$$\mathscr{J}^{\text{lan}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|y_i|} \sum_t m_{i,t}\, \mathcal{C} \left(r_{i,t}(\theta), \widehat{A}_{i,t}^{\text{lan}} \right), \tag{10}$$

where $m_{i,t} \in \{0, 1\}$ is a token-level mask that equals 0 for English medical terms and 1 otherwise, leveraging principled code-mixing (Li et al., 2025). To prevent excessive policy drift, we include a KL regularization term:

$$\text{KL} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|y_i|} \sum_t D_{\text{KL}} \big(\pi_\theta(\cdot \mid h_{i,t}) \,\|\, \pi_{\text{ref}}(\cdot \mid h_{i,t}) \big). \tag{11}$$

Final objective. The overall reinforcement learning objective is defined as

$$\mathscr{J}(\tau) = \big(1 - \lambda(\tau)\big)\, \mathscr{J}^{\text{acc}} + \lambda(\tau)\, \mathscr{J}^{\text{lan}} - \beta\, \text{KL}, \tag{12}$$

with the corresponding loss

$$\mathscr{L}_{\text{DSR}}(\tau) = -\mathscr{J}(\tau). \tag{13}$$

Policy optimization is performed using Group Relative Policy Optimization (Shao et al., 2024). Here, $\lambda(\tau) \in [0, 1]$ follows a cosine decreasing function over training iterations, so that auxiliary scaffolding reward dominates early optimization, while task-optimal reward governs the final policy.

3.3 Implementations

We implement the proposed DSR-RL objective via a three-stage training procedure, which progressively prepares the model for reinforcement learning and stabilizes Hindi medical reasoning.

Stage I: Language Adaptation via SFT (LA). The first stage focuses on adapting the model to generate medically grounded Hindi text. This stage stabilizes basic vocabulary and medical expressions, providing a language-aligned initialization for subsequent reasoning-oriented training. Each training example consists of a short factual response without explicit reasoning.

Stage II: Reasoning Cold-start via SFT (RC). With language generation stabilized, the second stage initializes structured medical reasoning. This stage provides a cold-start that shapes the model's reasoning behavior, serving as a stable initialization for subsequent reinforcement learning. Each training example consists of an explicit chain-of-thought followed by a final answer, concatenated as a single output.

Optimization objective. We optimize the model using standard supervised fine-tuning with a token-averaged negative log-likelihood for stage $k \in \{\text{LA}, \text{RC}\}$. We apply teacher forcing and compute losses only on model output tokens. Sequences longer than a fixed maximum length are truncated, and padding tokens are ignored via attention masks.

$$\mathcal{L}_k = \mathbb{E}_{(x,y) \sim \mathcal{D}_k} \Big[-\frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}) \Big] \tag{14}$$

Stage III: DSR-RL. We perform reinforcement learning according to the design described in Section 3.2, with the reference model initialized from the checkpoint obtained at the end of Stage II.

3.4 Data Engineering

Overview. We design the HiMed data engineering pipeline to construct reasoning data under two settings: culture-grounded Hindi reasoning for traditional Indian medicine, and terminology-faithful reasoning for Western medicine. Accordingly, HiMed comprises two parts: HiMed-Trad and HiMed-West. All our data construction processes comply with ethical standards.

3.4.1 HiMed-Trad Corpus and Bench

Scope and coverage. HiMed-Trad is constructed through a unified OCR-based pipeline designed for noisy, real-world Hindi medical scans. A defining feature of HiMed-Trad is its comprehensive coverage of all seven officially recognized Indian systems of medicine.

Data construction pipeline. Authoritative textbook materials are first digitized using DeepSeek-OCR (Wei et al., 2025). The raw OCR outputs are then automatically repaired and page-grounded calibration is applied to recover coherent, self-contained Hindi medical passages. Each passage is subsequently converted into instruction-style instances, with source-grounded medical rationales expressed natively in Hindi.

Evaluation split and leakage prevention. To ensure reliable evaluation, all instances derived from the same source passage are treated as an indivisible unit and assigned exclusively to either the training corpus or the benchmark. This design enforces a strict passage-level split with zero overlap.

Statistics and references. Dataset statistics are summarized in Table 3, with the full taxonomy and source references provided in Appendix H.

System	Corpus Size	Bench Size
Ayurveda	105,397	1000
Yoga	8,499	1000
Naturopathy	20,735	1000
Sowa-Rigpa	119	10
Homoeopathy	130,751	1000
Unani	18,536	1000
Siddha	2,620	1000
Total	**286,657**	**6,010**

Table 3: Coverage of HiMed-Trad.

3.4.2 HiMed-West Corpus and Bench

Task focus. HiMed-West targets Western medical reasoning under Hindi prompts, where accurate handling of standardized medical terminology is critical. We explicitly address this challenge by employing a carefully curated terminology lexicon.

Translation pipeline. We adopt a lexicon-guided translation pipeline that preserves standardized English medical terms when necessary. A high-coverage English–Hindi medical lexicon is constructed, and selected English medical corpora and benchmarks are translated using NLLB-200-3.3B (Team et al., 2022). For translated corpora, medical terms are accompanied by their original English forms in parentheses. In addition, we derive a specialized evaluation subset, HiMed-West Exam, from Indian national examinations on Western medicine: NEET UG.

Mixed-language supervision. To support effective cross-lingual transfer, only a category-stratified subset of data is translated into Hindi. This setting allows the model to be trained on both languages throughout the training process.

Evaluation safety and data sources. All translated training corpora are strictly disjoint from translated benchmarks to prevent leakage. Data sources and statistics are summarized in Table 4. GPQA-med refers to the Organic Chemistry, Molecular Biology, and Genetics subsets.

Source	Released Size	Translated Size
MedMCQA	182,822	91,411
Huatuo-o1 Corpus	19,704	9,852
Medreason	32,682	15,596
MMLU-Pro-Biology	717	717
MMLU-Pro-Health	818	818
GPQA-med	249	249
HiMed-Trad Corpus	286,657	–
HiMed-Trad Bench	6,010	–
HiMed-West Corpus	116,859	–
HiMed-West Bench	2254	–
HiMed-West Exam	470	–

Table 4: Sample composition of HiMed.

3.4.3 Data Quality

We apply multi-level quality control to ensure the reliability of the constructed data. Domain experts manually inspect randomly sampled outputs from key pipeline stages, yielding acceptance rates of 98.0% for OCR passages, 99.5% for HiMed-Trad instances, and 97.0% for translated outputs. Each inspection is conducted according to a predefined checklist. Detailed book and exam lists are provided in Appendix C and D. Complete pipeline details are provided in Appendix E, G, and I.

Model	MMLU-Pro-Biology			MMLU-Pro-Health			GPQA-med			HiMed-West Exam			HiMed-Trad	Avg. En	Avg. Hi
	En	Hi	Δ	En	Hi	Δ	En	Hi	Δ	En	Hi	Δ	Bench		
Baseline Models and Our Model															
GPT-4o	61.9	**55.6**	**6.3**	57.9	**47.0**	**10.9**	**41.8**	30.5	11.3	**78.1**	**78.5**	**-0.4**	62.0	**59.9**	**54.7**
UltraMedical-8B	66.8	41.7	25.1	56.1	26.3	29.8	33.3	20.9	12.4	54.7	40.2	14.5	45.1	52.7	34.8
MedReason-8B	65.8	43.7	22.1	58.7	27.6	31.1	27.3	29.7	**-2.4**	53.2	41.3	11.9	47.2	51.3	37.9
HuatuoGPT-o1-8B	71.0	49.1	21.9	**59.3**	27.6	31.7	37.8	32.9	4.9	59.1	42.8	16.3	56.1	56.8	41.7
Qwen2.5-7B-Instruct	**73.1**	41.8	31.3	54.4	20.3	34.1	28.9	19.7	9.2	62.0	41.5	20.5	53.3	54.6	35.3
LLaMA-3.1-8B-Instruct	66.7	43.5	23.2	52.6	19.7	32.9	31.3	22.5	8.8	52.6	34.3	18.3	51.0	50.8	34.2
HiMed-8B (ours)	69.9	53.8	16.1	57.8	32.5	25.3	38.6	**36.9**	1.7	58.7	46.6	12.1	**76.0**	56.3	49.2
Training Stage Ablation															
Language Adaptation only	68.1	47.0	21.1	54.4	23.6	30.8	33.7	27.7	6.0	53.4	38.3	15.1	70.0	52.4	41.3
Reasoning Cold-start only	68.3	46.2	22.1	53.1	24.7	28.4	34.9	26.9	8.0	52.8	39.6	13.2	70.3	52.3	41.5
DSR-RL only	66.7	42.0	24.7	50.5	19.8	30.7	28.9	19.3	9.6	48.4	35.0	13.4	51.3	48.6	33.5
LA + RC	69.5	49.7	19.8	56.4	27.6	28.8	35.3	30.9	4.4	56.2	42.6	13.6	74.9	54.4	45.1
LA + DSR-RL	69.3	46.7	22.6	51.7	19.3	32.4	36.1	33.7	2.4	51.6	34.1	17.5	69.3	52.2	40.6
RC + DSR-RL	69.0	41.4	27.6	51.2	29.1	22.1	36.1	34.5	1.6	57.9	42.3	15.6	74.1	53.6	44.3
Reward Ablation (DSR-RL)															
Accuracy reward only	70.4	51.3	19.1	58.6	30.2	28.4	38.6	32.1	6.5	59.0	44.7	14.3	75.1	56.7	46.7
Constant rewards ratio	67.5	49.0	18.5	53.8	25.4	28.4	36.1	32.5	3.6	54.0	39.6	14.4	72.0	52.9	43.7

Table 5: Unified evaluation across medical reasoning benchmarks. Accuracy (%) is reported. Within *Baseline Models and Our Model* segment, **bold** highlights the best scores, and underlines indicate the second-best.

4 Experiments

4.1 Experimental Setup

Model. HiMed-8B is initialized from LLaMA-3.1-8B-Instruct. Training is conducted on a single node equipped with 8 NVIDIA H200 GPUs. Across all stages, the model is trained for approximately 56 hours, including 2 hours for LA, 15 hours for RC, and 39 hours for DSR-RL. Details of training data usage and hyperparameter settings are provided in Appendix K.

Benchmarks. We evaluate models on multiple medical reasoning benchmarks. Western medicine is assessed using MMLU-Pro-Biology, MMLU-Pro-Health, GPQA-med, and HiMed-West Exam, while Indian systems of medicine are evaluated using HiMed-Trad Bench. All benchmarks are evaluated under the zero-shot setting.

Task format and evaluation metrics. All evaluations use a multiple-choice format. In the LA stage, we apply greedy decoding with a four-class classification head. For all baselines and the RC and DSR-RL stages, responses are decoded using a robust answer-extraction procedure with progressively relaxed matching to ensure fair and consistent evaluation across models with diverse generation styles. The procedure ranges from strict answer-pattern detection to option-letter and option-text matching, and finally similarity-based option selection. We report accuracy and the English–Hindi accuracy gap $\Delta = \text{En} - \text{Hi}$ as evaluation metrics. All reported results are averaged over three runs.

4.2 Main Results

Table 5 summarizes performance across Western and Indian medical reasoning benchmarks. HiMed-8B consistently outperforms all open-source baselines in Hindi settings and substantially narrows the English–Hindi accuracy gap identified in findings 1, relative to the LLaMA-3.1-8B-Instruct base model, indicating effective cross-lingual transfer of medical reasoning. On HiMed-Trad Bench, HiMed-8B achieves state-of-the-art results, outperforming both open-source baselines and GPT-4o. Overall, these results show that the proposed training framework not only improves cross-lingual medical reasoning accuracy, but also enhances alignment with Hindi medical expression and domain knowledge.

4.3 Ablation Study

We conduct ablation studies to assess the contribution of each training stage and reward component in DSR-RL. As shown in Table 5, removing any stage leads to a clear drop in Hindi performance. Models trained with only the task-optimal accuracy reward or with fixed reward ratios exhibit inferior Hindi performance compared to the full decaying-reward formulation, indicating that the two rewards play complementary roles during the weight-shifting process. To verify that these gains are not driven by superficial Hindi token substitution, we further conduct a human preference study focusing on native Hindi medical reasoning. Two medical experts perform comparisons between HiMed-8B and an otherwise identi-

cal RL variant trained without the language-form reward. Across both Western medicine and Indian systems of medicine prompts, experts consistently prefer HiMed-8B. Preferred responses exhibit symptom-to-mechanism reasoning articulated directly in Hindi and avoid answer-driven explanation patterns. These results indicate that the language-form scaffolding reward induces qualitatively different reasoning behaviors. Full evaluation details are provided in Appendix M.

4.4 Effectiveness of the Language-form Reward

We analyze the training dynamics of the two rewards used in DSR-RL. As shown in Figure 2, both rewards increase throughout training. In particular, the consistent rise of the R^{lan} indicates that the model progressively produces reasoning chains with higher proportions of Hindi tokens. These results confirm that the R^{lan} provides an effective and meaningful optimization signal.

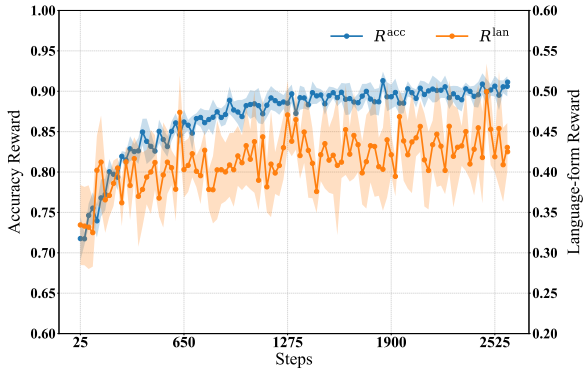

Figure 2: Training dynamics of rewards. The x-axis shows global steps, and the y-axis shows the reward signal. R^{acc} denotes the task-optimal accuracy reward, and R^{lan} denotes the auxiliary language-form reward.

4.5 Reliability of the Accuracy Reward Model

We fine-tune LLaMA-3.2-3B-Instruct model on a bilingual medical reasoning dataset as our accuracy verifier. At inference time, we apply softmax to the binary logits and use the probability of the True class as the confidence score. On the test set, the verifier achieves 0.969 Precision and 0.936 Recall. To assess reliability, we use GPT-5 to construct 300 challenging cases from two perspectives: (i) reversing key logical relations and (ii) slightly modifying numerical values in the reasoning. The verifier reaches 82.3% accuracy on this set. We further compare verifier judgments with annotations from two medical experts on 300 randomly sampled model responses and observe a 96.7% exact agreement rate. Remaining disagreements mainly arise from inherently ambiguous cases. These results support using the verifier as a reliable reward signal in DSR-RL.

5 Related Work

After presenting our methods and empirical results, we situate our work within the broader literature. We briefly review prior work most relevant to our study and defer a comprehensive discussion to Appendix B. Existing resources for medical language modeling remain predominantly English-centric, with Hindi coverage largely limited to small-scale or translated subsets (Narayanan et al., 2024). Prior research has investigated cultural alignment (Li et al., 2024; Pham et al., 2025), reasoning enhancement (Creswell et al., 2023; Liang et al., 2024), reward modeling (Lightman et al., 2024; Ma et al., 2023), and large-scale English medical reasoning LLMs (Chen et al., 2025; Singhal et al., 2023; Zhang et al., 2024). However, these lines of work are typically studied in isolation and rarely intersect with Hindi medical reasoning. In particular, existing corpora and benchmarks fail to jointly support Hindi language supervision, Indian systems of medicine, and high-quality chain-of-thought annotations. As a result, the challenges of culturally grounded, linguistically faithful medical reasoning in Hindi remain largely unaddressed in prior work.

6 Conclusion

In summary, this work identifies Hindi medical reasoning as a critical yet underexplored requirement for medical large language models and introduces **HiMed**, the first large-scale Hindi medical corpus and benchmark suite with native rationales. Building on this foundation, we propose a three-stage adaptation framework centered on **Decaying Scaffolding Reward Reinforcement Learning**, which yields a concrete instantiation in the form of **HiMed-8B**. Extensive experiments show that HiMed-8B consistently improves Hindi medical reasoning performance across benchmarks, validating the effectiveness of the proposed data and training design. More broadly, our study offers practical insights into adapting general-purpose LLMs to linguistically and culturally specialized medical domains.

Limitations

This work has a few limitations.

First, while HiMed substantially expands coverage of Hindi medical reasoning, it does not exhaustively represent all clinical scenarios or regional variations within India. Certain subdomains, rare conditions, and informal clinical expressions may remain underrepresented.

Second, our evaluations focus on multiple-choice medical reasoning benchmarks. Although this setting enables controlled comparison across models and languages, it does not fully capture open-ended clinical reasoning, longitudinal decision-making, or real-world physician–patient interactions. Extending evaluation to free-form and interactive settings remains an important direction for future work.

Third, the proposed training framework is instantiated on a single base model scale (LLaMA-3.1-8B-Instruct). While our results demonstrate consistent gains at this scale, further investigation is needed to assess scalability to larger models and generalization to other low-resource languages.

Finally, although we emphasize Hindi reasoning, our approach does not replace the need for clinical oversight. As with all medical language models, HiMed-8B is intended for research and decision support rather than autonomous clinical use.

Ethical Considerations

While our proposed model is a medical language model with advanced reasoning capabilities, it may still generate hallucinated, incomplete, or inaccurate outputs. As such, it is not suitable for direct deployment in real-world clinical settings. We explicitly restrict the use of HiMed-8B to research and evaluation purposes, and prohibit its application in clinical decision-making, diagnosis, treatment planning, or other safety-critical scenarios where errors could result in harm. Users bear ethical responsibility for respecting these limitations and for avoiding inappropriate downstream use.

This work emphasizes Hindi medical reasoning to support linguistic and cultural alignment in medical AI research. Importantly, accuracy on benchmarks related to Indian systems of medicine reflects consistency with officially examined curricula, authoritative textbooks, and standardized assessment materials. Such results should not be interpreted as clinical efficacy, biomedical validation, or endorsement of any specific medical system. Our benchmarks are designed to evaluate reasoning fidelity within established educational and examination contexts rather than real-world therapeutic effectiveness.

We further acknowledge that medical knowledge is culturally situated and historically contingent. By constructing datasets that cover both Western medicine and Indian systems of medicine, our goal is not to adjudicate between medical paradigms, but to enable transparent and faithful modeling of how medical reasoning is expressed within different linguistic and institutional frameworks. The inclusion of traditional medical content does not imply clinical recommendation and should not be used to guide health decisions without qualified professional oversight.

Finally, all datasets in HiMed are derived from publicly available or licensed sources, national examination materials, or scanned textbooks used for educational purposes. We do not include patient records or personal health information, and no identifiable individual data is involved. To further ensure this, we explicitly ask GPT to ignore these personal information during data generation. In this study, all data, code, and models we use and we release are under the CC-BY 4.0 license or Apache license 2.0. We have verified that their usage complies with the original license agreements and access conditions. Furthermore, participants involved in the human evaluation phase were explicitly informed that their evaluation results would be used solely for academic purposes. A total of eight annotators from two countries participated in this study, and each annotator was proficient in both English and Hindi.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511:1–511:23.

AI4Bharat. 2024. Indicqa: Question answering dataset for indic languages. https://huggingface.co/datasets/ai4bharat/IndicQA. License: CC-BY-4.0; includes Assamese, Bengali, Hindi, Kannada, Marathi, Malayalam, Punjabi, Oriya, Tamil, Telugu.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large

language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Kajol Bagga. 2025. Multilingual healthcare text dataset (hi, en, pu). https://www.kaggle.com/datasets/kajolagga/multilingual-healthcare-text-dataset-hi-en-pu. Accessed: 2026-01-04.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022b. Constitutional AI: harmlessness from AI feedback.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2025. Long chain-of-thought reasoning across languages. *arXiv preprint arXiv:2508.14828*.

Shailaja Chandra and Kishor Patwardhan. 2018. Allopathic, ayush and informal medical practitioners in rural india–a prescription for change. *Journal of Ayurveda and integrative medicine*, 9(2):143–150.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.

Sribala Vidyadhari Chinta, Zichong Wang, Xingyu Zhang, Thang Doan Viet, Ayesha Kashif, Monique Antoinette Smith, and Wenbin Zhang. 2024. Ai-driven healthcare: A survey on ensuring fairness and mitigating bias. *arXiv preprint arXiv:2407.19655*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

CSTT. 2009. औषधि प्रतिकूल प्रतिक्रिया शब्दावली *(Drug Adverse Reaction Glossary)*. वैज्ञानिक तथा तकनीकी शब्दावली आयोग (Commission for Scientific and Technical Terminology), नई दिल्ली (New Delhi). Accessed on 2025-08-30.

CSTT. 2010. रोग-निदान एवं विकृति-विज्ञान शब्दसंग्रह *(Clinical Diagnostics and Pathology Glossary)*. वैज्ञानिक तथा तकनीकी शब्दावली आयोग (Commission for Scientific and Technical Terminology), नई दिल्ली (New Delhi). Accessed on 2025-08-30.

CSTT. 2018a. कैंसर विज्ञान शब्दावली *(Oncology Glossary)*. टाटा स्मारक अस्पताल (Tata Memorial Hospital), मुंबई (Mumbai). Accessed on 2025-08-30.

CSTT. 2018b. शरीर रचना विज्ञान शब्द-संग्रह *(Glossary of Anatomy)*. वैज्ञानिक तथा तकनीकी शब्दावली आयोग (Commission for Scientific and Technical Terminology), नई दिल्ली (New Delhi). Accessed on 2025-08-30.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Alexander d'Elia, Mark Gabbay, Sarah Rodgers, Ciara Kierans, Elisa Jones, Irum Durrani, Adele Thomas, and Lucy Frith. 2022. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. *Family Medicine and Community Health*, 10(Suppl 1):e001670.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

Vijay Devane, Mohd Nauman, Bhargav Patel, Aniket Mahendra Wakchoure, Yogeshkumar Sant, Shyam Pawar, Viraj Thakur, Ananya Godse, Sunil Patra, Neha Maurya, Suraj Racha, Nitish Kamal Singh, Ajay Nagpal, Piyush Sawarkar, Kundeshwar Vijayrao Pundalik, Rohit Saluja, and Ganesh Ramakrishnan. 2025. Bhashabench v1: A comprehensive benchmark for the quadrant of indic domains.

Government of NCT of Delhi Directorate of AYUSH. Ashtāṅga ayurveda overview. https://ayush.delhi.gov.in/ayush/ayurveda. Accessed 2025-09-20.

John W Ely, Jerome A Osheroff, Mark H Ebell, George R Bergus, Barcey T Levy, M Lee Chambliss, and Eric R Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *Bmj*, 319(7206):358–361.

Pharmacopoeia Commission for Indian Medicine & Homoeopathy. Official website. https://pcimh.gov.in/.

Central Council for Research in Ayurvedic Sciences (CCRAS). 2024. Sowa-rigpa (amchi): Overview and programme note. Technical report.

Central Council for Research in Homoeopathy (CCRH). Official website. https://ccrhindia.ayush.gov.in/.

Central Council for Research in Siddha (CCRS). Official website. https://siddhacouncil.com/ccrs/?page_id=16.

Central Council for Research in Unani Medicine (CCRUM). Official website. https://ccrum.res.in/.

Central Council for Research in Yoga & Naturopathy (CCRYN). Council overview and programmes. https://namayush.gov.in/content/central-council-research-yoga-naturopathy-ccryn. Accessed 2025-09-20.

Jay P. Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M., Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, V. Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing hindi instruction-tuned LLM. *arXiv preprint arXiv:2401.15006*.

S. C. Sen Gupta and S. C. Kapoor. 1955. *Dictionary of Medicine (English to Hindi)*. Hindi Sahitya Sammelan, Prayag, India.

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. 2025. Adapters for altering LLM vocabularies: What languages benefit the most? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

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*, Virtual Event, Austria.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yogini S. Jaiswal and Leonard L. Williams. 2017. A glimpse of ayurveda – the forgotten history and principles of indian traditional medicine. *Journal of Traditional and Complementary Medicine*, 7(1):50–53.

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. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2627–2638. ACM.

kagglekirti123. 2025. Ayurgenixai: Ayurvedic dataset. https://www.kaggle.com/datasets/kagglekirti123/ayurgenixai-ayurvedic-dataset. Accessed: 2026-01-04.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

CLAWS Lab. 2024. Xlinghealth: A cross-lingual healthcare question answering benchmark. https://huggingface.co/datasets/claws-lab/XLingHealth. License: Apache-2.0; includes English, Spanish, Chinese, Hindi.

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. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5848–5864. Association for Computational Linguistics.

Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Mataric, Daniel J. McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *arXiv preprint arXiv:2403.14814*.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. Cross-lingual optimization for language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15100–15119, Vienna, Austria. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. 2025. The impact of language mixing on bilingual LLM reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32519–32536, Suzhou, China. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Renqian Luo, Luyang Sun, Yingce Xia, and et al. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning.

Government of India Ministry of Ayush. 2023. Replies to parliament questions (budget session 2023): Yoga & naturopathy and related modalities. Technical report, Ministry of Ayush. Written replies in Lok Sabha; item dated 04 Aug 2023 referenced therein.

Rosa L. Molina and Jennifer Kasper. 2019. The power of language-concordant care: a call to action for medical schools. *BMC Medical Education*, 19(1):378.

Murthyrudra. 2025. Milu-cleaned: cleaned version of the milu benchmark. https://huggingface.co/datasets/murthyrudra/milu-cleaned. 80,000 MCQs over 11 Indic languages, 8 domains and 41 subjects; includes a validation set.

Vaishak Narayanan, Prabin Raj K. P., and Saifudheen Nouphal. 2024. Suvach - generated hindi QA benchmark. *arXiv preprint arXiv:2404.19254*.

Kolkata National Institute of Homoeopathy (NIH). Official website. https://www.nih.nic.in/.

Chennai National Institute of Siddha (NIS). Official website. https://www.nischennai.org/.

Bengaluru National Institute of Unani Medicine (NIUM). Official website. https://www.nium.in/.

Ministry of Ayush. a. Sowa-rigpa overview (ayush education portal). https://ayushedu.bisag-n.gov.in/AYUSH_EDU. Accessed 2025-09-20.

Ministry of Ayush. b. Unani overview (ayush education portal). https://ayushedu.bisag-n.gov.in/AYUSH_EDU/unani.

Ministry of Ayush and MDNIY. 2025. Common yoga protocol (june 2025 edition). Technical report, Morarji Desai National Institute of Yoga, Ministry of Ayush.

Government of India. 2020. The national commission for homoeopathy act, 2020 (act no. 15 of 2020). https://nch.org.in/. Act notified; see NCH official portal for text and subsequent regulations.

National Institute of Sowa-Rigpa (NISR). Official website. https://infohub.ayush.gov.in/autonomous/19/. Accessed 2025-09-20.

World Health Organization. 2022. Who benchmarks for training in yoga. In development; see WHO Traditional Medicine programme updates.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan

Lowe. 2022. Training language models to follow instructions with human feedback.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Jisu Park, Lucas Bandarkar, Artem Vazhentsev, Minjoon Choi, Swaroop Mishra, Chitta Baral, and Daniel Khashabi. 2025. Xtom: Exploring the multilingual theory of mind for large language models. *arXiv preprint arXiv:2506.02461*.

PCIM&H. a. Downloads: Unani pharmacopoeia of india (upi) / national formulary of unani medicine (nfum). https://www.portal.pcimh.gov.in/content/downloads.

PCIM&H. b. Homoeopathic pharmacopoeia of india (hpi) / homoeopathic pharmacopoeia laboratory (hpl) – publications portal. https://pcimh.gov.in/show_content.php?lang=1&level=1&lid=57&ls_id=59.

PCIM&H. c. Siddha pharmacopoeial publications and siddha formulary of india. https://pcimh.gov.in/show_content.php?lang=1&level=1&lid=55&ls_id=57.

PCIM&H. d. The unani pharmacopoeia of india – publication list. https://pcimh.gov.in/show_content.php?lang=1&level=1&lid=56&ls_id=58.

Stephen R. Pfohl, Aleksandra Foryciarz, Harsha Nori, Michael Chen, and Tatsunori Hashimoto. 2024. A toolbox for surfacing health equity harms and biases in LLM-generated medical answers. *Nature Medicine*, 30:1731–1740.

Viet Thanh Pham, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2025. CultureInstruct: Curating multicultural instructions at scale. 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 9207–9228, Albuquerque, New Mexico. Association for Computational Linguistics.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, and 81 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

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. *arXiv preprint arXiv:2402.13963*.

Hamed Rahimian and Sanjay Mehrotra. 2019. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Katrina T. Reaume, Maria Diaz, Sameer R. Patel, and Shoshana Cohen. 2025. Patient-physician language concordance and cardiovascular outcomes. *JAMA Network Open*, 8(1):e243803.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Shintaro Sakai, Jisun An, Migyeong Kang, and Haewoon Kwak. 2025. Somatic in the east, psychological in the west?: Investigating clinically-grounded cross-cultural depression symptom expression in llms.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, and 11 others. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

SNOMED International. 2024. SNOMED CT. Accessed: 2024-06-20.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and Yu Rong. 2025. ReasonMed: A 370K multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26457–26478, Suzhou, China. Association for Computational Linguistics.

Llama Team. 2024. The llama 3 herd of models.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

The State Education Department and The University of the State of New York. 2018. High school level chemistry glossary (english–hindi): Translation of chemistry terms based on the coursework for chemistry grades 9 to 12. Updated October 2018.

Luong Quoc Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7601–7614. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2024. Milu: A multi-task indic language understanding benchmark. *arXiv preprint arXiv:2411.02538*.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized cultural knowledge is conserved and controllable in large language models.

Guoxin Wang, Minyu Gao, Shuai Yang, Ya Zhang, Lizhi He, Liang Huang, Hanlin Xiao, Yexuan Zhang, Wanyue Li, Lu Chen, Jintao Fei, and Xin Li. 2025a. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. 2025b. Language mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2637–2665, Suzhou, China. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024a. Apollo: An lightweight multilingual medical LLM towards democratizing medical AI to 6b people. *arXiv preprint arXiv:2403.03640*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024c. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

C. Woodyard. 2011. Exploring the therapeutic effects of yoga and its ability to increase quality of life. *International Journal of Yoga*, 4(2):49–54.

World Health Organization. 2010a. *Benchmarks for training in Naturopathy*. WHO, Geneva.

World Health Organization. 2010b. Benchmarks for training in traditional/complementary and alternative medicine: Benchmarks for training in naturopathy. Publication date: 5 June 2010. ISBN: 9789241599658.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Changin Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025a. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Linjuan Wu, Hao-Ran Wei, Baosong Yang, and Weiming Lu. 2025b. From english to second language mastery: Enhancing LLMs with cross-lingual continued instruction tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 23006–23023.

Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, Jin Lu, Wei Zhang, Tuo Zhang, Lu Zhang, Dajiang Zhu, Xiang Li, Wei Liu, Quanzheng Li, Andrea Sikora, and 3 others. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21125–21147, Vienna, Austria. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jinfang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Huichi Zhou, Zehao Xu, Munan Zhao, Kaihong Li, Yiqiang Li, and Hongtao Wang. 2025. Moral reasoning across languages: The critical role of low-resource languages in llms. *arXiv preprint arXiv:2504.19759*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*.

A Pilot Evaluation

For both MMLU-Pro and RET, we randomly sampled 150 test cases using a fixed random seed and translated them into Hindi to facilitate a controlled English–Hindi comparison under identical question content, following the translation pipeline described in Appendix I. For MMLU-Pro, we only included biology and health subsets. Specifically, RET comprises 100 questions from the 2023 AIAPGET exam, evenly drawn from four disciplines, and the remaining 50 questions from BNYS First Year, Part II examination papers. All RET questions were formatted as multiple-choice items with standardized options before translation and evaluation. GPT-4o was asked to output simply one option as the answer. HuatuoGPT-o1-8B and LLaMA3-8B-UltraMedical were deployed locally. For these two models, we adopted the option with the highest predicted probability as the model's answer. This follows each model's standard inference interface and avoids additional prompt engineering that could confound cross-model comparison. We reported the number of correct predictions out of the 150 test cases as the accuracy. The Δ value denotes the performance gap, calculated as the difference between Hindi and English accuracy. All evaluations are conducted in a zero-shot setting. The reported results are obtained through single run.

B Related Work

Datasets. The dual nature of the Hindi medical ecosystem (Chandra and Patwardhan, 2018) imposes three key requirements on corpora for training Hindi medical reasoning LLMs: Hindi language coverage, inclusion of Indian systems of medicine, and support for chain-of-thought training. No existing resource satisfies all three requirements. MedMCQA (Pal et al., 2022), although derived from the NEET PG and AIIMS examinations, is publicly released almost entirely in English and focuses on Western medicine. UltraMedical (Zhang et al., 2024), BioASQ-QA (Krithara et al., 2023), PubMedQA (Jin et al., 2019), MedNLI (Romanov and Shivade, 2018), ReasonMed (Sun et al., 2025), and MedReason (Wu et al.,

2025a) provide large-scale biomedical MCQ, QA, or NLI data, but are exclusively English. General multilingual corpora such as IndicInstruct (Gala et al., 2024), AI4Bharat-IndicNLP (Kunchukuttan et al., 2020), AI4Bharat-IndicQA (AI4Bharat, 2024), IndicLLM-Suite (Khan et al., 2024), XLingHealth (Jin et al., 2024; Lab, 2024), and MILU-cleaned (Verma et al., 2024; Murthyrudra, 2025) contain Hindi text, but either lack medical specialization or rely primarily on translated content. AyurGenixAI (kagglekirti123, 2025) covers Indian medicine but remains predominantly English, while the Multilingual Healthcare Text Dataset (Bagga, 2025) contains diverse medical content without providing a ready-to-use QA or MCQ format. Moreover, only MedMCQA, ReasonMed, and MedReason explicitly include CoT annotations. Overall, no existing corpus adequately covers both Western and Indian medicine in Hindi with high-quality CoT supervision, leaving a substantial gap for downstream model training.

Benchmarks. Benchmark imposes similar requirements of native Hindi coverage and inclusion of Indian medicine. Native Hindi evaluation is crucial to avoid translation bias and cultural information loss (Jin et al., 2024; Qiu et al., 2024; Wang et al., 2024a). However, few existing benchmarks meet these criteria. MedQA-USMLE (Jin et al., 2020), MMLU-Pro-Med (Hendrycks et al., 2021; Wang et al., 2024b), HLE-med (Phan et al., 2025), and HealthBench (Arora et al., 2025) are entirely English-based. This limitation extends to advanced biomedical reasoning benchmarks such as GPQA-med (Rein et al., 2023) and MedXpertQA (Zuo et al., 2025). BhashaBench-Ayur (Devane et al., 2025) consists primarily of OCR outputs from examinations, but merely focuses on Ayurveda, leaving a huge gap on other streams such as Yoga. Taken together, current benchmarks remain English-centric and fail to support comprehensive reasoning evaluation across both Western and Indian systems of medicine.

Cultural alignment. Recent work explores cultural grounding through region-specific data curation and instruction tuning (Li et al., 2024; Pham et al., 2025), or through principle-based approaches that promote normative adherence (Bai et al., 2022b). Culture-aware benchmarks have also been proposed (Wang et al., 2024c; Karinshak et al., 2024). However, most of these efforts focus on reducing cultural bias or assessing cultural values, and remain largely English-centric. Culturally aligned methods for medical reasoning in Hindi, especially those spanning both Western and Indian systems of medicine, are still largely unexplored.

Reasoning enhancement. A broad range of techniques have been proposed to improve reasoning, including chain-of-thought prompting (Wei et al., 2022), self-consistency (Wang et al., 2023), verifier-guided decoding (Creswell et al., 2023), Tree-of-Thought (Yao et al., 2023), debate frameworks (Liang et al., 2024), and process supervision (Lightman et al., 2024). These methods improve robustness and accuracy on complex reasoning tasks, but are rarely adapted to multilingual or medical settings. None explicitly addresses the linguistic and cultural fidelity required for Hindi medical reasoning.

Reward modeling. Reward modeling is central to LLM alignment and can be broadly categorized into three paradigms. (1) Outcome reward models score final answers or pairwise preferences and form the basis of standard RLHF pipelines (Ouyang et al., 2022). (2) Process reward models (PRMs) provide step-level supervision for reasoning chains (Lightman et al., 2024), with recent work integrating PRM signals directly into reinforcement learning (Ma et al., 2023). (3) LLM-as-a-Judge (Zhu et al., 2025; Liu et al., 2023) replaces human annotation with strong LLM evaluators, and has been applied both for evaluation and as a reward source in RLHF or RLAIF pipelines (Bai et al., 2022a; Cobbe et al., 2021; Chen et al., 2024). Multi-objective training has been explored in mathematical reasoning (Shao et al., 2024) and biomedical reasoning (Wang et al., 2025a), typically by combining reward signals or enforcing structured trade-offs. Distributionally robust optimization (DRO) (Rahimian and Mehrotra, 2019) further mitigates reward imbalance by optimizing worst-case performance. Despite these advances, reward modeling in medical domains remains largely accuracy-centric, with limited attention to language nativeness or cultural fidelity.

Medical Reasoning LLMs. Recent research has increasingly focused on enhancing the medical reasoning capabilities of large language models (LLMs). Representative systems include Med-PaLM2, BioGPT, PMC-LLaMA, BioMistral, HuatuoGPT-o1 and other domain-adapted LLMs

(Luo et al., 2022; Singhal et al., 2023; Wu et al., 2023; Chen et al., 2025; Labrak et al., 2024), which demonstrate strong performance on exam-style and reasoning-oriented benchmarks such as USMLE, MedQA, and MMLU (Jin et al., 2020; Hendrycks et al., 2021; Wang et al., 2024b). Existing approaches typically improve medical reasoning through three major methodological directions: (1) instruction tuning, which leverages curated or synthetic medical question–answer and diagnostic datasets to strengthen domain-specific reasoning (Jin et al., 2020); (2) reasoning modeling, such as Chain-of-Thought prompting or stepwise diagnostic reasoning templates, to encourage multi-step inference (Wei et al., 2022); and (3) preference tuning, including preference-based learning and reinforcement learning, to improve consistency and medical plausibility of generated reasoning paths (Chen et al., 2025).

Despite these advances, the vast majority of existing medical reasoning LLMs remain English-centric, with underlying medical knowledge and reasoning paradigms deeply rooted in Western clinical practice. Cross-lingual medical reasoning scenarios have received little systematic attention.

Cross-Lingual Transfer in Large Language Models. Cross-lingual transfer in LLMs typically combines language-adaptive pre-training, instruction tuning or alignment, and representation analysis. Language-Adaptive Pre-training (LAPT) continues pre-training on target-language corpora to build core linguistic competence, often complemented by vocabulary-oriented methods such as explicit vocabulary expansion (Cui et al., 2023) or adapter-based embedding learning with fixed model weights (Han et al., 2025). To transfer instruction-following and reasoning, prior work constructs multilingual instruction data and leverages cross-lingual in-context learning, translation-based strategies, and distillation (Wu et al., 2025b). Preference alignment has also been extended to multilingual settings by transferring preferences from English-aligned models via implicit rewards or modified DPO objectives on translation-augmented data, encouraging same-language responses (Yang et al., 2025; Lee et al., 2025). Finally, representation diagnostics suggest that middle layers are relatively language-agnostic, motivating explicit middle-layer alignment during task fine-tuning (Liu and Niehues, 2025).

C Books

This section lists representative books and official documents used in constructing the HiMed-Trad corpus. The resources span multiple Indian systems of medicine, including Ayurveda, Yoga, Naturopathy, Unani, Siddha, Sowa-Rigpa, and Homoeopathy. The listed materials include classical treatises, official guidelines, and modern textbooks, reflecting both historical foundations and contemporary clinical practice.

C.1 Ayurveda

- **नाड़ी-दर्शन**. A text describing the Ayurvedic practice of pulse diagnosis.
- **Aahar Chikitsa**. A book on dietary therapy within the Ayurvedic system of medicine.
- **Anubhut Yog**. A work describing traditional Ayurvedic formulations and practices based on experiential knowledge.
- **Journal of Ayurveda Case Reports**. A peer-reviewed journal publishing case reports in Ayurveda, issued by the All India Institute of Ayurveda.
- **Essential Drugs List – Ayurveda**. An official document listing essential Ayurvedic medicines, including indications and usage contexts.
- **Charak Samhita**. A classical Ayurvedic text describing core principles and clinical practices.
- **चरक-संहिता: द्वितीय खंड**. The second volume of the Charak Samhita, a foundational Ayurvedic text.
- **Ayurvedic Standard Treatment Guidelines**. Standardized treatment guidelines prepared by the Ministry of AYUSH, Government of India.
- **रसराज महोदधि**. A comprehensive text covering a wide range of Ayurvedic medical practices.
- **संपूर्ण चिकित्सा (अष्टांग हृदयम पर आधारित)**. A compendium of Ayurvedic practices based on the Ashtanga Hridayam tradition.
- **Scientific Basis for Ayurvedic Therapies**. A book discussing scientific perspectives on Ayurvedic therapeutic methods.

- **Sidhparikshapadati Part I**. A text focusing on Ayurvedic diagnostic principles and treatment approaches.
- **Sushruta Samhita**. A classical Ayurvedic text with a focus on surgery and medical procedures.
- **The Complete Book of Ayurvedic Home Remedies**. A guide to household-level Ayurvedic remedies.
- **अगद तंत्र (विष चिकित्सा) प्रथम भाग**. A text on Agad Tantra, the Ayurvedic discipline concerned with toxicology.
- **सुश्रुतसंहिता: शारीरस्थान हिंदी अनुवाद सहित**. A Hindi translation of the Sharirasthana section of the Sushruta Samhita, covering anatomy and physiology.
- **स्वास्थ्य अमृतावली**. A guide to health maintenance based on Ayurvedic principles.
- **प्रायोगिक वनऔषधि विज्ञान**. A book on the practical application of forest-based herbal medicine in Ayurveda.

C.2 Yoga

- **योगासन चिकित्सा**. A book on the therapeutic applications of yoga postures.
- **श्री योगसूत्र**. A Hindi exposition of the Yoga Sutras, a foundational text of yoga philosophy.
- **योगचिकित्सा: अनुपान के साथ**. A work discussing yoga therapy in conjunction with medicinal practices.
- **योग के रहस्य**. A book exploring yogic practices and their philosophical foundations.

C.3 Naturopathy

- **यौन स्वास्थ्योपयोगी**. A text addressing sexual health through naturopathic practices.
- **प्राकृतिक चिकित्सा क्यों?**. A book discussing the principles and benefits of naturopathy.
- **रोगों की नई चिकित्सा: New Science of Healing**. A work presenting naturopathic approaches to disease treatment.
- **Ankh-Ka: The Art of Naturopathy**. A book on holistic health and natural healing methods.
- **Upwash Chikitsa**. A text on fasting-based therapeutic practices in naturopathy.
- **स्वास्थ्य और जलचिकित्सा**. A book on naturopathy and hydrotherapy.

C.4 Unani

- **युनानी चिकित्सा के आधारभूत सिद्धांत**. A work outlining foundational principles of Unani medicine.
- **Unanani Chikitsasar**. A book describing Unani medical theory and practice.
- **संयुक्त प्रान्त आयुर्वेदिक तथा यूनानी तिब्बी चिकित्सा पद्धति अधिनियम, 1939**. A legislative document concerning Ayurvedic and Unani medical systems.

C.5 Siddha

- **सिद्ध वनौषधि**. A book on medicinal plants used in the Siddha system.
- **सिद्ध प्रयोग – Part I**. A Hindi commentary on Siddha medical practices.

C.6 Sowa-Rigpa

- **नेपाली बौद्ध परम्परामा सोवारिग्पा पद्धतिको अवस्था**. A study of the Sowa-Rigpa tradition within Nepali Buddhist contexts.

C.7 Homoeopathy

- **होम्योपैथिक चिकित्सा**. An introductory text on homeopathic medicine.
- **होम्योपैथिक मटेरिया मेडिका (रिपर्टरी सहित)**. A comprehensive reference on homeopathic remedies with repertory.
- **होमियोपैथिक पारिवारिक चिकित्सा (Paribarik Chikitsa)**. A guide to family-oriented homeopathic treatments.
- **होम्योपैथिक औषधियों का सचित्र विवरण**. A pictorial reference of homeopathic medicines.
- **होम्योपैथी चिकित्सा**. A general overview of homeopathic therapeutic practices.
- **होम्योपैथिक दर्शन**. A book discussing philosophical foundations of homeopathy.
- **सरल होम्योपैथिक चिकित्सा**. A practical guide to homeopathic treatment.

- **पशु रोगों में होम्योपैथी: एक वैकल्पिक चिकित्सा**. A work on veterinary applications of homeopathy.

- **सरल होम्योपैथी इलाज**. A handbook of simplified homeopathic treatments.

D Exams

This section lists national- and university-level examination papers used as primary sources for constructing the HiMed benchmarks and corpora. All materials are publicly available examination papers and do not contain personal or private information.

National Eligibility cum Entrance Test (NEET-UG). This examination serves as the uniform entrance test for admission to undergraduate medical programs across India. Papers from the years 2018, 2020, 2021, 2023, and 2024 were used.

All India AYUSH Post Graduate Entrance Test (AIAPGET). AIAPGET is an entrance examination for postgraduate programs in Ayurveda, Unani, Siddha, and Homoeopathy. We incorporate papers from the Ayurveda, Unani, and Siddha disciplines.

Banaras Hindu University Research Entrance Test (RET) – Swasthavritta & Yoga. This Research Entrance Test (RET) paper is used for the Swasthavritta & Yoga discipline at Banaras Hindu University. The examination consists of two parts: research methodology and subject-specific knowledge. The latter includes concepts from Ayurvedic Swasthavritta, public health, Yoga, and naturopathy.

University of Patanjali – Bachelor of Naturopathy and Yogic Science (BNYS), First Year, Part II. This collection includes examination papers from the BNYS First Year, Part II examinations conducted in December 2023. The papers cover foundational subjects such as Human Anatomy and Human Physiology, each divided into two parts. Additional papers focus on the Philosophy of Nature Cure and the Fundamental Principles of Integrated Systems of Medicine, incorporating concepts from Ayurveda, Homoeopathy, and Unani.

E OCR Process

This section describes the OCR pipeline used to construct **HiMed-Trad Corpus**, **HiMed-Trad Bench**, and **HiMed-West Exam**. Segment merging and LLM-as-a-judge merging are applied only to the HiMed-Trad, as exam questions are already provided in well-separated question-level formats and therefore do not require cross-segment reconstruction.

E.1 DeepSeek-OCR and Basic Cleaning

We first convert each raw PDF into high-resolution page images to support subsequent cross-modal verification. We then apply DeepSeek-OCR to extract page-wise text. The OCR outputs are exported as raw, unprocessed MultiMarkdown (MMD) files and retained without manual correction, in order to preserve original structure and OCR artifacts for systematic downstream handling. These outputs often contain OCR inconsistencies, including broken words, spurious line breaks, residual layout markers, and page-level noise. A basic cleaning stage normalizes punctuation and whitespace and attaches document-level metadata. We also keep page indices and source identifiers alongside each intermediate segment to maintain traceability back to the scanned pages. The goal is to produce a noise-reduced yet structurally faithful representation of the scanned material.

E.2 Segment Merging

OCR systems frequently fragment semantically continuous content across adjacent lines or pages. To restore coherence, we compare adjacent fragments using embedding-based cosine similarity with all-MiniLM-L6-v2 (Hearst, 1997; Reimers and Gurevych, 2019; Wang et al., 2020). Fragments above a similarity threshold are merged, while headings and list markers are preserved as independent units. After merging, each segment is stored as a self-contained record with its merged text and the corresponding page span, enabling later alignment and verification. This step reconstructs medically meaningful segments while maintaining the original narrative structure.

E.3 LLM-as-a-Judge Merging

For HiMed-Trad Corpus, each merged segment is classified by GPT-4o along four dimensions: (i) *Hindi validity*, whether the text is primarily written in Devanagari Hindi, (ii) *medical relevance*, whether the content expresses meaningful medical knowledge, (iii) *referential ambiguity*, whether unresolved references block medical interpretation,

Branch	Raw OCR	Cleaned	Merged	4-way Filter	Calibrated	Final "No-Problem"
Ayurveda	75,164	46,040	32,420	14,424	9,085	9,054
Yoga	11,582	5,698	3,090	1,922	844	839
Naturopathy	17,683	7,200	3,799	2,218	1,527	1,451
Sowa-Rigpa	68	57	37	21	10	10
Homoeopathy	64,752	39,953	18,400	12,673	9,316	9,226
Unani	6,447	5,148	2,900	2,477	1,487	1,486
Siddha	3,125	2,122	1,385	671	465	453

Table 6: Number of Hindi medical segments for each Indian medical branch after each stage of the OCR pipeline.

and (iv) *structural role*, whether the segment functions as a title or heading.

Segments not primarily written in Hindi are discarded. Segments flagged as ambiguous are further processed via a backward merging strategy: we iteratively concatenate the ambiguous segment with the closest preceding segment until an unambiguous segment is concatenated. Merged segments satisfying the pattern Hindi=True, Medical=True, Ambiguity=False, Heading=False are accepted into the high-confidence pool. We persist both the 4-way decision and the resulting merged unit in the intermediate files, so downstream calibration and scoring operate only on the high-confidence pool.

Prompt: LLM-as-a-Judge Merging

You are a classification tool. Analyze the following text and output exactly 4 lines:

is_hindi: True/False
is_medical: True/False
has_ambiguity: True/False
is_title_or_heading: True/False

Definitions:

- is_hindi: True if text is primarily Hindi in Devanagari.
- is_medical: True if it contains medical knowledge (Western/Indian).
- has_ambiguity: True only if the text contains a strong, clearly identifiable, critical referential ambiguity that satisfies all conditions below.

Conditions for has_ambiguity=True:

1. The ambiguity concerns a key subject/object/symptom/treatment/causal relation and blocks correct medical interpretation.
2. The ambiguous element cannot be resolved from the text itself, even with generous natural-language inference.
3. The ambiguity spans the entire text (the text never provides enough information to clarify the referent).
4. It is not a normal omission/abbreviation/vagueness/stylistic shortening/common medical phrasing.
5. It is not a minor unclear phrase that does not affect the main meaning.

If the text is understandable overall (even with minor unclear parts), then has_ambiguity must be False.

Rules:

- Output **only** the 4 lines above.
- No explanations.
- Use Python-style booleans: True/False.

Text:
{text}

E.4 LLM Calibration

Accepted segments are aligned with their corresponding page images and re-evaluated by GPT-4o, which is instructed to correct only OCR-level imperfections. This stage fixes minor transcription errors while explicitly prohibiting semantic alterations, additions, or deletions. Cross-checking against page images ensures that the calibrated text remains maximally faithful to the original scanned document.

Prompt: LLM Calibration

You are an OCR post-correction assistant.
You are given OCR text that may span one or more scanned pages. You are also given all corresponding page images.

Picture:

- The corresponding page images are attached as image files.

Rules:

- Only correct characters, punctuation, spacing, and diacritics.
- Do **not** add new content or delete meaningful content.
- Do **not** translate.
- Output **only** the corrected text.

OCR text:
{text}
Page image:
{image}

E.5 LLM-as-a-Judge Scoring

GPT-4o assigns each calibrated segment one of three quality levels: NO_PROBLEM (readable and semantically intact), POSSIBLE_ISSUE (minor irregularities with preserved meaning), or DEFINITE_ISSUE (severe corruption obstructing interpretation). Only NO_PROBLEM segments are retained for downstream annotation and instruction generation; the other two categories are archived for potential manual recovery. Notably, the LLM-as-a-judge is used strictly for structural and quality control and does not introduce new content or alter medical meaning.

Prompt: LLM-as-a-Judge Scoring

You are an OCR text quality evaluator.
Your main goal is to determine whether the text contains serious, meaning-breaking OCR corruption. Default to NO_PROBLEM unless there is clear evidence of major corruption.

Classification rules:

1. NO_PROBLEM:
   - The text is readable, understandable, and mostly coherent.
   - Minor typos, spacing issues, missing matras, imperfect Hindi, or small artifacts are not major OCR problems.
   - If a normal Hindi reader can infer the meaning easily, choose NO_PROBLEM.
2. POSSIBLE_ISSUE:
   - One or two suspicious fragments may be OCR errors, but overall meaning is still understandable.
3. DEFINITE_ISSUE:
   - The text is seriously corrupted (broken words, incomplete fragments, nonsense sequences) and the meaning cannot be recovered.

Output format:

- Line 1: exactly one of NO_PROBLEM / POSSIBLE_ISSUE / DEFINITE_ISSUE
- Line 2: a short English explanation

Do not output Markdown, JSON, or extra text.

Text to evaluate:
{text}

E.6 Manual Inspection

To validate the reliability of the automated pipeline, two domain experts independently evaluated 200 randomly sampled segments. They confirmed that 98.0% of retained segments were medically correct, coherent, and contextually complete, deemed suitable for instruction generation. These results indicate that the OCR pipeline provides robust and reliable textual foundations for building Hindi medical instruction data. The experts are hired from local translation company. The experts are paid 500 currency units per hour, which is notably higher than the local translators' hourly wage. In total, 2,000 currency units are spent at this inspection stage.

E.7 Annotator Instructions for Manual Inspection

This manual inspection is conducted as a targeted quality audit of the automated OCR and data processing pipeline, rather than as full human correction or annotation. The goal is to verify whether the retained text segments are suitable for downstream instruction generation in Hindi medical reasoning tasks.

Annotator background. All annotations are performed independently by two domain experts with formal training in medicine or biomedical sciences and demonstrated familiarity with Hindi medical terminology.

Sampling protocol. Each annotator is provided with the same randomly sampled set of text segments drawn from the retained outputs of the OCR pipeline. Samples are selected across different source documents and medical topics. No information about downstream model training or benchmark usage is disclosed to the annotators.

Evaluation criteria. For each text segment, annotators assess the following criteria:

- **Medical correctness.** The medical content should be factually correct according to standard textbooks or officially examined curricula. Minor stylistic issues are acceptable as long as they do not alter medical meaning.
- **Semantic completeness.** The segment should be self-contained and convey a complete medical concept or explanation without requiring missing context from adjacent passages.
- **Coherence and readability.** The text should be logically coherent and readable in Hindi, without severe OCR artifacts that hinder understanding.

- **OCR alignment fidelity.** The textual content should faithfully reflect the original source material. Minor character-level OCR errors are acceptable if they do not affect semantic interpretation.

Decision rule. Annotators assign a binary judgment to each segment:

- **Accept**: The segment satisfies all evaluation criteria and is suitable for downstream instruction generation.
- **Reject**: The segment contains medical errors, missing context, or OCR artifacts that compromise semantic fidelity.

Disagreement handling. Annotations are conducted independently. In cases of disagreement, the segment is conservatively marked as rejected. No post-hoc adjudication or correction is performed.

Scope and limitations. This inspection focuses on verifying the usability and reliability of OCR outputs as textual foundations. It does not assess clinical safety, therapeutic efficacy, or real-world medical applicability of the content.

F Instruction Pool

F.1 Method

The instruction pool comprises high-quality medical questions from several public datasets: HealthBench (Arora et al., 2025), HealthSearchQA (Singhal et al., 2022), MedMCQA (Pal et al., 2022), MedReason (Wu et al., 2025a), and ReasonMed (Sun et al., 2025). To integrate these heterogeneous sources and improve generality, we employ a transformation pipeline in which GPT-4o abstracts the core intent shared by similar questions into a single standardized instruction.

Central to our approach is a five-type question taxonomy (Ely et al., 1999) developed from an analysis of thousands of questions posed by practicing physicians. This taxonomy aligns with structured representations common in medical question answering, such as the PICO framework, and principles of semantic unification (Demner-Fushman and Lin, 2007; Abacha and Demner-Fushman, 2019). It thus reinforces the distinctions among diagnostic, therapeutic, etiological, prognostic, and conceptual information needs.

We define five categories as follows:

- **Diagnosis.** Questions concerned with identifying a disease or condition from observed findings (e.g., "Could this patient have [disease]?"; "Is test [procedure] indicated in [context]?").
- **Treatment.** Questions about clinical management, therapies, and medications (e.g., "What is the drug of choice for [disease]?"; "How should I treat [disease]?"; "What is the dose of [medication]?").
- **Etiology.** Questions about causes, origins, or risk factors underlying a symptom or disease (e.g., "What causes [symptom]?"; "What factor leads to [finding]?").
- **Prognosis.** Questions concerning expected disease course, severity, or complications.
- **Medical Knowledge.** General conceptual/definitional questions foundational to medical understanding (e.g., "What is [disease]?").

In addition, we generalize instructions using placeholders grounded in the top-level hierarchies of SNOMED CT (SNOMED International, 2024), the world's most comprehensive clinical terminology system. Each placeholder corresponds to a high-level clinical concept, making the abstraction scheme clear and robust. Table 7 lists the placeholders and definitions.

Prompt: Forming Instruction

You are a professional medical AI data scientist, proficient in multilingual processing. Your task is to analyze medical records and accurately extract the core medical consultation intent. You need to generate standardized outputs in three formats based on the complexity of each case and perform semantic deduplication on instructions sharing the same core intent.

Processing Steps and Requirements

1. **Information Analysis and Intent Extraction**
   - Carefully read all information points provided by the user, such as symptoms, history, test indicators, doctor's advice, demographics, and location.
   - Identify and distill the user's core medical consultation intent.
2. **Standardized Output Generation**
   - Based on question complexity, generate `Instruction` in one of the following formats:
     - **A. Conversation Summary Format** (multiple user utterances):
       Use a narrative summary: "In the conversation, the user provided [Information 1], [Information 2], ..., ultimately asking [Core Question]."

Term	Definition
Symptom	A physical or mental sign that may indicate a disease.
Disease	A specific illness or disorder with distinct symptoms and causes.
Cause	The reason why a disease happens, or the source of a health problem.
Medication	A substance (usually a drug) used to treat or prevent illness.
Procedure	A medical intervention or operation, including diagnostic and therapeutic care.
Context	Circumstances or background information relevant to the patient situation.
Age	An attribute of a person.
Gender	An attribute of a person.
Department	A specific medical department or specialty.
Treatment	A therapeutic intervention intended to cure, manage, or alleviate a disease.
Doctor	A qualified healthcare professional providing diagnosis, treatment, or guidance.

Table 7: Placeholder definitions used for instruction generalization.

  - **B. QA Standard Format** (one user question):
    Use a direct professional question with placeholders (e.g., [symptom], [disease], [procedure]).
  - **C. MCQ Format** (requires choice):
    Ask in second person and list options (A/B/C/...), using placeholders (e.g., [option]).

3. **Intent Classification**
   - Assign one primary label: Diagnosis, Treatment, Etiology, Prognosis, or Medical Knowledge.
4. **Semantic Deduplication**
   - Identify instructions with identical core intent and retain only the clearest, most standard one for each group.
5. **Output Format**
   - Return a Markdown table with columns: Instruction, Category, Form.

Final Output: Return the Markdown table only.

F.2 Questionnaire

We conduct a survey to understand how people formulate health-related questions when seeking medical advice from doctors or AI assistants. The questionnaire is designed to elicit diverse symptom descriptions, diagnostic inquiries, and treatment-related questions across multiple scenarios. The survey is administered in both English and Hindi to accommodate a diverse respondent population. Below we present the full questionnaire.

Questionnaire

We are a research team developing AI to better answer health-related questions. Please share a few questions you might ask a doctor or an AI assistant.

हम एक शोध टीम हैं जो स्वास्थ्य संबंधी प्रश्नों का बेहतर जवाब देने के लिए AI (कृत्रिम बुद्धिमत्ता) विकसित कर रहे हैं। कृपया कोई ऐसे प्रश्न साझा करें जो आप एक डॉक्टर या AI (सहायक से पूछ सकते हैं।

1. Imagine you wake up feeling unwell. You have sudden blurred vision in your left eye. You decide to message a doctor or a reliable AI health assistant for advice.

   How would you describe your symptoms to a doctor or AI to get help with diagnosis?

   कल्पना कीजिए कि आप बीमार महसूस करते हुए उठते हैं। आपकी बाईं आंख में अचानक धुंधला दिखाई देने लगा है। आप सलाह के लिए डॉक्टर या एक विश्वसनीय AI (स्वास्थ्य सहायक) को मैसेज करने का फैसला करते हैं।

   निदान में मदद पाने के लिए आप डॉक्टर या AI (को अपने लक्षण कैसे बताएंगे?)

2. After your recent physical examination at your company/school, you received your report, which showed that several indicators (e.g., white blood cells, blood pressure) were beyond the normal range. You are worried. If you want to ask the doctor/AI why your results are abnormal and what might be causing these changes, what would you say?

   आपकी कंपनी/स्कूल में हाल की शारीरिक जांच के बाद, आपको अपनी रिपोर्ट मिली, जिसमें दिखाया गया कि आपके कई मानक (जैसे सफेद रक्त कोशिकाएँ, रक्तचाप) सामान्य सीमा से बाहर थे। आप चिंतित हैं। यदि आप डॉक्टर या AI (से पूछना चाहते हैं कि "मेरे परिणाम असामान्य क्यों हैं और इन बदलावों का कारण क्या हो सकता है?" तो आप क्या पूछेंगे?)

3. If you unfortunately contract chikungunya fever (a new virus), which causes persistent fever, and you are very worried that you may not live long or have serious long-term effects, what would you ask the doctor?

   यदि आप दुर्भाग्य से चिकनगुनिया बुखार (एक नया वायरस) की चपेट में आ जाते हैं, जिससे लगातार बुखार आता है, और आप बहुत चिंतित हैं कि कहीं आपकी उम्र कम न हो जाए या गंभीर दीर्घकालिक प्रभाव न हों, तो आप डॉक्टर से क्या पूछेंगे?

4. If you suspect you have a necrotizing skin infection and cannot go to the hospital, what questions would you ask an AI assistant for help?

   यदि आपको संदेह है कि आपको नेक्रोटाइज़िंग स्किन इन्फेक्शन (त्वचा का गलना/मरना) है और आप अस्पताल नहीं जा सकते, तो मदद के लिए आप AI (सहायक से कौन से प्रश्न पूछेंगे?)

5. Suppose you are traveling in the Amazon region and have been experiencing watery diarrhea and cramps for two weeks. You are eager to find a way

to relieve the symptoms. How would you ask a doctor for treatment methods? (Write down questions.)

कल्पना कीजिए कि आप अमेज़न क्षेत्र में यात्रा कर रहे हैं और आपको दो सप्ताह से पानी जैसा दस्त और मरोड़ हो रहे हैं। आप लक्षणों से राहत पाने का तरीका ढूंढने के लिए बेताब हैं। आप डॉक्टर से उपचार के तरीके कैसे पूछेंगे? (प्रश्न लिखें)

We ultimately collect 103 valid responses, from which 37 new instructions are extracted and categorized into five types, following generic instruction frameworks derived from prior seminal work (Abacha and Demner-Fushman, 2019; Demner-Fushman and Lin, 2007). Statistics are reported in Section F.3.

F.3 Statistics and Manual Inspection

Category	Former Work	Questionnaire	Total
Diagnosis	24	7	31
Etiology	22	7	29
Prognosis	26	6	32
Treatment	26	6	32
Med. Know.	26	7	33
Total	124	33	157

Table 8: Instruction counts by category and source.

Table 8 provides the instruction counts by category and source. The instruction pool is further reviewed manually with four focuses:

1. **Deduplication.** Instructions are grouped by category and semantically similar items are consolidated. For duplicates, only the clearest and most comprehensive instruction is retained. After deduplication, the remaining instructions account for 82% of the original pool.

2. **Complexity and clarity assessment.** Two experts independently assess each instruction for clarity, which requires the intent to be unambiguous, and executability, defined as the model's ability to reasonably produce a coherent and accurate response. Only instructions approved by all experts are retained, yielding a pass rate of 61%.

F.4 Annotator Instructions for Instruction Pool Review

This manual inspection is conducted to assess the quality and usability of the instruction pool prior to model training. The goal is to ensure that retained instructions are non-duplicative, clear, and executable for medical reasoning tasks, rather than to perform content rewriting or optimization. Two experts are hired from professional medical institute. The experts are paid 500 currency units per hour, which is notably higher than the local doctor's hourly wage. In total, 1,000 currency units are spent at this inspection stage.

Annotator background. The review is performed independently by two domain experts with backgrounds in medicine or biomedical sciences and proficiency in Hindi medical language. Annotators are not informed of downstream model performance or experimental outcomes.

Review procedure. Each instruction is reviewed within its assigned category. Annotators follow the criteria below and provide binary judgments as specified.

Deduplication criterion. Annotators identify semantically duplicate or near-duplicate instructions within the same category. Two instructions are considered duplicates if they ask the same medical question or require substantially overlapping reasoning, even if surface wording differs. For duplicate groups, annotators retain a single instruction that is the clearest, most complete, and least ambiguous. All other duplicates are removed.

Clarity and executability assessment. Each remaining instruction is evaluated independently by both annotators according to the following criteria:

- **Clarity.** The instruction should have a well-defined intent, with no ambiguous references, underspecified conditions, or unclear task requirements.

- **Executability.** The instruction should be reasonably answerable by a language model using medical reasoning, without requiring external context, private information, or real-world clinical decision-making.

Decision rule. Annotators assign a binary decision to each instruction:

- **Accept**: The instruction satisfies both clarity and executability criteria.

- **Reject**: The instruction is ambiguous, underspecified, redundant, or not reasonably answerable.

Only instructions accepted by both annotators are retained. In cases of disagreement, the instruction is conservatively rejected.

Scope and limitations. This review assesses instruction quality for benchmarking and training purposes only. It does not evaluate clinical correctness, therapeutic safety, or real-world medical applicability of the instructions.

G HiMed-Trad Generation

G.1 Method

We construct HiMed-Trad via a staged pipeline that converts generation-ready merged segments into a structured, instruction-ready corpus.

Stage I: subject and type labeling. For each passage, we assign one or more intent categories from the five types (Diagnosis, Treatment, Etiology, Prognosis, Medical Knowledge), and one or more instance formats from MCQ, QA, Dialogue. Labeling is performed using GPT-4o. For multi-subject or multi-type passages, we expand them into atomic entries by taking the Cartesian product of category × format, so that each final entry has exactly one category and one format.

Stage II: instruction generation with controlled templates. From labeled entries, we randomly select an instruction template using a predefined template pool and optional few-shot exemplars used only as style references.

Given (*text*, *subject*, *type*, *template*, *few-shot*), an LLM is prompted to generate instruction instances.

We enforce structured outputs and apply automatic validators to reject malformed generations, such as missing MCQ options, schema violations, or empty rationales when required.

Prompt:Template-Guided Instruction Generation

Role

You are an expert question writer for traditional medicine.

Your task is to generate high-quality instruction-style data (QA, MCQ, or Dialogue) grounded in traditional medicine texts, using predefined templates and optional few-shot exemplars. Ignore all the possible names or personal identifiers.

Input

For each generation instance, you are given:

- text: source paragraph
- subject: medical subject
- type: one of ["MCQ","QA","Dialogue"]
- template: a compatible instruction template
- few-shot: optional few-shot exemplars

Templates are NOT selected by another model. Instead, for each (subject, type) pair, a compatible template is randomly sampled to reduce selection bias and increase diversity.

Overall Generation Procedure

1. Select the given template and follow its structure strictly.
2. Generate three difficulty levels: EASY, MEDIUM, and HARD.
3. Ensure all outputs are self-contained, grounded in the source text, and consistent with traditional medicine.
4. Produce structured outputs that satisfy the required schema.

Malformed generations (e.g., missing MCQ options, schema violations, or empty rationales when required) will be automatically rejected.

Few-Shot Style References

The following few-shot blocks are provided as style references only (do NOT copy sentences verbatim):

- MCQ_FEW_SHOT {...}
- QA_FEW_SHOT {...}
- DIALOGUE_FEW_SHOT {...}

Difficulty Specification

For each instruction type, you MUST generate all three difficulty levels in the following fixed order: EASY, MEDIUM, HARD.

- **EASY**: single fact, no scenario, no reasoning required.
- **MEDIUM**: two-step reasoning, combining about two points; brief scenario optional.
- **HARD**: three to five-step reasoning; a realistic scenario is REQUIRED and must be grounded in traditional medicine.

Type-Specific Instructions

(A) MCQ

- Generate THREE multiple-choice questions (EASY, MEDIUM, HARD).
- Each question MUST have exactly five options: A., B., C., D., E.
- There must be ONE correct option.
- Provide:
  - the correct option letter
  - a clear reasoning
- MCQ is **NOT** a dialogue:

- Do NOT use "User:" or "Assistant:"
- Do NOT create multi-turn conversations

- Each item must be a single, self-contained question.
- Avoid any mention of text, paragraph, or source.

(B) QA

- Generate THREE short-answer questions (EASY, MEDIUM, HARD).
- Each item MUST consist of:
  - one direct question
  - one concise correct answer
  - reasoning
- QA is **NOT** a dialogue:
  - NO "User:" or "Assistant:" labels
  - NO multi-turn format
- Avoid any mention of text, paragraph, or source.

(C) Dialogue

- Generate THREE multi-turn dialogues (EASY, MEDIUM, HARD).
- Each dialogue MUST:
  - include two or more turns
  - contain explicit User: and Assistant: labels
  - end with a question from the User
- Provide:
  - the dialogue (up to the final user question)
  - the correct answer
  - reasoning
- The scenario must clearly belong to traditional medicine.

Output Format

Return ONLY the following structure (no explanations, no markdown):

<EASY><q>...</q><a>...<cot>...</cot>
<MEDIUM> ...
<HARD> ...

- MCQ: <q> must include options A. B. C. D. E.
- QA: must be a single direct question (no dialogue)
- Dialogue: must include multi-turn User:/Assistant: format

If the task is absolutely impossible, return: <FAIL>.

Stage III: automatic quality rating and filtering. In the third stage of data construction, we apply an automatic quality inspection step to assess and filter generated instruction instances.

Each generated sample, consisting of a source paragraph, question, answer, and chain-of-thought rationale, is evaluated by a dedicated data quality rater prompted as a medical expert in traditional medicine.

The rater assigns continuous scores in four dimensions: contextual grounding, medical correctness, reasoning clarity, and Hindi language quality.

Based on a small pilot audit by two domain experts, we set the acceptance threshold to an average score of 0.725 (mean over the four dimensions), and retain only instances above this threshold for downstream use.

Here we provide the prompt used in this stage.

Prompt: Data Quality Rater

You are an expert data quality rater for a medical Q-A-CoT dataset in traditional medicine.
You will be given one JSON object with fields:

- text: source paragraph (in Hindi)
- subject: one of ["diagnosis","etiology","medical knowledge","prognosis","treatment"]
- type: one of ["MCQ","QA","Dialogue"]
- question: generated question
- answer: generated answer
- cot: chain-of-thought reasoning

You must evaluate the question/answer/cot on four dimensions , each scored from 0.00 to 1.00:
1. **grounded_in_context** :

- Score 1.00 if all information in question/answer/cot is directly derivable from the source text
- Score lower if there is any hallucinated information or external knowledge not present in the text
- Score 0.00 if the content is completely unrelated to the source text

2. **medical_correctness** :

- Score 1.00 if the medical information is correct according to the source text
- Score lower if there are minor inaccuracies or misinterpretations

- Score 0.00 if the medical information is incorrect or contradicts the source text

3. **reasoning_clarity**:

- Score 1.00 if the reasoning steps are logical, clear, and well-structured
- Score lower if reasoning is somewhat unclear or has minor logical gaps
- Score 0.00 if reasoning is illogical, confusing, or missing

4. **language_quality**:

- Score 1.00 if the Hindi language is fluent, natural, and grammatically correct
- Score lower if there are minor grammatical errors or awkward phrasing
- Score 0.00 if the language is severely broken or incomprehensible

Return ONLY a JSON object with 4 float scores in [0.00, 1.00]:

```
{
  "grounded_in_context": 0.00,
  "medical_correctness": 0.00,
  "reasoning_clarity": 0.00,
  "language_quality": 0.00
}
```

Do NOT add any explanations or extra keys. Return only the JSON object.

G.2 Statistics

Table 9 summarizes the corpus scale after each major stage. We report two types of statistics here: document- and page-level statistics, which reflect OCR coverage, and entry-level statistics, which reflect downstream usability.

Item	Count
Source documents	43
Total scanned pages	12,450
OCR-good MMD files	43
Segmented passages	21,932
QA instances	93,859
MCQ instances	102,494
Dialogue instances	96,314

Table 9: Summary statistics of HiMed-Trad.

G.3 Manual Inspection

Manual inspection is conducted as a targeted quality audit rather than full human correction, with the goal of verifying that the corpus is usable while preserving scalability and reproducibility. 200 samples are randomly selected and inspected. Two experts are hired from professional medical institute. The experts are paid 500 currency units per hour, which is notably higher than the local doctor's hourly wage. In total, 3,000 currency units are spent at this inspection stage. Manual inspection is conducted on samples that have already passed automatic filtering, with the goal of validating the reliability of the quality rater rather than re-filtering the data.

Sampling protocol. We randomly sample entries across subjects, types, and instruction formats (QA, MCQ, and dialogue). To avoid bias toward easy cases, samples are stratified by passage length and script composition, ensuring coverage of both short factual passages and longer, reasoning-intensive contexts, as well as pure Hindi and code-mixed inputs.

Inspection criteria. Each sampled instance is independently reviewed by domain-aware annotators with a medical background. The inspection focuses on three aspects: (i) *semantic fidelity*, i.e., whether the question, answer, and rationale are faithful to the source passage without introducing unsupported content; (ii) *medical soundness*, i.e., whether the medical statements are correct and internally consistent; and (iii) *linguistic naturalness*, i.e., whether the Hindi expressions are fluent, coherent, and appropriate for medical communication. Importantly, annotators do not rewrite or correct instances, but only judge whether they are acceptable for inclusion.

Acceptance and outcomes. An instance is marked as *passed* if all inspected aspects meet the acceptance standard. Across all inspected samples drawn from instances that had already passed the automatic quality threshold, 99.5% are deemed acceptable without manual edits, indicating strong agreement between automated filtering and human judgment. The few rejected cases are primarily attributed to subtle linguistic infelicities or minor mismatches between rationales and source passages, rather than factual medical errors.

G.4 Annotator Instructions for Post-filter Manual Inspection

This manual inspection serves as a validation audit of the automatic quality rating and filtering stage. Annotators are instructed to assess whether instances that passed automatic filtering are acceptable for inclusion as-is. No rewriting, correction, or re-ranking of instances is permitted.

Annotator background. All inspections are conducted independently by domain-aware annotators with a medical background and proficiency in Hindi medical language. Annotators are blinded to the automatic quality scores and thresholds.

Inspection setup. For each sampled instance, annotators are provided with: (i) the source passage, (ii) the generated question, (iii) the generated answer, and (iv) the chain-of-thought rationale. Judgments must be based solely on these materials.

Evaluation criteria. Annotators evaluate each instance along the following three aspects, which correspond to the dimensions used by the automatic quality rater:

- **Semantic fidelity.** The question, answer, and rationale must be faithful to the source passage. No hallucinated facts, unsupported claims, or content external to the passage should be introduced.
- **Medical soundness.** All medical statements must be correct, internally consistent, and aligned with standard medical knowledge or officially examined curricula.
- **Linguistic naturalness.** The Hindi language should be fluent, coherent, and appropriate for medical communication. Minor stylistic imperfections are acceptable if they do not affect comprehension.

Decision rule. Annotators assign a binary judgment to each instance:

- **Pass**: All evaluation criteria are satisfied, and the instance is acceptable for inclusion without modification.
- **Fail**: One or more criteria are violated.

Annotators must not edit the content or provide corrective feedback. In cases of disagreement, the instance is conservatively marked as failed.

Scope and limitations. This inspection evaluates the reliability of automated quality filtering for dataset construction. It does not assess clinical safety, real-world applicability, or downstream model behavior.

H HiMed-Trad Taxonomy

To provide a balanced and authoritative foundation, we draw on both peer-reviewed scholarship and official documents across the major Indian systems of medicine. For Ayurveda, we reference the historical and conceptual overview (Jaiswal and Williams, 2017) together with the Ashtāṅga Ayurveda materials from the Directorate of AYUSH (Directorate of AYUSH). In the case of Yoga, we rely on evidence from clinical and quality-of-life studies (Woodyard, 2011), as well as the official *Common Yoga Protocol* periodically issued by the Ministry of Ayush (of Ayush and MDNIY, 2025). Training standards are contextualized with the ongoing WHO benchmarks for Yoga (Organization, 2022). For Naturopathy, we include parliamentary records that list its principal modalities (Ministry of Ayush, 2023), together with the Central Council for Research in Yoga and Naturopathy, alongside WHO's 2010 training benchmarks (for Research in Yoga & Naturopathy , CCRYN; World Health Organization, 2010a). The Sowa-Rigpa tradition is represented by the National Institute of Sowa-Rigpa, Ministry of Ayush portals, and pharmacopoeial initiatives by PCIM&H (of Sowa-Rigpa , NISR; of Ayush, a; for Research in Ayurvedic Sciences , CCRAS; for Indian Medicine & Homoeopathy). For Homoeopathy, we cite both regulatory frameworks such as the *National Commission for Homoeopathy Act* (of India, 2020) and institutional sources including CCRH, NIH Kolkata, and the Homoeopathic Pharmacopoeia of India (for Research in Homoeopathy , CCRH; National Institute of Homoeopathy , NIH; PCIM&H, b). Similarly, Unani medicine is documented through AYUSH portals, CCRUM, NIUM, and the official pharmacopoeia and formulary volumes (of Ayush, b; for Research in Unani Medicine , CCRUM; National Institute of Unani Medicine , NIUM; PCIM&H, d,a). Finally, Siddha medicine is covered via the Central Council for Research in Siddha, the National Institute of Siddha, and the Siddha Pharmacopoeia and Formulary of India (for Research in Siddha , CCRS; National Institute of Siddha , NIS; PCIM&H, c). Our taxonomy-level comprehensiveness serves two purposes. It avoids bias toward a single tradition and supports fine-grained error analysis. For instance, whether an LLM struggles more with Ayurveda, Unani, or Siddha.

I Translation Process

We here introduce the detailed process of translation, which is critical for the construction of HiMed-West Corpus and HiMed-West Bench.

I.1 Lexicon Table

To guarantee terminological precision and domain adaptability, we constructed a high-fidelity English–Hindi medical lexicon. The construction process consists of utilizing authoritative glossaries and augmenting them with data-driven extractions.

I.1.1 Authoritative Medical Glossaries

The foundational stratum of our lexicon was curated from three distinct streams, comprising a total of six authoritative volumes.

General Medical Dictionary. We include a broad-coverage English–Hindi medical dictionary, *Dictionary of Medicine* (Gupta and Kapoor, 1955). The dictionary is organized alphabetically by English headwords and provides Hindi equivalents, covering core topics such as anatomy, physiology, diseases, symptoms, examinations, and treatments.

CSTT Specialized Glossaries. To align with official Hindi terminology standards, we incorporate four domain glossaries released by the *Commission for Scientific and Technical Terminology (CSTT)*, Government of India (CSTT, 2009, 2018a, 2010, 2018b). These volumes provide standardized, authority-sanctioned Hindi medical terms and serve as our highest-priority reference during lexicon consolidation.

Educational Science Glossary. To complement clinical resources with foundational biochemical and pharmaceutical nomenclature, we additionally include the *Chemistry Glossary (High School Level)* published by the *New York State Education Department (NYSED)* (The State Education Department and The University of the State of New York, 2018).

I.1.2 Digitization

Except for the general dictionary, the remaining five volumes were obtained as scanned PDFs and digitized. We apply *DeepSeek-OCR* to each page and export the raw results as unprocessed Multi-Markdown (MMD) files. We intentionally keep the OCR text minimally normalized to preserve layout cues and OCR artifacts.

I.1.3 Corpus-based Augmentation (MedReason)

To complement static glossaries and match the vocabulary distribution of *HiMed-West*, we augment the lexicon by extracting medical entities directly from the *MedReason* corpus, targeting long-tail clinical terms that are absent from authoritative resources.

I.2 Manual Correction and Translation.

Two professional Hindi–English bilingual annotators perform line-by-line correction of OCR outputs and translation of extracted outputs from corpus-based augmentation. The correction focuses on mixed-script errors and confusions among visually similar Devanagari characters, applied directly to the text to recover faithful content and reduce downstream noise.

Final Lexicon Consolidation. We merge three sources: (i) digitized authoritative glossaries and dictionary (Section I.1) and (ii) manual-translated corpus-derived terms. Conflicts are resolved with priority to official CSTT terminology. The final consolidated lexicon contains **45,902** English–Hindi term pairs. Authoritative glossaries contribute **62%** of entries, while corpus augmentation contributes the other **38%**.

I.3 Annotator Instructions for Lexicon Construction and Correction

This annotation task focuses on constructing and validating a high-fidelity English–Hindi medical lexicon. Annotators are instructed to perform deterministic correction and translation of medical terms, rather than creative paraphrasing or free-form rewriting. The annotators are experts hired from local translation company. The experts are paid 400 currency units per hour, which is notably higher than the local translators' hourly wage. In total, 30,000 currency units are spent at this inspection stage.

Annotator background. All annotations are performed by professional Hindi–English bilingual annotators with prior experience in medical or biomedical translation. Annotators are familiar with Devanagari script conventions and common OCR artifacts.

Input materials. For each entry, annotators are provided with one of the following: (i) OCR-extracted glossary entries from authoritative medical volumes, or (ii) English medical terms extracted from the MedReason corpus for lexicon augmentation.

Correction and translation guidelines. Annotators follow the rules below:

- **OCR correction.** Correct mixed-script errors, broken characters, and confusions among visually similar Devanagari glyphs. Do not modernize spelling or alter stylistic variants unless required to restore correct meaning.
- **Terminological translation.** Translate each English medical term into its most appropriate Hindi equivalent. When an official CSTT term exists, it must be used preferentially.
- **One-to-one mapping.** Each lexicon entry must correspond to a single English term and a single Hindi rendering. Do not introduce explanations, definitions, or multiple alternatives.
- **Conservativeness.** If a term cannot be translated confidently without external context, annotators must flag it for exclusion rather than speculate.

Conflict resolution. In cases of disagreement between sources, annotators defer to CSTT terminology. Ambiguous or low-confidence entries are excluded from the final lexicon.

Scope and limitations. This process aims to ensure terminological precision and consistency for translation and benchmarking. It does not evaluate clinical validity or recommend medical usage.

I.4 Method

We propose a high-throughput, lexicon-guided translation pipeline that targets medical-domain fidelity while strictly preserving the structural constraints of benchmark datasets. The system combines rule-based terminology injection with neural machine translation (NMT), and is engineered for efficient large-scale processing.

I.4.1 Lexicon-Guided Hybrid Architecture

Our translation engine adopts a three-stage hybrid design: *lexicon injection*, *neural translation*, and *post-editing*. To enforce terminological consistency, we perform dictionary-based injection prior to neural inference. Specifically, using the lexicon in Section I.1, we compile medical terms into regex patterns and deterministically replace matched English terms (with word-boundary constraints) with their standardized Hindi equivalents before translation. This yields a code-mixed intermediate that constrains the NMT model to preserve domain-specific terminology.

Neural translation is performed with **NLLB-200-3.3B** (Team et al., 2022). For scalable processing, we segment text with *BlingFire* and apply length-aware dynamic batching to reduce padding overhead. Finally, we optionally append the original English term in parentheses as *Hindi Term (English Term)* to facilitate bilingual inspection and medical evaluation.

I.4.2 Adaptive Dataset Processing

To preserve evaluation formats across heterogeneous benchmarks, we adopt a format-adaptive pipeline that routes **structured MCQs** and **unstructured inputs** through different processing paths.

Structured inputs (MCQs). For MCQs, a structure-aware parser decomposes each instance into the instruction, question stem, and labeled options using regular expressions. Only the textual spans are translated, while option labels and structural markers are kept intact, preventing label corruption or reordering.

Unstructured inputs. For free-form prompts, we translate the input directly with the batch engine in Section I.4 to preserve discourse flow without unnecessary decomposition.

Fault-tolerant execution. All jobs run under a unified fault-tolerant executor with checkpointing and schema validation, ensuring resumable processing and that the translated outputs strictly conform to the original JSON format for downstream evaluation.

I.4.3 Manual Audit: NLLB vs. GPT-4o Translation Outputs

Comparison and implications. Using the same audit rubric, the NLLB-based pipeline requires

fewer manual fixes than the GPT-4o-based pipeline and exhibits no translation-induced hallucinations. In practice, most remaining NLLB issues are localized and can often be mitigated deterministically via lexicon injection. By contrast, GPT-4o exhibits more frequent format violations and hallucinations under the same audit rubric. Overall, NLLB provides more stable, format-preserving translations, making it preferable for large-scale benchmark construction. This audit is conducted by two experts hired from local translation company, with 500 currency units hourly wage, higher than average level. In total, 2000 currency units are spent at this inspection stages. In another two experts' manual inspection consisting of 200 samples, our pipeline's results achieve an acceptance rate of 97.0%, indicating the usability and reliability of our method and practice. In total, 4,000 currency units are spent at this inspection stages.

Metric	GPT-4o	Ours
Need Fix	23	2
Hallucination	19	0

Table 10: Manual audit results of 50 samples under a unified rubric.

I.4.4 Annotator Instructions for Final Translation Quality Audit

This manual audit evaluates the quality and reliability of translated benchmark instances produced by the lexicon-guided translation pipeline. Annotators are instructed to assess usability and fidelity of translations as-is, rather than to edit or improve them.

Annotator background. Audits are conducted by bilingual reviewers with medical knowledge and proficiency in both English and Hindi. Annotators are blinded to the translation method used for each instance.

Audit setup. For each sampled instance, annotators are provided with: (i) the original English input, (ii) the translated Hindi output, and (iii) the original task format (e.g., MCQ structure, option labels).

Evaluation criteria. Annotators assess each instance according to the following criteria:

- **Terminological fidelity.** Medical terms should be translated consistently with the constructed lexicon. No critical terminology should be mistranslated, omitted, or hallucinated.
- **Semantic equivalence.** The Hindi translation should preserve the meaning of the original English input without introducing unsupported information or altering intent.
- **Format preservation.** Structural elements such as MCQ options, labels, and JSON schemas must remain intact. Any format violation is grounds for rejection.
- **Linguistic adequacy.** The Hindi text should be fluent and comprehensible. Minor stylistic imperfections are acceptable if meaning and structure are preserved.

Decision rule. Annotators assign a binary judgment:

- **Accept**: The translation satisfies all criteria and is usable for benchmarking.
- **Reject**: One or more criteria are violated.

Annotators do not modify translations. No adjudication or correction is performed. In cases of disagreement, the instance is conservatively rejected.

Scope and limitations. This audit assesses translation reliability for dataset construction and evaluation only. It does not assess downstream model performance or clinical applicability.

I.5 Software Packages

Libraries (versions). `torch` 2.9.0+cu128; `transformers` 4.57.3; `pandas` 2.3.3; `openpyxl` 3.1.5; `sentencepiece` 0.2.0; `blingfire` 0.1.8.

J Decontamination

To ensure that benchmark performance reflects genuine generalization rather than data memorization, we apply decontamination at multiple levels.

First, **source-level separation** is enforced by construction: all books and official documents used for corpus generation are strictly disjoint from examination papers used for benchmarking, and no source appears in both splits.

Second, at the **passage level**, all data derived from the same OCR passage (as identified by a shared source document and page index) are assigned exclusively to either the training corpus or the benchmark, preventing any partial overlap across splits.

Third, for **translated benchmarks**, both the original questions and their translated or paraphrased variants are explicitly excluded from training data construction, ensuring that benchmark items are not indirectly exposed during training.

K Training Details

K.1 Training Data Overview Across Stages

Hyperparameter selection. Unless otherwise noted, all hyperparameters are determined using the training datasets described in this section, and we do not tune hyperparameters on any evaluation datasets.

Data mixing. Across all stages, we shuffle the training instances globally, mixing all data sources, task types, and languages into a single randomized stream for optimization. Table 11 summarizes the training data composition for each stage.

Dataset	LA	RC	DSR-RL
HiMed-West	46.5K	52K	10K
HiMed-Trad	102K	102K	10K
MedMCQA	46.5K	39K	
Huatuo-o1 Corpus	–	9K	–
GSM8K	–	8K	–
MedReason	–	–	5K
DailyDialog	11K	–	–
Persona-Chat	17K	–	–

Table 11: Training data composition.

In-domain pool and stage-wise split. For our Hindi traditional medical data, we use HiMed-Trad throughout the three-stage pipeline via uniform random sampling: 101.6k instances are used in LA, 101.6k instances are used in RC, and 10k instances are reserved for DSR-RL. This stage-wise allocation keeps the in-domain distribution consistent across stages while matching the intended training schedule.

Language Adaptation (LA). For LA, we use 102k in-domain instances from HiMed-Trad. We additionally incorporate MedMCQA with 46.5k English instances and 46.5k Hindi instances, all used with chain-of-thought rationales removed. We further include DailyDialog with 5.5k English instances and 5.5k Hindi instances, as well as Persona-Chat with 8.9k English instances and 8.9k Hindi instances.

Reasoning Cold-start (RC). RC is trained in two phases. Phase 1 runs for one epoch on 183.4k instances, including 101.6k in-domain examples from HiMed-Trad and 81.8k MedMCQA examples with restored chain-of-thought rationales, comprising 40.9k English and 40.9k Hindi instances. Phase 2 uses 101.6k HiMed-Trad instances, together with Huatuo-o1 Corpus comprising 9.9k English instances and 9.0k Hindi instances after filtering samples whose chain-of-thought plus answer exceed 1536 tokens, and MedMCQA with 79.1k chain-of-thought annotated instances, including 39.5k English and 39.6k Hindi examples. In addition to these medical datasets, we include 7.5k GSM8K instances, split into 3.7k English and 3.7k Hindi examples.

DSR-RL. For DSR-RL, we use 10k in-domain instances sampled from HiMed-Trad. We additionally include MedReason with 10k Hindi instances and 5k English instances.

K.2 R_1 Reward Model Training Details

Base Model and Training Data We adopt Llama-3.2-3B-Instruct as the base model and construct the supervised fine-tuning corpus from two primary data sources. We extract 10k English medical reasoning instances from the Huatuo-o1 (Chen et al., 2025) dataset using a fixed random seed, and additionally sample 10,000 further instances that are translated into Hindi through our customized translation pipeline. We further incorporate the MedReason (Wu et al., 2025a) dataset, sampling 7.5k English instances and translating another 7.5k instances into Hindi in the same manner. To build the reward supervision labels, we use the unfine-tuned Llama-3.1-8B-Instruct model to generate Chain-of-Thought outputs for each problem and employ GPT-5 to compare these outputs with the corresponding reference CoT, producing ground-truth annotations on whether the two reasoning traces are semantically equivalent. The resulting bilingual dataset is used to train our reward model for CoT semantic equivalence evaluation. For the annotation of 300 randomly sampled model responses, two experts are hired from professional medical institute. The experts are paid 500 currency units per hour, which is notably higher than the local doctor's hourly wage. In total, 3,000 currency units are spent at this inspection stage. Each expert was independently presented with 300 model-generated responses, paired with the corresponding question and ground-truth answer. The task was to assess whether the model's final answer was factually and medically correct given the

question context. Experts were instructed to focus on the correctness of the conclusion rather than the surface fluency, verbosity, or stylistic quality of the reasoning. Intermediate reasoning steps were considered only insofar as they affected the validity of the final answer. Annotations were binary, labeled as True (the answer is correct) or False (the answer is incorrect or unsupported). In cases where the response involved minor wording variations, paraphrases, or alternative but medically equivalent formulations, experts were instructed to mark the answer as True as long as the medical conclusion was sound. If the case was inherently ambiguous, under-specified, or dependent on unstated assumptions that prevented a definitive judgment, experts were allowed to make their best professional judgment based on standard medical practice. The two experts performed annotations independently without access to each other's labels or to the verifier's predictions.

K.3 Language Adaptation Training Details

We conduct language adaptation via supervised causal language modeling on Stage-1 prompt–response pairs. Each sample is formatted with a Llama-3 style chat template, and we compute the training loss only on assistant tokens by masking the prompt tokens with -100. We apply dynamic padding using the eos_token_id (without introducing a new pad token) and tail-truncate sequences to a maximum length of 4096 tokens. Training is implemented with the Accelerate framework and DeepSpeed, using bf16 mixed precision and gradient checkpointing. We optimize with AdamW and a cosine learning-rate schedule with linear warmup. The best checkpoint is selected based on an EMA-smoothed training loss. Table 12 summarizes the hyperparameters used for language adaptation.

K.4 Reasoning Cold-start Training Details

We initialize the model's reasoning capability via a cold-start supervised fine-tuning stage on chain-of-thought annotated data. Each training sample is formatted using a Llama-3 style chat template, where the user message is the original prompt and the assistant response is structured as: "## Thinking" followed by the annotated rationale (Complex_CoT), and "## Response" followed by the final answer (ground_truth). We compute the training loss only on assistant tokens by masking all prompt tokens with -100.

LA-Hyperparameter	Settings
Max sequence length	4096
Micro batch size (per GPU)	32
Gradient accumulation steps	1
Global batch size	256
Optimizer	AdamW
Learning rate	5e-6
Weight decay	0.01
Warmup ratio	0.03
LR scheduler	cosine
Epochs	3
Padding strategy	EOS padding
EMA decay (for selection)	0.9

Table 12: Language adaptation hyperparameter settings.

We employ dynamic padding with eos_token_id (without introducing a new pad token) and tail-truncate sequences to a maximum length of 1536 tokens. Training is implemented with Accelerate and DeepSpeed under bf16 mixed precision and gradient checkpointing, using AdamW with a cosine learning-rate schedule and linear warmup. The best checkpoint is selected based on an EMA-smoothed training loss. Table 13 summarizes the hyperparameters used in this cold-start stage.

RC-Hyperparameter	Settings
Max sequence length	1536
Micro batch size (per GPU)	8
Gradient accumulation steps	2
Global batch size	128
Optimizer	AdamW
Learning rate	1e-6
Weight decay	0.01
Warmup ratio	0.03
LR scheduler	cosine
Epochs	10
Padding strategy	EOS padding
EMA decay (for selection)	0.9

Table 13: Reasoning cold-start training hyperparameter settings.

K.5 Reinforcement Learning Training Details

We adapt the GRPO implementation from the trl framework for reinforcement learning training. Our setup includes (i) a Task-optimal reward verifier that has been validated through reliability testing, and (ii) a token-level auxiliary scaffolding reward to provide suitable low-resource language form. For LLaMA-3.1-8B-Instruct, we inject LoRA adapters into the projection modules q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj in every decoder block. The details of training hyperparameters is summa-

rized in Table 14, which we found the best during the experiments.

RL-Hyperparameter	Settings
Batch Size	8
Learning Rate	5e-6
LoRA_r	16
LoRA_alpha	32
LoRA_dropout	0.05
Task-optimal reward threshold	0.5
Reward Start Ratio	0.1:0.9
Reward End Ratio	0.9:0.1
Annealing Strategy	cosine
GRPO_clip	0.2
KL_beta	0.001
GRPO_num_generations	8

Table 14: Reinforcement Learning Hyperparameter settings.

K.6 Software Packages

Libraries (versions). torch 2.9.0; transformers 4.57.3; accelerate 1.12.0; deepspeed 0.18.2; tqdm 4.67.1; jinja2 3.1.6; swanlab 0.7.2.

L Data Structures

We organize the training corpus and benchmarks using a simple and consistent data format. Each training instance is associated with a stable identifier, a single medical intent category, and one instruction format. Training instances further include a question, its corresponding answer, and a reasoning rationale, all instantiated from predefined templates and explicitly grounded in the same source passage. Benchmark items are stored in a lighter format, containing only an identifier, the question text, and the gold answer.

M Human Evaluation of Native Hindi Medical Reasoning

This appendix describes the human evaluation protocol used to assess whether the language scaffolding reward in DSR-RL improves native Hindi medical reasoning.

Models compared. We compare two reinforcement learning variants trained from the same Stage-II checkpoint: (1) **HiMed-8B**, trained with the full decaying-reward scaffolding framework including the language-form reward, and (2) **RL w/o Scaffolding**, trained using the same reinforcement learning procedure and data but without the language-form reward. All other training configurations, including model architecture, data, optimization settings, and training duration, are kept identical.

Evaluation data. We randomly sample 200 prompts from the evaluation pool, covering both Western medicine and Indian systems of medicine. For each prompt, both models generate a full response including an explicit medical rationale. Model outputs are anonymized and randomly ordered to avoid positional or identity bias.

Annotators. Two annotators with formal medical training and native or near-native proficiency in Hindi independently perform the evaluation. Annotators are blinded to model identity and training configuration. Annotators are hired from professional medical institute. The experts are paid 500 currency units per hour, which is notably higher than the local doctor's hourly wage. In total, 4,000 currency units are spent at this inspection stage.

Evaluation criterion. Annotators perform a blind pairwise comparison and are asked the following question for each response pair:

> *Which response exhibits more native Hindi medical reasoning?*

Judgments are guided by the following rubric, focusing on the *organization of medical reasoning* rather than surface language quality:

- **Hindi-based reasoning organization:** Whether medical reasoning steps (e.g., interpretation of symptoms, causal relations, or diagnostic logic) are explicitly structured within Hindi sentence forms rather than expressed as minimal or answer-driven statements.
- **Symptom–mechanism linkage in Hindi:** Whether the response articulates connections between symptoms, underlying mechanisms, and conclusions directly in Hindi, instead of omitting reasoning or relying on implicit English-style shortcuts.
- **Avoidance of English-centric reasoning patterns:** Whether the response avoids answer-first or template-like explanations that could plausibly be translated verbatim from English (e.g., "The correct answer is B because …").
- **Reasoning coherence:** Whether the medical reasoning forms a logically connected

and internally consistent progression in Hindi, where each step (e.g., symptom interpretation, causal inference, and diagnostic conclusion) follows coherently from the previous one rather than appearing as isolated or fragmented statements.

- **Reasoning faithfulness:** Whether the reasoning process is faithfully grounded in the information explicitly stated in the prompt, with conclusions supported by articulated symptom–mechanism relations in Hindi, rather than introducing unstated assumptions, external knowledge, or implicit reasoning shortcuts.

Annotators are instructed *not* to judge factual correctness, verbosity, or stylistic fluency, and *not* to penalize principled code-mixing for medical terminology, numerals, or standardized symbols. A *Tie* option is explicitly provided and selected when neither response clearly exhibits more native Hindi medical reasoning according to the rubric.

Agreement and analysis. Annotators independently select one of three options for each response pair: *Response A*, *Response B*, or *Tie*, without access to any model identity information. Preferences are mapped back to their originating models only after annotation is completed. Preference rates are computed on non-tie decisions, while tie rates are reported separately to avoid forcing distinctions when no clear difference is perceived.

Across annotators, HiMed-8B is consistently preferred over the RL variant without the language scaffolding reward. Specifically, among non-tie judgments, HiMed-8B is preferred in 64.5% and 64.7% of cases by the two annotators, respectively. Using a conservative panel aggregation that requires annotator agreement and assigns *Tie* otherwise, HiMed-8B is preferred in 67.4% of non-tie cases, with a tie rate of 32.5%.

Inter-annotator agreement, measured by Cohen's κ under the three-way decision setting, is 0.61, indicating substantial agreement. A two-sided binomial test on non-tie panel decisions shows that the preference for HiMed-8B is statistically significant. These results demonstrate that annealing the language scaffolding reward leads to more native Hindi medical reasoning behaviors as perceived by human experts.

N Use of AI Assistants

AI assistants were used in a limited and auxiliary manner during the preparation of this manuscript. Specifically, large language models were employed to support language polishing, grammatical refinement, and clarity improvements in portions of the text written in English. All technical content, including research ideas, methodological design, experimental setup, results, and interpretations, was conceived, implemented, and validated by the authors. AI assistants were not used to generate experimental data, conduct analyses, derive conclusions, or make scientific judgments. All code, model training, evaluations, and statistical analyses were performed by the authors.

O OCR Result Inspection UI

Here we provide the UI used in the OCR manual inspection. The interface is designed around a side-by-side verification workflow.

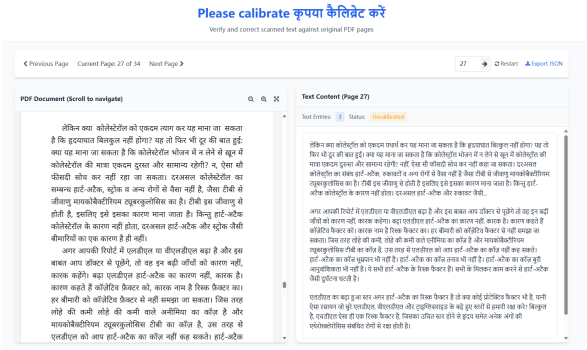

Figure 3: UI used in the OCR manual inspection

The original scanned PDF page is displayed on the left, while the corresponding OCR-extracted text for the same page is shown on the right. This layout enables annotators to directly compare source content and recognized text at the page level, quickly identify discrepancies, and perform targeted corrections. Page navigation, calibration status indicators, and export functionality are integrated to support efficient, traceable, and systematic manual OCR calibration across documents.