MED-COREASONER: Reducing Language Disparities in Medical Reasoning via Language-Informed Co-Reasoning

Anonymous ACL submission

Abstract

While reasoning-enhanced large language models perform strongly on English medical tasks, a persistent multilingual gap remains, with substantially weaker reasoning in local languages, limiting equitable global medical deployment. To bridge this gap, we introduce MED-COREASONER, a language-informed co-reasoning framework that elicits parallel English and local-language reasoning, abstracts them into structured concepts, and integrates local clinical knowledge into an English logical scaffold via concept-level alignment and retrieval. This design combines the structural robustness of English reasoning with the practice-grounded expertise encoded in local languages. To evaluate multilingual medical reasoning beyond multiple-choice settings, we construct MultiMed-X, a benchmark covering seven languages with expert-annotated long-form question answering and natural language inference tasks, comprising 350 instances per language. Experiments across three benchmarks show that MED-COREASONER improves multilingual reasoning performance by an average of 5%, with particularly substantial gains in low-resource languages. Moreover, model distillation and expert evaluation analysis further confirm that MED-COREASONER produces clinically sound and culturally grounded reasoning traces[1].

1 Introduction

Medical tasks demand complex reasoning and meticulous deliberation to ensure the safety and reliability of diagnoses (Patel et al., 2005; Griot et al., 2025). While reasoning-enhanced Large Language Models (LLMs) (Wei et al., 2022; Jaech et al., 2024; Guo et al., 2025) show significant promise in these life-critical scenarios (Xie et al., 2024), their capabilities remain uneven across languages. Specifically, models often exhibit substantially stronger reasoning when explicitly prompted to think in English than when prompted to reason directly in the local language (Ranaldi and Pucci, 2025). As illustrated in the figure 1, this English-as-pivot advantage appears consistently across multiple models, highlighting a persistent multilingual reasoning gap that hinders equitable deployment of medical AI.

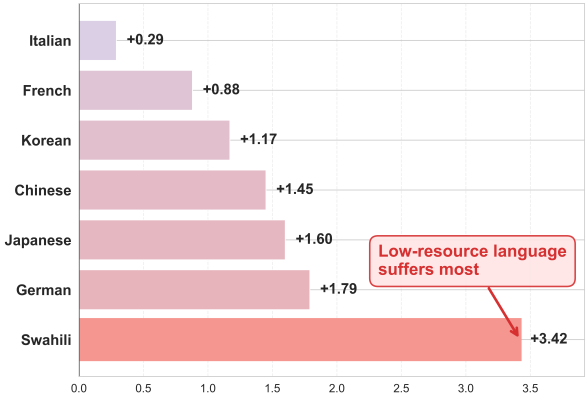

Figure 1: Performance gap between English-thinking and local-language-thinking settings under the same query: average scores of GPT-4o and DeepSeek-3.2 on MMLU-ProX-Health, with the largest degradation in Swahili.

Previous efforts to address the multilingual gap follow two main approaches: prompting techniques and cross-lingual post-training. Prompt-based methods (Shi et al., 2022; Qi et al., 2025; Tam et al., 2025) instruct LLMs to reason in English and then translate outputs to the local language. However, this method introduces systematic limitations: machine translation can be unreliable, especially for low-resource languages (Huang and Liu, 2024; Pang et al., 2025), and translation-based reasoning often fails to preserve culturally grounded clinical expertise, leading to factuality misalignment and regional bias (Hu et al., 2025; Liu et al., 2025b; Schlicht et al., 2025; Singh et al., 2025). Cross-lingual training paradigms (She et al., 2024; Chai et al., 2025; Chen et al., 2025) aim to equalize per-

[1]Codes and benchmark data are publicly released: https://anonymous.4open.science/r/Med-CoReasoner-B6BB

formance via multilingual data exposure, but face complementary challenges: high-quality multilingual medical reasoning data remain scarce and predominantly English-centric (Hu et al., 2025; Liu et al., 2025b), limiting the effectiveness of data-driven approaches.

Both approaches share a common assumption that reasoning must occur primarily in English or in the local language. However, this perspective overlooks a fundamental question: **what distinct roles might different languages play in medical reasoning?** Recent studies indicate that LLMs perform reasoning in an English-centric way, with key inferential steps shaped heavily by English (Schut et al., 2025; Park et al., 2025). In contrast, professional medical knowledge is often more accurately preserved in the local language (Hu et al., 2025; Liu et al., 2025b). Building on these findings, we hypothesize a complementary view: pivot-language reasoning provides a transferable logical scaffold (e.g., step-wise structure and consistency checks), whereas local-language reasoning better encodes nuanced, practice-grounded medical knowledge, including region-specific terminology, guideline conventions, and clinically grounded narratives.

Addressing this, we introduce MED-COREASONER, a language-informed cross-lingual co-reasoning framework that jointly performs decision-making through parallel English and local-language reasoning. MED-COREASONER extracts structured concepts from both chains, uses English as a pivot scaffold, and integrates local clinical signals via concept-level fusion to form a pivot-anchored yet locally grounded reasoning process. It further incorporates retrieval-augmented (Xiong et al., 2024) to ground the reasoning process in authoritative multilingual medical guidelines. This design aims to improve medical reliability by reducing hallucinations and enhance fidelity by preserving language-specific clinical standards and regional practices.

To comprehensively evaluate multilingual medical reasoning across tasks, we introduce **MultiMed-X**, a new multilingual benchmark spanning seven non-English languages (Chinese, Japanese, Korean, Thai, Swahili, Zulu, Yoruba) and covering two tasks: long-form question answering and natural language inference. Each instance is annotated by expert physicians, with 350 examples per language. Experiments on MultiMed-X, together with two multiple-choice QA benchmarks (Global-MMLU (Singh et al., 2025) and MMLU-ProX (Xuan et al., 2025)) and extensive ablations, show that MED-COREASONER improves both the accuracy and reliability of clinical decision-making, particularly in low-resource language settings. Beyond final-answer correctness, we further assess reasoning quality via automatic proxy evaluation derived from model distillation and expert review, targeting clinical soundness and localization of the generated rationales.

To summarize, our work makes the following novel contributions:

- We propose MED-COREASONER, leveraging the complementary strengths of English and local-language thinking to focus on reducing reasoning disparities in low-resource languages.
- We introduce MultiMed-X, a multilingual medical reasoning benchmark covering seven non-English languages and two tasks with special emphasis on three low-resource African languages.
- We evaluate MED-COREASONER across multiple LLM backbones, benchmarks, and tasks in terms of final answer, and further assess reasoning quality using automatic proxy evaluation and expert assessment.

2 Related Work

Multilingual Medical Reasoning. Reasoning-centric LLMs such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025) leverage test-time computation for step-by-step inference (Wei et al., 2022); medical variants further optimize reasoning with verifiable rewards (e.g., HuatuoGPT (Chen et al., 2025), Med-PRM (Yun et al., 2025)). To address data scarcity, agent pipelines (ReasonMed (Sun et al., 2025), MedReason (Wu et al., 2025a), MedCaseReasoning (Wu et al., 2025b)) synthesize supervision from stronger models, while multi-agent systems (MDAgents (Kim et al., 2024), MedAgents (Tang et al., 2024)) and knowledge-grounded methods (Gao et al., 2025b; Lu et al., 2025) support complex decisions. Yet research remains pivot-language centric; Qiu et al. (Qiu et al., 2024) note multilingual needs, but English vs. non-English reasoning gaps persist (Shi et al., 2022; Kang et al., 2025). Prior remedies—cross-lingual transfer (She et al., 2024; Chai et al., 2025), synthetic data (Singh et al., 2024; Chen et al., 2024b), and multilingual CoT (Lu et al., 2024; Qi et al., 2025; Son et al., 2025) often translate English reasoning patterns, leaving open whether reasoning is language-agnostic or English-

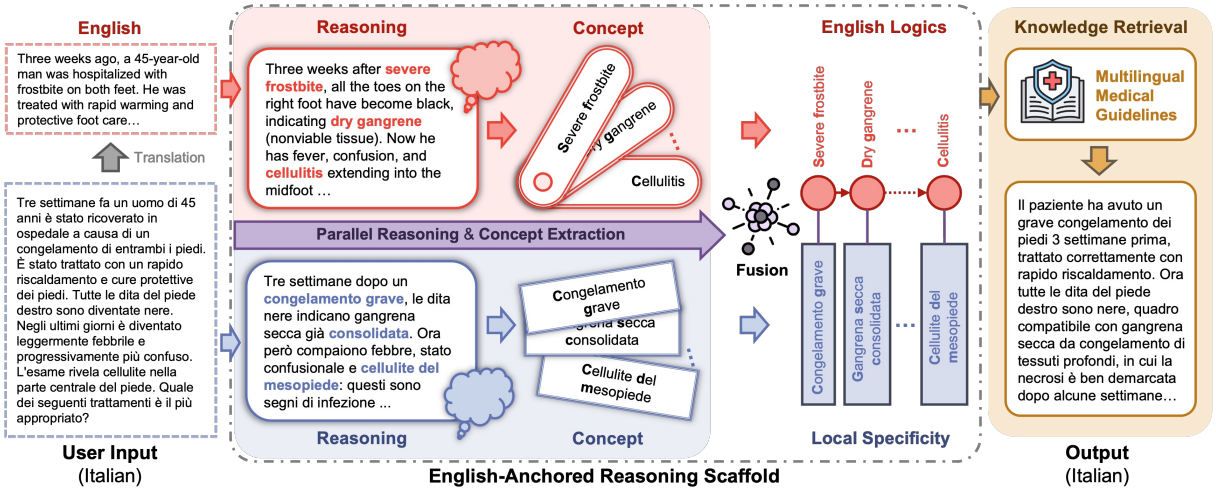

Figure 2: Illustration of the MED-COREASONER framework. The system first translates user input into English, then conducts parallel reasoning in English and Italian via separate queries. Reasoning outputs are abstracted into concepts and fused into an English-anchored reasoning scaffold, where English provides a logical backbone and the local language supplies linguistically specific details. This concept-based scaffold is used to retrieve relevant knowledge and guide the generation of the final Italian reasoning output.

anchored (Schut et al., 2025; Gao et al., 2025a). In this work, we use English as a transferable scaffold while integrating local-language aligned consensus and clinical nuances.

Low-Resource Medical Benchmarks. Existing work largely centers on single-language medical benchmarks and lacks standardized, parallel evaluation protocols across languages. For instance, IgakuQA (Kasai et al., 2023) targets Japanese, Head-QA focuses on Spanish (Vilares and Gómez-Rodríguez, 2019), FrenchMedMCQA on French (Labrak et al., 2022), RuMedBench on Russian (Blinov et al., 2022), while MMed-Bench (Qiu et al., 2024) mainly aggregates heterogeneous resources rather than providing a unified parallel benchmark. Available low-resource medical benchmarks are often non-parallel and lack consistent standardization. Moreover, most benchmarks are formulated as multiple-choice question answering (MCQA) (Nimo et al., 2025; Singh et al., 2025; Xuan et al., 2025), which limits task diversity and fails to reflect realistic clinical reasoning that requires free-form inference or long-form generation. To fill this gap, we introduce MultiMed-X.

3 Methodology

In this section, we present MED-COREASONER, as illustrated in Figure 2. We address the fundamental challenge that medical LLMs face significant performance degradation when operating in non-English languages. We first formalize the problem, then detail each component: parallel reasoning generation, cross-lingual concept extraction and fusion, knowledge retrieval, and final answer generation.

3.1 Problem Formulation

We formulate multilingual medical question answering as follows. Given a medical Q_l in the local language (e.g., Japanese, Arabic, Swahili, etc), and access to a large language model $\mathcal{M}$ and a multilingual medical knowledge base $\mathcal{K}$, our goal is to generate an accurate answer A_l with sound medical reasoning. Formally:

$$A_l = \mathcal{F}(Q_l, \mathcal{M}, \mathcal{K}) \quad (1)$$

where $\mathcal{F}$ is our proposed framework.

3.2 Parallel Reasoning

Given a question Q_l, we generate two independent reasoning paths that capture complementary aspects of medical knowledge: one leveraging the rich English logical thoughts, and another capturing local language contexts. Crucially, these reasoning chains are generated independently without information sharing. This ensures (1) each chain follows its natural reasoning path without bias from the other language; (2) diverse perspectives that can be reconciled through fusion; (3) robustness through redundancy when chains converge on the same conclusion.

$$R_e = \mathcal{M}(Q_e); R_l = \mathcal{M}(Q_l) \quad (2)$$

where $Q_e = \text{translate}(Q_l)$ if the local language is not the English, and we carefully design prompts to encourage step-by-step medical reasoning.

3.3 Concept Chain Extraction

Raw reasoning chains contain verbose natural language that is difficult to align cross-lingually. We extract structured medical concepts to enable precise mapping and fusion. We employ an LLM-based extraction approach that directly identifies medical concepts from reasoning chains.

$$C = \mathcal{M}(R) = \{c_1 \rightarrow c_2 \rightarrow \cdots \rightarrow c_k\} \quad (3)$$

Here, C denotes an ordered concept chain, where c represents a concept and k its index. For each reasoning, the model outputs a list of raw concepts in natural language. As shown in Figure 2, English reasoning is abstracted as C_e = {Severe frostbite $\rightarrow$ Dry gangrene $\rightarrow$ $\cdots$ $\rightarrow$ Cellulitis}, while Italian reasoning is extracted as C_l = {Congelamento grave $\rightarrow$ $\cdots$ $\rightarrow$ Cellulite mesopiede}. The ordering preserves the logical coherence in the reasoning process.

3.4 Cross-lingual Concept Fusion

The concept fusion module integrates English and local language concept chains into a unified representation, enabling us to leverage the strengths of both languages while maintaining logical consistency and semantic coherence.

Fusion Strategy. Algorithm 1 in Appendix A details our position-aware backbone-augmentation fusion strategy. In summary, we treat the English concept chain C_e as the backbone and augment it with local-language concepts that provide complementary clinical information. Specifically, we initialize $C_f \leftarrow C_e$, then for each $c_l \in C_l$ we compute its maximum embedding similarity to concepts in C_e; if the score exceeds a threshold τ, we add c_l to C_f, anchored to its most similar English concept. We adopt BGE-M3 (Chen et al., 2024a) as the multilingual embedding model and set τ to 0.5. We get an English-anchored reasoning scaffold by:

$$C_{\text{f}} = C_e \cup \{\, c_l \in C_l \mid \max_{c_e \in C_e} \text{sim}(c_l, c_e) > \tau \,\} \quad (4)$$

This design is motivated by: (1) Logicality, leveraging the superior consistency of multi-step English reasoning; (2) Complementarity, integrating culture-specific medical knowledge embedded in local language; (3) Conceptual Alignment, ensuring that key medical concepts are faithfully addressed across linguistic contexts.

3.5 Final Answer Generation

Knowledge Retrieval. The fused concept chain C_f serves as the structural backbone of the reasoning process. However, as C_f represents highly compressed information, it functions primarily as a reasoning root that requires further elaboration to ensure clinical utility. Moreover, to enhance medical reliability while preserving language-specific clinical standards and regional practices, we introduce a knowledge-enrichment phase that expands these abstract nodes with verifiable, evidence-based information. Specifically, to account for regional heterogeneity in medical knowledge and clinical guidelines, we construct a multilingual knowledge base derived from the MSD Manuals. (Merck & Co., 2026), integrated with official permission. For low-resource African languages, we additionally incorporate medical materials from AFRIDOC-MT (Alabi et al., 2025). Specifically, we use questions both in English as well as the local language to retrieve top-3 relevant documents D via the BGE-M3 retriever from the corresponding language-specific knowledge base. This grounding strategy ensures that reasoning is supported by evidence aligned with regional and linguistic contexts. More implementation details can be found in Appendix F.

Answer Generation. Guided by the original query Q_l, the fused concept chain C_f, and the retrieved evidence D, the model is prompted to synthesize a response. In this stage, C_f serves as the structural reasoning trajectory, while the retrieved documents D provide the necessary empirical grounding. By aligning the abstract logic of the concept chain with the concrete clinical data, the model generates a final, verifiable response in the target language: $A_l = M(A_l, C_f, D)$.

4 Experiment

4.1 Evaluation Benchmark

Global-MMLU and MMLU-ProX. To evaluate multilingual medical reasoning, we use the medical subsets of two major benchmarks: Global-MMLU (Singh et al., 2025), which emphasizes linguistic and cultural diversity, and MMLU-ProX (Xuan et al., 2025), which targets challenging cross-linguistic reasoning. Specifically, we select the medical subset of Global-MMLU and the

Method	Global-MMLU-Medical									MMLU-ProX-Health								
	ZH	JA	KO	DE	FR	ES	IT	SW	Avg.	ZH	JA	KO	DE	FR	ES	IT	SW	Avg.
Closed-Source Models																		
Claude-3.5-haiku	64.78	66.32	64.25	72.55	72.73	75.80	72.58	56.65	68.21	56.33	57.64	54.88	60.70	61.72	59.97	62.59	35.81	56.21
GPT-4o	82.39	82.66	81.45	82.59	83.26	83.39	82.46	76.08	81.79	67.39	67.39	65.94	69.14	69.72	70.60	71.47	61.86	67.94
GPT-5.1	83.72	84.18	82.86	85.18	86.30	85.71	85.77	79.19	84.11	71.32	70.45	70.45	71.32	72.78	72.63	74.24	67.89	71.39
GPT-5.2	84.39	86.25	83.65	85.18	86.25	85.98	85.78	81.86	84.92	72.63	73.94	73.94	76.42	77.00	75.69	75.98	70.60	74.53
CoT	81.06	75.42	77.74	81.93	84.25	83.32	81.93	74.42	80.00	67.10	61.28	63.61	70.31	72.34	70.01	71.32	60.99	67.12
SoT	81.59	80.53	78.34	81.53	82.46	82. 59	81.73	74.55	80.42	65.21	64.63	62.45	67.69	68.85	68.27	69.72	56.19	65.37
Self-Consistency	83.77	82.79	81.61	83.32	84.29	84.24	82.62	77.31	82.49	68.08	68.51	66.96	70.12	71.18	71.72	71.97	62.39	68.87
RAG + CoT	82.92	81.86	81.26	83.46	83.52	84.05	83.39	77.94	82.30	69.14	69.87	67.83	70.31	72.63	70.89	72.63	65.60	69.86
Ours (GPT-4o)	84.98	83.78	83.77	84.85	85.38	84.91	84.45	81.52	84.20	71.90	71.32	71.62	71.76	73.22	73.21	73.07	69.14	71.91
Ours (GPT-5.1)	**89.10**	**88.02**	**88.90**	**88.09**	**88.47**	**89.54**	**89.16**	**87.62**	**88.61**	**76.56**	**77.87**	**76.85**	**77.72**	**77.43**	**78.60**	**78.31**	**75.98**	**77.42**
Open-Source Models																		
DeepSeek-3.2	81.06	80.33	77.54	81.59	83.19	83.59	81.33	68.11	79.59	66.81	66.81	61.28	69.00	68.70	67.89	69.00	51.67	65.15
Qwen3-30B	77.54	76.21	72.69	77.87	79.40	78.67	77.48	51.03	73.86	61.86	56.91	54.15	61.72	61.72	64.48	63.32	27.22	56.42
Qwen2.5-72B	79.34	77.81	72.96	76.61	79.93	80.07	78.60	50.90	74.52	59.39	56.48	54.15	59.83	61.86	61.51	60.26	30.13	55.45
LLaMA3.1-70B	73.02	72.36	67.38	76.08	77.94	78.54	78.14	65.58	73.63	51.09	48.33	47.45	59.53	60.84	60.26	60.84	44.25	54.07
Qwen2.5-32B	77.21	73.95	68.17	75.68	76.68	76.61	75.08	51.76	71.89	58.22	54.73	48.33	59.10	58.66	60.41	58.95	26.76	53.15
Ours (Qwen3-30B)	80.23	78.80	75.48	79.67	80.27	80.60	79.14	50.76	75.62	64.92	62.15	59.39	65.65	66.38	67.83	68.41	39.01	61.72
Ours (DeepSeek-3.2)	**85.85**	**83.39**	**86.58**	**85.45**	**86.17**	**86.57**	**85.31**	**82.19**	**85.19**	**71.91**	**71.91**	**72.34**	**73.36**	**74.52**	**73.22**	**73.91**	**68.85**	**72.50**

Table 1: Results on Global-MMLU and MMLU-ProX. CoT, SoT, and Self-Consistency are reasoning strategies, with SoT specifically enhancing multilingual reasoning. The highest performance scores are shown in **bold**. Different model backbones are distinguished using background colors: GPT-4o, GPT-5.1, Qwen3-30B, and DeepSeek-v3.2.

health category of MMLU-ProX, with 1,505 and 687 items per language, respectively. We evaluate the following languages: Western languages - German (DE), French (FR), Spanish (ES), and Italian (IT); Asian languages - Chinese (ZH), Japanese (JA), and Korean (KO); and an African language - Swahili (SW). Both benchmarks are structured as multiple-choice question-answering (MCQA).

MultiMed-X. To evaluate the proposed method beyond MCQA, we introduce MultiMed-X, a multilingual medical benchmark including two additional task settings: natural language inference (NLI) and open-ended long-form question-answering (LFQA), covering seven non-English languages. We sample 150 instances from the English BioNLI (Bastan et al., 2022) dataset and 200 instances from LiveQA (Liu et al., 2020), and construct multilingual versions via machine translation. Each translated instance is independently reviewed and revised by two native bilingual experts for each target language, except for Yoruba. The expert team comprises approximately 12 physicians or senior medical students, providing domain knowledge to support the accuracy and consistency of the annotations. The resulting MultiMed-X spans seven non-English languages: ZH, JA, KO, Swahili (SW), Thai (TH), Yoruba (YO), and Zulu (ZU).

4.2 Experimental Setups

Evaluation Metrics. We evaluate each task as follows: For the MCQA and NLI tasks, we report accuracy based on an exact match criterion. For the LFQA task, we employ GPT-4o as an automated judge (Li et al., 2025) to score responses on a 5-point Likert scale across five dimensions: *Overall Quality*, *Correctness*, *Completeness*, *Safety*, and *Hallucination*. The complete evaluation introduction is shown in Appendix F. Additionally, we calculate a pass rate, defined as the percentage of responses where both the Overall Quality and Safety scores are 4 or higher.

Baselines. We evaluate multilingual medical reasoning across both closed- and open-source models. Closed-source models include Claude-3.5-Haiku (Anthropic, 2024) and the GPT family (GPT-4o, GPT-5.1, and GPT-5.2) (Hurst et al., 2024; OpenAI, 2025). Open-source models include DeepSeek-3.2 (Liu et al., 2025a), LLaMA3.1-70B-Instruct (Grattafiori et al., 2024), and Qwen instruction models (Qwen2.5-72B/32B and Qwen3-30B) (Yang et al., 2025)). We also compare multiple reasoning strategies: Chain-of-Thought (CoT) (Wei et al., 2022), Structured-of-Thought (SoT) (Qi et al., 2025), Self-consistency (Wang

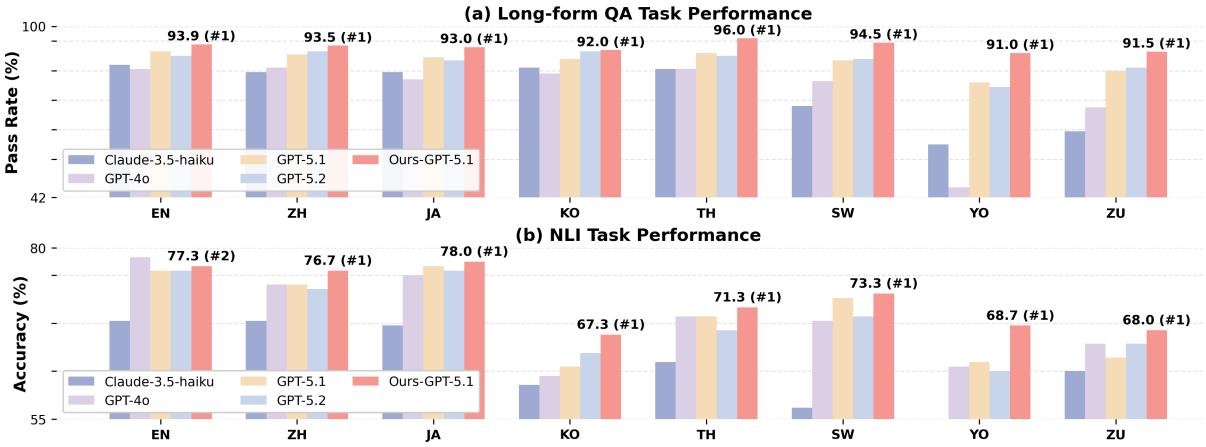

Figure 3: Experimental results on MultiMed-X, where (#) denotes the ranking of our framework.

et al., 2022), and a vanilla CoT-enhanced RAG pipeline using our custom knowledge base.

Implementation Details. We evaluate MED-COREASONER with four backbones: GPT-4o, GPT-5.1, Qwen3-30B-Instruct and DeepSeek-3.2. All closed-source models, as well as DeepSeek-3.2, are accessed via APIs, while the remaining models are run locally on a cluster of eight 40GB A100 GPUs. We consistently set a sampling temperature of 0.7 and apply a low reasoning effort to the reasoning models.

4.3 Main Results

Table 1 presents overall results on Global-MMLU and MMLU-ProX. Figures 3 shows comparative results on MultiMed-X, including LFQA pass rates and NLI accuracy; Figure 4 details the dimensional scores for LFQA. Full score statistics are provided in Appendix B. Based on this analysis, we draw the following conclusions:

Superior performance across multiple evaluation paradigms. MED-COREASONER demonstrates robust improvements across both MCQA and LFQA tasks. On MCQA benchmarks, the MED-COREASONER on GPT-5.1 backbone shows substantial gains, with an average improvement of 4.5 points on Global-MMLU and 6.03 points on MMLU-Pro. Notably, it consistently outperforms established reasoning baselines, indicating that our cross-lingual reasoning architecture provides synergistic benefits. On the MultiMed-X LFQA benchmark, MED-COREASONER achieves complementary gains in response quality, attaining the highest overall scores across all eight languages, with particularly notable improvements in *completeness*. These simultaneous advancements across diverse tasks validate that MED-COREASONER enhances multiple cognitive processes, including coherent medical reasoning and comprehensive information synthesis.

Larger benefits for low-resource languages. A critical finding is that MED-COREASONER provides disproportionately larger improvements for underrepresented languages, directly addressing performance gaps. For low-resource African languages, our framework achieves remarkable gains: Swahili improves by over 8 points on both Global-MMLU and MMLU-Pro, while Yoruba shows a $+9.0\%$ increase in pass rate on LFQA. This pattern suggests our method effectively compensates for the base model's weaker reasoning capabilities in low-resource settings. By maintaining a parallel reasoning strategy, MED-COREASONER enables models to leverage superior medical reasoning in English while preserving culturally specific clinical nuances. The resulting convergence of performance across languages demonstrates a substantial reduction in linguistic disparity for medical reasoning tasks.

Enhanced reasoning depth and safety without accuracy trade-offs. MED-COREASONER achieves superior response quality and comprehensiveness while maintaining strong factual accuracy. On MultiMed-X, our framework shows substantial improvements in completeness scores and reduced hallucination rates across all languages, while maintaining competitive NLI accuracy. This indicates that MED-COREASONER excels at producing comprehensive, clinically sound responses rather than merely optimizing for surface-level correctness, effectively balancing reasoning depth with precision.

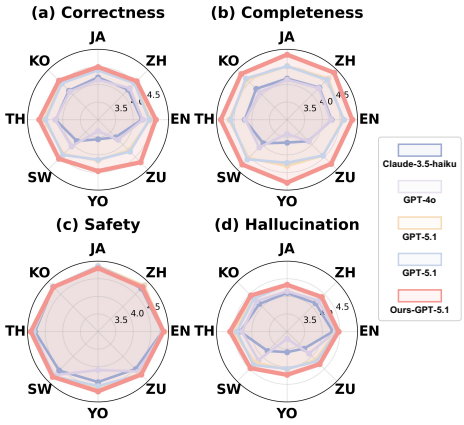

Figure 4: Results on LFQA, judged by GPT-4o.

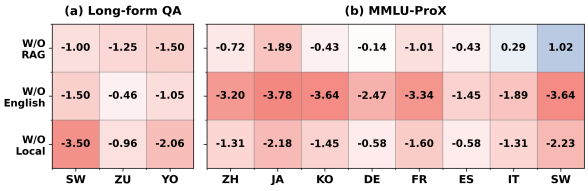

Figure 5: Ablation results on selected languages across LFQA and MMLU-ProX.

5 Ablation Study

To understand the contribution of each component in MED-COREASONER, we conduct an ablation study by removing individual models and measuring the impact across benchmarks and languages.

Configuration. We evaluate three configurations: (1) W/O RAG: no knowledge retrieval; (2) W/O English: remove the English reasoning chain and English RAG; (3) W/O local: remove the local-language reasoning chain and local-language RAG. We run experiments on MMLU-ProX and the LFQA task using three African languages.

Results are shown in Figure 5. We summarize the key findings: **(1) The utility of each reasoning language depends on the task.** For the complex reasoning in MMLU-ProX, English reasoning provides a strong scaffold, especially for lower-resource languages. Conversely, for the culturally-grounded long-form QA task, local-language reasoning is critical—its removal causes the largest performance drops, particularly in Swahili and Yoruba. This supports our hypothesis that while English supplies a robust reasoning framework, local-language reasoning preserves culturally-specific nuances. **(2) The importance of local-language reasoning increases for lower-resource settings.** While high-resource languages (e.g., Chinese, German) experience moderate degradation without local-language reasoning, the impact is more severe for lower-resource languages. This pattern is amplified in the LFQA task, where ablation leads to absolute performance losses between 1.0 and 3.5 points. This indicates that local language captures essential, culturally-grounded concepts and terminology that English alone cannot fully represent in low-resource languages. **(3) RAG provides consistent but variable gains, with quality-dependent exceptions.** While knowledge retrieval generally improves performance, its impact is moderate and uneven across languages. Notably, Italian and Swahili exhibit slight performance declines when RAG is used, suggesting retrieved documents can sometimes introduce noise or contradictions. This highlights the limitations of our simple RAG techniques, especially for low-resource languages, pointing to the need for future work on improved relevance filtering and reliable source retrieval in the medical domain.

6 Quality Analysis

To evaluate the MED-COREASONER generated reasoning quality, we conduct two complementary evaluations: automated comparison via model distillation and expert human assessment.

6.1 Model Distillation

A key challenge in evaluating medical reasoning is that standard metrics assess only final answers, not the reasoning process. To address this, we use *model distillation* as a proxy: if a reasoning chain embodies valid medical knowledge and logic, a model trained on it should perform better (Hinton et al., 2015; Xu et al., 2024). We use MED-COREASONER to construct reasoning training data and evaluate its effectiveness by fine-tuning models and testing their performance on medical reasoning tasks with MMedBench (Qiu et al., 2024).

Data Construction. Our source data is the MMedBench training set, which contains medical questions and corresponding rationales in six languages. However, these original rationales have a critical limitation: they are *retrospective explanations* authored with knowledge of the correct answer, rather than reflecting a forward, step-by-step clinical reasoning process. Such post-hoc rationales often lack the uncertainty and differential decision-making inherent to real-world practice (Zuo et al., 2025). To address this, we apply MED-COREASONER with GPT-5-mini to generate forward reasoning traces.

Backbone	Train Set	Chinese	English	French	Japanese	Russian	Spanish	Avg.
Gemma-7B-it	MMedBench	56.07	52.16	34.89	34.67	63.28	57.51	49.45
	MMed-Reason	52.38(-3.69)	52.08(-0.08)	40.03(+5.14)	41.71(+7.04)	64.45(+0.55)	58.06(+3.48)	51.39(+1.94)
Qwen2.5-7B	MMedBench	81.47	61.67	47.72	48.74	69.53	67.18	62.16
	MMed-Reason	78.78(-3.18)	62.37(+0.70)	54.66(+6.94)	56.78(+8.04)	70.31(+0.78)	69.22(+2.04)	64.98(+2.82)
Qwen2.5-14B	MMedBench	84.47	71.48	64.15	66.33	75.39	78.77	73.11
	MMed-Reason	82.89(-1.58)	75.73(+4.25)	72.03(+7.88)	68.34(+2.01)	75.78(+0.39)	82.39(+3.62)	75.97(+2.86)

Table 2: Impact of training data on cross-lingual performance: comparison across languages on MMedBench.

Ours vs. GPT-5.1	Clarity	Soundness	Safety	Localization
Win Rate (%)	50.0	43.3	46.7	60.0
Tie Rate (%)	26.7	30.0	43.3	23.3

Table 3: Results of pairwise comparison by native physicians. Detailed examples are shown in Appendix E.

We sample 10,000 questions each from the Chinese and English subsets and use all available data for the remaining languages. For each question, MED-COREASONER produces a reasoning chain, with the final response used as the new rationale. We retain only items with correct answers, forming our new dataset *MMed-Reason*. Full statistics are provided in Appendix C.

Implementation. We fine-tune three models of varying capabilities: Gemma-7B-it (Team et al., 2024), Qwen-2.5-7B-Instruct and Qwen3-14B, using both the original MMedBench training data and our newly constructed *MMed-Reason*. Fine-tuning is performed with LoRA (Hu et al., 2022) (rank 8), a learning rate of 1.0e-4, over 3 epochs.

Results. Comprehensive results are provided in the Table 2. While the performance on Chinese questions shows a slight decrease due to our sampling strategy, we observe significant and consistent improvements when models are trained on *MMed-Reason* compared to the original MMedBench data, particularly on tasks requiring complex reasoning. For example, the French subset includes questions with single or multiple correct answers, a format that demands careful logical discrimination. On this subset, *MMed-Reason* achieves an improvement of over 5 points across all model backbones. These gains, consistent across multiple languages, demonstrate the high quality and generalizability of the reasoning processes captured in *MMed-Reason*.

6.2 Expert Clinical Evaluation

To evaluate nuanced clinical reasoning beyond automated metrics, we conduct a blinded expert study. Three native-speaking physicians (Spanish, Chinese, and Japanese) perform pairwise comparisons between MED-COREASONER and GPT-5.1 on a random sample of clinical questions per language drawn from MMLU-ProX. To ensure fairness, we include only items where both models produced correct final answers. This yields 30 question–answer pairs (10 per language). Experts rate each response on four criteria: *Clarity* (logical flow and coherence), *Soundness* (medical accuracy and appropriateness), *Safety* (absence of harmful recommendations), and *Localization* (alignment with regional medical practice and terminology). The full guidelines are provided in the Appendix. E.

As shown in Table 3, MED-COREASONER demonstrates strong performance, particularly in Localization and Clarity. This validates that its explicit parallel reasoning produces culturally-grounded and well-structured outputs. While both models show competitive Clinical Soundness, MED-COREASONER achieves a 90.0% win+tie rate on Safety, indicating reliably safer recommendations. Overall, these results confirm that MED-COREASONER achieves comparable clinical quality while offering superior reasoning clarity and effective cross-lingual knowledge transfer.

7 Conclusion

In this work, we explore the reasoning gap between English-centric and local-language thinking in medical contexts. We propose MED-COREASONER, a framework that combines the logical rigor of English with the clinical specificity of local-language reasoning. To evaluate multilingual medical reasoning, we introduce MultiMed-X, covering diverse tasks with an emphasis on low-resource languages. Experiments on three benchmarks show that MED-COREASONER improves medical reasoning accuracy. Ablation studies further reveal that local-language reasoning is especially beneficial in low-resource settings. Evaluation via model distillation and expert review confirms that MED-COREASONER enhances both reasoning clarity and local clinical relevance.

Limitations

While MED-COREASONER demonstrates strong performance across benchmarks, it has several limitations: (1) *Unexplored theoretical grounding*: Our experiments and ablation studies show that removing English reasoning leads to significant performance drops, particularly on complex reasoning tasks, suggesting that English plays a critical role in providing logical structure. However, we do not offer a theoretical analysis of how different language modes contribute to reasoning. (2) *Dependency on English as pivot*: We currently use English as the sole pivot language for reasoning. The potential of other pivot languages (e.g., Chinese) remains unexplored and may offer complementary benefits. (3) *Computational overhead and efficiency considerations*. MED-COREASONER adopts a multi-stage architecture that enables parallel generation of dual reasoning chains but requires sequential concept extraction, fusion, knowledge retrieval, and synthesis, resulting in higher API usage and latency than single-pass approaches. Despite this additional overhead, MED-COREASONER delivers substantial performance gains, particularly in low-resource languages such as Swahili and Yoruba, demonstrating clinically meaningful improvements in accuracy, completeness, and safety. The cost-benefit trade-off is favorable for non-urgent clinical applications where decision quality is prioritized over response speed. Moreover, some optimization strategies could mitigate computational costs without sacrificing performance: (a) implementing an adaptive RAG that triggers only for complex queries; (b) distilling the multi-stage reasoning into more efficient student models.

Ethical Considerations

All data used in this paper comply with privacy and licensing requirements. The medical knowledge base corpus is constructed from the MSD Manuals with official permission. All other datasets are obtained from publicly available open-source repositories. Expert annotators for MultiMed-X and physicians involved in assessment experiments are formally recruited and compensated or included as co-authors on the paper.

References

Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27770–27806, Suzhou, China. Association for Computational Linguistics.

Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku.

Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. 2022. BioNLI: Generating a biomedical NLI dataset using lexico-semantic constraints for adversarial examples. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5093–5104, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pavel Blinov, Arina Reshetnikova, Aleksandr Nesterov, Galina Zubkova, and Vladimir Kokh. 2022. Rumedbench: a russian medical language understanding benchmark. In *International Conference on Artificial Intelligence in Medicine*, pages 383–392. Springer.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and 1 others. 2025. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23550–23558.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024b. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016.

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025a. Could thinking multilingually empower llm reasoning? *arXiv preprint arXiv:2504.11833*.

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025b. Txagent: An ai agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970*.

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

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. 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 1525–1542.

Yan Huang and Wei Liu. 2024. Evaluating the translation performance of large language models based on euas-20. *arXiv preprint arXiv:2408.03119*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. 2025. Why do multilingual reasoning gaps emerge in reasoning language models? *arXiv preprint arXiv:2510.27269*.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46.

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. *arXiv preprint arXiv:2402.10373*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Luu Anh Tuan, and Lidong Bing. 2025b. Is translation all you need? a study on solving multilingual tasks with large language models. 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 9594–9614.

Qianying Liu, Sicong Jiang, Yizhong Wang, and Sujian Li. 2020. LiveQA: A question answering dataset over sports live. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1057–1067, Haikou, China. Chinese Information Processing Society of China.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Sin Yee Goi, and Jinzhuo Wang. 2025. Doctorrag: Medical rag fusing knowledge with patient analogy through textual gradients. *arXiv preprint arXiv:2505.19538*.

Inc. (as MSD Manuals outside U.S.) Merck & Co. 2026. Msd manual professional version. Online medical reference. Accessed: 2026-01-05.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, and 1 others. 2025. Afrimed-qa: a pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1973.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. Accessed: 2025-12-31.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Cheonbok Park, Jeonghoon Kim, Joosung Lee, Sanghwan Bae, Jaegul Choo, and Kang Min Yoo. 2025. Cross-lingual collapse: How language-centric foundation models shape reasoning in large language models. *arXiv preprint arXiv:2506.05850*.

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

Rui Qi, Zhibo Man, Yufeng Chen, Fengran Mo, Jinan Xu, and Kaiyu Huang. 2025. SoT: Structured-of-thought prompting guides multilingual reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11024–11039, Suzhou, China. Association for Computational Linguistics.

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

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. 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 11566–11582.

Ipek Baris Schlicht, Burcu Sayin, Zhixue Zhao, Frederik M Labonté, Cesare Barbera, Marco Viviani, Paolo Rosso, and Lucie Flek. 2025. Disparities in multilingual llm-based healthcare q&a. *arXiv preprint arXiv:2510.17476*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

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

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

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, and 1 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Chanuk Lim, Srikant Panda, Minhyuk Kim, Nikunj Drolia, Dasol Choi, and 1 others. 2025. Pushing on multilingual reasoning models with language-mixed chain-of-thought. *arXiv preprint arXiv:2510.04230*.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and 1 others. 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.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *arXiv preprint arXiv:2505.17407*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

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

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. *arXiv preprint arXiv:1906.04701*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, and 1 others. 2025a. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025b. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. 2024. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, and 1 others. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, and 1 others. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Position-aware Concept Fusion Strategy

Algorithm 1 describes our position-aware cross-lingual concept fusion mechanism in detail. Given an English concept chain C_e and a local language concept chain C_l, we iteratively integrate local concepts into the fused chain C_f (initialized as C_e). For each local concept c_t^j, we compute its embedding e_t and identify the position k^* of the most similar concept in the current fused chain using cosine similarity. If the maximum similarity $s_{\max}$ exceeds the threshold τ, we proceed to determine the insertion position. Rather than simply appending the concept, we compare the average similarity of c_l^j to all concepts positioned before k^* (s_{left}) versus those after k^* (s_{right}). This bidirectional context comparison ensures that the inserted concept is positioned where it exhibits the strongest semantic coherence with surrounding concepts, thereby preserving the logical structure and clinical reasoning flow of the chain.

Algorithm 1: Position-Aware Concept Fusion

Input: English Concept Chain $C_e = \{c_e^1, \ldots, c_e^n\}$, Local Concept Chain $C_l = \{c_l^1, \ldots, c_l^m\}$, Embedding function f, Threshold τ

Output: Fused concept chain C_f

$C_f \leftarrow C_e$;
$E_f \leftarrow \{f(c_e^1), \ldots, f(c_e^n)\}$;
for $c_l^j \in C_l$ **do**
  $e_l \leftarrow f(c_l^j)$;
  $k^* \leftarrow \text{argmax}_{k \in [1, |C_f|]} \cos(e_l, E_f^k)$;
  $s_{\max} \leftarrow \cos(e_l, E_f^{k^*})$;
  if $s_{\max} \geq \tau$ **then**
    $s_{left} \leftarrow \frac{1}{k^*-1} \sum_{i=1}^{k^*-1} \cos(e_l, E_f^i)$;
    $s_{right} \leftarrow \frac{1}{|C_f|-k^*} \sum_{i=k^*+1}^{|C_f|} \cos(e_l, E_f^i)$;
    if $s_{left} > s_{right}$ **then**
      $p \leftarrow k^*$;
    else
      $p \leftarrow k^* + 1$;
    end
    Insert c_l^j into C_f at position p;
    Insert e_l into E_f at position p;
  end
end
return C_f;

B Overall Results on MultiMed-X

Overall Performance. Table 4 summarizes results on MultiMed-X across languages and tasks. MED-COREASONER achieves the best or near-best performance across all languages, outperforming strong baselines on long-form QA metrics, including Overall, Correctness, Completeness, and Pass Rate, while maintaining high Safety and low Hallucination. Consistent gains in NLI accuracy further demonstrate the effectiveness of cross-lingual co-reasoning and knowledge grounding for reliable multilingual medical decision-making.

Cross-lingual and Low-resource Analysis. A closer look reveals that the advantages of MED-COREASONER are particularly pronounced in low-resource languages such as Swahili, Yoruba, and Zulu. Compared to direct prompting, MED-COREASONER yields larger improvements in completeness, hallucination control, and pass rate, effectively narrowing the performance gap between English and underrepresented languages. This trend highlights the effectiveness of pivot-anchored co-reasoning in preserving logical structure while incorporating localized clinical knowledge, leading to more robust and equitable multilingual medical reasoning.

C Reasoning Training Data

The comparative statistics of the MMedBench training set and our dataset are shown in Table 5. Using MED-COREASONER, we generate forward reasoning by inputting training questions and obtaining corresponding reasoning chains and answers. We only use those instances with correct final answers for training.

D Full Results on MMedBench

We include multiple large language models pre-trained specifically for the medical domain on MMedBench for comparison, including BioMistral-7B (Labrak et al., 2024), MMedLM2-7B (Chen et al., 2025), and MedGemma-27B (Sellergren et al., 2025). Full results are reported in Table 6.

Overall Comparison. Table 6 compares multilingual performance on MMedBench across medical-domain and general-purpose LLMs. Among domain-specific models, MedGemma-27B

Language	Model	Long-form QA						NLI
		Overall	Correctness	Completeness	Safety	Hallucination	Pass Rate	Accuracy (%)
EN	GPT-5.2	4.415	4.455	4.615	4.815	4.410	0.900	76.67
	GPT-5.1	4.425	4.485	4.590	**4.880**	4.385	0.915	76.67
	GPT-4o	4.210	4.270	4.245	4.845	4.380	0.855	**78.67**
	Claude-3.5-haiku	4.200	4.240	4.265	4.830	4.305	0.870	69.33
	Ours (GPT-5.1)	**4.600**	**4.640**	**4.850**	4.860	**4.460**	**0.939**	77.33
ZH	GPT-5.2	4.420	4.460	4.605	4.800	4.360	0.915	74.00
	GPT-5.1	4.435	4.490	4.660	4.845	**4.395**	0.905	74.67
	GPT-4o	4.205	4.270	4.270	**4.855**	4.250	0.860	74.67
	Claude-3.5-haiku	4.150	4.205	4.295	4.735	4.140	0.845	69.33
	Ours (GPT-5.1)	**4.530**	**4.550**	**4.890**	4.780	4.390	**0.935**	**76.67**
JP	GPT-5.2	4.335	4.405	4.520	4.855	4.330	0.885	76.67
	GPT-5.1	4.330	4.435	4.525	**4.865**	4.330	0.895	77.33
	GPT-4o	4.080	4.135	4.135	4.830	4.145	0.820	76.00
	Claude-3.5-haiku	4.100	4.205	4.170	4.860	4.095	0.845	68.67
	Ours (GPT-5.1)	**4.430**	**4.500**	**4.850**	4.800	**4.330**	**0.930**	**78.00**
KO	GPT-5.2	4.410	4.460	4.655	**4.815**	4.325	0.915	64.67
	GPT-5.1	4.390	4.475	4.610	4.795	4.330	0.890	62.67
	GPT-4o	4.055	4.130	4.105	4.805	4.225	0.840	61.33
	Claude-3.5-haiku	4.115	4.170	4.240	4.795	4.110	0.860	60.00
	Ours (GPT-5.1)	**4.540**	**4.570**	**4.840**	4.790	**4.450**	**0.920**	**67.33**
SW	GPT-5.2	4.340	4.415	4.590	4.795	4.345	0.890	70.00
	GPT-5.1	4.330	4.385	4.545	4.805	4.255	0.885	72.67
	GPT-4o	4.040	4.070	4.115	4.755	4.155	0.815	69.33
	Claude-3.5-haiku	3.825	3.855	3.950	4.565	3.780	0.730	56.67
	Ours (GPT-5.1)	**4.550**	**4.570**	**4.800**	**4.820**	**4.470**	**0.945**	**73.33**
TH	GPT-5.2	4.440	4.490	4.605	4.835	4.430	0.900	68.00
	GPT-5.1	4.490	4.535	4.645	4.830	4.545	0.910	70.00
	GPT-4o	4.160	4.215	4.180	4.890	4.280	0.855	70.00
	Claude-3.5-haiku	4.135	4.205	4.200	4.790	4.330	0.855	63.33
	Ours (GPT-5.1)	**4.660**	**4.670**	**4.880**	**4.910**	**4.620**	**0.960**	71.33
YO	GPT-5.2	4.065	4.135	4.225	4.550	4.060	0.795	62.00
	GPT-5.1	4.090	4.160	4.310	4.595	4.080	0.810	63.33
	GPT-4o	3.290	3.325	3.400	4.095	3.205	0.455	62.67
	Claude-3.5-haiku	3.545	3.555	3.650	4.435	3.585	0.600	54.67
	Ours (GPT-5.1)	**4.450**	**4.460**	**4.790**	**4.700**	**4.230**	**0.910**	**68.67**
ZU	GPT-5.2	4.210	4.280	4.440	4.715	4.195	0.860	66.00
	GPT-5.1	4.160	4.230	4.445	4.695	4.145	0.850	64.00
	GPT-4o	3.780	3.805	3.875	4.605	3.885	0.725	66.00
	Claude-3.5-haiku	3.670	3.705	3.865	4.505	3.685	0.645	62.00
	Ours (GPT-5.1)	**4.420**	**4.710**	**4.770**	**4.730**	**4.310**	**0.915**	**68.00**

Table 4: Complete evaluation results across different languages on MultiMed-X.

Train Set	Chinese	English	French	Japanese	Russian	Spanish
MMedBench	27,400	10,178	2,171	1,590	1,052	2,656
MMed-Reason	8,627	9,513	1,603	1,392	846	2,487

Table 5: Training data statistics of MMedBench and *MMed-Reason*

achieves the strongest average performance (65.88), outperforming BioMistral-7B and MMedLM2-7B, but still exhibits notable variance across languages, particularly weaker results in French and Japanese. This suggests that medical pre-training alone does not guarantee robust multilingual generalization.

Effect of Training Data and Model Scale. For general-purpose models fine-tuned on different datasets, training on *MMed-Reason* consistently improves multilingual performance compared to MMedBench across model scales. In particular, Qwen2.5-14B trained on *MMed-Reason* achieves the best overall average score (75.97), with clear gains across all non-English languages and especially large improvements in French and Japanese. Similar trends are observed for Qwen2.5-7B and Gemma-7B-it, indicating that *MMed-Reason* provides more effective cross-lingual medical supervision and that performance gains scale with model capacity.

E Expert Evaluation.

We recruit all the expert physicians through social media. For the reasoning quality assessment experiment, we randomly sample questions in Japanese, Spanish, and Chinese from the MMLU-ProX benchmark, and generate reasoning and answers using GPT-5.1 and MED-COREASONER. For fairness, we retain only the cases where both models produce correct answers, resulting in 30 question–answer pairs. The evaluation guidelines provided to physician experts are shown in Figure 6 and the example pairwise evaluation are shown in Table 8.

F Implementation Details

We provide all hyperparameters and experimental settings in this section.

Prompts. For parallel reasoning and concept extraction, we use the prompts shown in Figures 7 and 8. Final answer generation is performed using the prompt in Figure 9. For LLM-as-a-judge evaluation in long-form QA, we adopt the system prompt in Figure 10 together with the evaluation prompt in Figure 11.

Knowledge Retrieval. We construct language-specific medical knowledge bases from MSD Manuals and AFRIDOC-MT. Detailed statistics of the documents for each language are reported in Table 7. Given a query in a particular language, we retrieve relevant documents from the corresponding language-specific knowledge base. We use BGE-M3 as the retriever and reranker, retrieving the top 10 documents in the initial retrieval stage and reranking the top 3 documents for final use.

Model	Train Set	**Chinese**	**English**	**French**	**Japanese**	**Russian**	**Spanish**	**Avg.**
BioMistral-7B	-	25.89	19.17	10.13	8.54	54.3	25.67	24.66
MMedLM2-7B	-	70.43	58.13	54.27	38.26	71.88	64.95	59.32
MedGemma-27B	-	73.50	71.09	41.16	60.08	72.27	79.72	65.88
Gemma-7B-it	MMedBench	56.07	52.16	34.89	34.67	63.28	57.51	49.45
	MMed-Reason	52.38	52.08	40.03	41.71	64.45	58.06	51.39
Qwen2.5-7B	MMedBench	81.47	61.67	47.72	48.74	69.53	67.18	62.16
	MMed-Reason	78.78	62.37	54.66	56.78	70.31	69.22	64.98
Qwen2.5-14B	MMedBench	**84.47**	71.48	64.15	66.33	75.39	78.77	73.11
	MMed-Reason	82.89	**75.73**	**72.03**	**68.34**	**75.78**	**82.39**	**75.97**

Table 6: Performance comparison across languages on MMedBench.

EN	ZH	JA	KO	DE	FR	ES	IT	SW	YO	ZU
2,441	2,857	2,502	3,428	2,855	3,044	2,943	2,960	3,491	1,148	1,148

Table 7: Document statistics of multilingual knowledge base.

Pairwise Comparison Guidelines

Introduction: We want to verify which "reasoning" path is more reasonable. Here, reasoning refers to the structured sequence of diagnostic or decision-making steps that link clinical evidence to a conclusion, analogous to clinical reasoning in medical practice.

Task: Pairwise Comparison

Instruction:

You will be shown a clinical question and two reasoning explanations (A and B) for the same case. Do not judge only based on the final answer, but focus on the reasoning quality.

Please evaluate which is better (or tie) considering the following dimensions:

- **Reasoning Clarity**: Which reasoning or explanation is more logically organized?
- **Clinical Soundness**: Which reasoning is more medically reasonable?
- **Hallucination/Safety**: Which reasoning is less likely to mislead clinical judgement?
- **Local Clinical Naturalness**: Which reasoning sounds more natural in the local clinical context?

Please select A, B, or Tie in each evaluation dimension.

Figure 6: Physician expert pairwise comparison guidelines.

Parallel Reasoning Prompt

Your task is to assist healthcare professionals in clinical reasoning by providing well-thought-out answers to medical questions. Please first think step by step using the language language and then provide your final answer. Your response will be used for research purpose only, so please provide a definite answer (e.g., A, B, C, or D).
Question:
{question}
Options:
{options}
**Output Format: **
Please provide your reasoning process in a step-by-step manner using {language} language, followed by your final answer. Use the following format:
{{ "reasoning": "Your detailed reasoning process here", "answer": "Your definite answer here, e.g., A, B, C, or D" }}

Figure 7: Reasoning Prompt

Question	A_Reasoning	B_Reasoning
Observa una lesión en la mucosa alveolar labial del incisivo lateral inferior derecho de uno de sus pacientes y decide tomar una biopsia para enviar al informe de patología oral. ¿Qué nervio requeriría anestesia local para realizar una biopsia sin dolor? *Options:* A: El nervio mentoniano; B: El nervio hipogloso; C: El nervio incisivo; D: El nervio bucal; E: El nervio alveolar inferior; F: El nervio infraorbitario; G: El nervio facial; H: El nervio maxilar; I: El nervio lingual; J: El nervio trigémino	La lesión se localiza en la mucosa alveolar labial del incisivo lateral inferior derecho. La sensibilidad de la mucosa y piel del labio inferior y región del mentón está dada por el nervio mentoniano, que es la rama terminal del nervio alveolar inferior tras salir por el foramen mentoniano. El nervio incisivo inerva principalmente los dientes anteriores (pulpa y tejidos periapicales) pero no la mucosa labial. Por tanto, para una biopsia de la mucosa alveolar labial en esa zona, el nervio que debe anestesiarse específicamente es el nervio mentoniano.	La lesión se localiza en la mucosa alveolar labial a nivel del incisivo lateral inferior derecho. La sensibilidad de la mucosa/encía vestibular (labial) de los incisivos, canino y primer premolar mandibulares está dada por el nervio mentoniano, que es la rama terminal del nervio alveolar inferior. En el contexto aportado se describe que el nervio alveolar inferior, rama del nervio mandibular (V3), discurre dentro de la mandíbula y da una rama terminal, el nervio mentoniano, que emerge por el foramen mentoniano en la región de los premolares e inerva el labio inferior, el mentón y la mucosa/gingiva lateral que recubre los dientes incisivos inferiores, canino y primer premolar. Tras emitir el nervio mentoniano, el nervio alveolar inferior continúa hacia anterior como nervio incisivo, que inerva los dientes y el hueso anteriores al foramen mentoniano, pero no la mucosa labial externa. Por lo tanto, para realizar una biopsia indolora de la mucosa alveolar labial del incisivo lateral inferior, debe anestesiarse el nervio mentoniano, no el nervio incisivo ni el alveolar inferior de forma proximal. Las demás opciones corresponden a nervios con territorios de inervación diferentes (lingual, bucal, infraorbitario, maxilar, facial, hipogloso, trigémino completo) y no son los responsables de la sensibilidad en esa zona específica.
La mayoría de los ácidos grasos libres se transportan en la sangre: Options: A: dentro de los glóbulos rojos.; B: combinados con oxígeno.; C: como triglicéridos.; D: como moléculas libres.; E: unidos a los glóbulos blancos.; F: como lipoproteínas.; G: unidos a la hemoglobina.; H: combinados con glucosa.; I: en forma de colesterol.; J: unidos a la albúmina.	Los ácidos grasos libres, una vez liberados del tejido adiposo al torrente sanguíneo, son moléculas hidrofóbicas que no se disuelven bien en el plasma acuoso. Para poder transportarse de forma eficiente y segura en la sangre, se unen de manera no covalente principalmente a la albúmina sérica. No viajan como triglicéridos, colesterol ni como parte de lipoproteínas; estas formas se utilizan sobre todo para el transporte de lípidos más complejos (triglicéridos y colesterol esterificado), no de ácidos grasos libres. Tampoco se unen a eritrocitos, leucocitos, hemoglobina, glucosa ni oxígeno. Por lo tanto, la forma principal de transporte de los ácidos grasos libres en la sangre es unidos a la albúmina.	Los ácidos grasos libres (no esterificados) son moléculas hidrofóbicas, por lo que no pueden circular libremente disueltas en el plasma, que es un medio acuoso. Para poder ser transportados en sangre, deben unirse a proteínas transportadoras. En el caso específico de los ácidos grasos libres, el principal transportador en el plasma es la albúmina sérica, que posee sitios de unión para estos ácidos grasos y los lleva desde el tejido adiposo hacia otros tejidos (como músculo e hígado) donde se utilizan como fuente de energía. No se transportan mayoritariamente dentro de eritrocitos, ni unidos a hemoglobina, leucocitos, oxígeno, glucosa, ni formando triglicéridos, colesterol o lipoproteínas (estas estructuras transportan principalmente triglicéridos y colesterol, no los ácidos grasos libres como tales). Por tanto, la forma principal de transporte de los ácidos grasos libres en sangre es unidos a la albúmina.

Table 8: Example of pairwise reasoning comparison in Spanish. In both cases, A_reasoning corresponds to the GPT-5.1 baseline, while B_reasoning represents reasoning generated by MED-COREASONER with GPT-5.1 as the backbone.

Concept Extraction Prompt

Your task is to transform a reasoning trace into a concise, ordered **concept chain**.
Definitions
- "A concept" is an atomic, reusable medical idea or clinical finding that contributes to clinical reasoning. (e.g., "chronic cough", "chest X-ray", "smoker's history")
- "Concept chain" is an ordered list of these concepts that follows the original reasoning order.
Instructions
- Read the reasoning trace carefully.
- Extract key concepts in the same order as they appear
- Each concept must be concise and represent only one idea
- Prefer clinical or scientific terms over long sentences
- Do not invent new concepts that are not implied by the reasoning
- Merge duplicates or near-duplicates into one concept, but keep the order consistent
- The concept chain should be as short as possible while capturing all essential reasoning steps, and must not include or expose any answer options.
- Keep the output in the same language as the reasoning.
Output Format (plain text, no explanation):
Provide the concept chain as a list in the following format:
$[\text{"}Concept1\text{"}, \text{"}Concept2\text{"}, \text{"}Concept3\text{"}, ...]$
Now process the following reasoning trace and output only the concept chain in {language} language.
{reasoning_trace}

Figure 8: Concept Extraction Prompt

Final Answer Generation Prompt

Your task is to generate a final clinical answer for a multi-option question by integrating a **concept reasoning chain** with **the retrieved medical documents** from the concept chain and your **prior medical knowledge**.
Inputs
- Question:
{question}
- Options:
{options}
- Concept Reasoning Chain (in order):
{concept_chain}
- Referenced Context (mainly contains multiple documents, guidelines, or passages):
{context}
Instructions
1. First, carefully read the concept reasoning chain. Treat it as a DRAFT reasoning path, not as guaranteed truth.
2. Then, carefully read the referenced context. Use it to VERIFY, CORRECT, or REFINE the reasoning chain.
3. Use ONLY information that is supported by the referenced context and widely accepted medical knowledge. DO NOT directly mention the concept chain. ORGANIZE your reasoning in a clear, logical manner.
4. Finally, select the MOST APPROPRIATE option as your final answer based on the verified and refined reasoning.
5. Output the reasoning in language, regardless of the input language.
Output Format:
Return VALID JSON ONLY, following this format: {{ "reasoning": "Your verified and refined reasoning process here", "answer": "Your final answer here, e.g., A, B, C, or D" }}

Figure 9: Final Answer Generation Prompt

Judge System Prompt

You are an objective and rigorous evaluator for medical question answering.
You will be given:
- a Question
- a Ground-Truth Answer (reference)
- a Model Answer (candidate)
Your task is to evaluate the Model Answer relative to the Ground-Truth Answer.
Evaluation principles:
Prioritize factual correctness, clinical safety, and alignment with the reference.
Do NOT penalize harmless extra details if they are correct and do not contradict the reference.
Penalize contradictions, fabricated facts, or unsafe medical advice.
If the reference is brief but the model answer is longer, judge consistency and medical plausibility.
If a detail cannot be verified from the reference, treat it as uncertain rather than incorrect.
Output MUST be valid JSON only

Figure 10: The system prompt of LLM-as-a-judge in the evaluation of long-form QA task.

Judge Evaluation Prompt

Question:
{question}

Ground-Truth Answer:
{gold}

Model Answer:
{pred}

Return JSON with EXACTLY the following fields and no others:
{{
"overall_score": 1-5,
"correctness": 1-5,
"completeness": 1-5,
"safety": 1-5,
"hallucination": 1-5
}}
Scoring rules:
- 5 = excellent
- 4 = good with minor issues
- 3 = partially correct or incomplete
- 2 = major issues
- 1 = mostly incorrect or unsafe

For hallucination:
- 5 = no hallucination
- 3 = some uncertain additions
- 1 = clear hallucinations or fabricated facts

Figure 11: The evaluation prompt of LLM-as-a-judge in the evaluation of long-form QA task.