ComPLUG: Compact English-Guided Reasoning for Multilingual Medical LLMs

Anonymous ACL submission

Abstract

Robust multilingual reasoning is critical for safe and equitable AI-driven care in medicine, yet the capabilities of existing large language models (LLMs) in this setting remain largely untested. Although recent advances have enabled multilingual progress in domains like mathematics and code, it is still unknown which strategies are most effective for medical tasks. To address this gap, we conduct a comprehensive evaluation of prompting and fine-tuning methods across five languages using medical question-answering benchmarks. Our analysis yields three core findings: (1) current LLMs fall short in multilingual medical reasoning, with general-purpose models lacking domain expertise and medical models often producing English responses due to low language adherence; (2) reasoning via English pivoting substantially improves accuracy, highlighting the model's stronger reasoning capabilities in English; and (3) such approaches incur high inference cost, revealing a lack of solutions that balance performance and efficiency in multilingual clinical settings. To overcome this, we introduce ComPLUG, a compact reasoning framework that preserves the benefits of English-guided reasoning while eliminating the need to generate verbose English outputs. By encoding intermediate English reasoning into latent representations, ComPLUG reduces token length by 65.80% and inference time by 62.59% while outperforming baselines. Our work provides the first systematic analysis of multilingual medical reasoning and delivers a practical solution for real-world deployment.

1 Introduction

Multilingual reasoning refers to the ability of models to perform complex inferential tasks across different languages. Recent advances in training-free prompting techniques and cross-lingual fine-tuning have significantly improved the multilingual capabilities of large language models (LLMs) (Shi et al., 2023; Qin et al., 2023; Zhu et al., 2024b; Zhang et al., 2024b; Yong et al., 2025). However, much of this progress has centered on mathematical problem solving, with high-stakes domains such as medicine remaining comparatively underexplored. In clinical contexts, robust multilingual reasoning is crucial, as both clinicians and patients must accurately interpret and act on nuanced medical information in their native languages. Breakdowns in this process can result in miscommunication, misdiagnoses, and adverse health outcomes. Despite its importance, practical and empirically validated solutions for multilingual medical reasoning remain scarce.

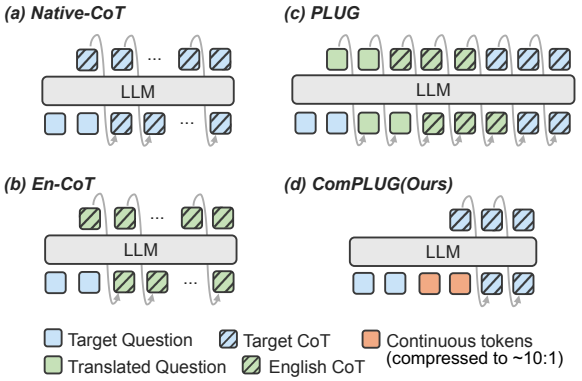

Figure 1: Approaches to multilingual reasoning. (a) Native-CoT generates chain-of-thought (CoT) reasoning directly in the user's native language but results in substantial performance degradation. (b) En-CoT generates reasoning in English, taking advantage of the model's proficiency in English reasoning, but fails to produce responses in the target language. (c) PLUG achieves high accuracy by generating English reasoning and then translating the final answer into the native language, but suffers from inefficiency due to verbose intermediate outputs. (d) ComPLUG retains the benefits of English reasoning while enhancing efficiency by encoding intermediate reasoning into compact latent representations.

To address this gap, we conduct a systematic evaluation of multilingual medical reason-

ing in LLMs, examining a diverse set of existing approaches. Specifically, we compare general-purpose multilingual models such as Mistral Nemo and Aya 101 (Sreenivas et al., 2024; Üstün et al., 2024; Dubey et al., 2024), medical–specialized models (Kim et al., 2025; Zhang et al., 2024a), prompting-based strategies, and supervised fine-tuning methods (She et al., 2024; Shen et al., 2025; Zhang et al., 2024b). Their performance is evaluated on a suite of medical licensing exam–style benchmarks developed in five languages: French, Swedish, Chinese, Japanese, and Korean (Labrak et al., 2022; Hertzberg and Lokrantz, 2024; Liu et al., 2023; Kasai et al., 2023; Kweon et al., 2024). Each benchmark is constructed in the native language of its respective country, ensuring a realistic assessment.

Our analysis yields three key findings. First, existing models remain inadequate for multilingual medical reasoning. General-purpose multilingual models lacked domain specialization and underperformed on medical benchmarks. Medical models, on the other hand, were often fine-tuned using English medical reasoning data and consequently exhibited low language adherence. Even when explicitly prompted in non-English languages, they frequently generated reasoning and responses in English, limiting their applicability in multilingual clinical settings.

Second, current LLMs perform reasoning most effectively in English, particularly in the medical domain. The Native-CoT prompting method, in which the chain of thought is presented in the user's native language, resulted in substantial performance degradation (see Figure 1a). In contrast, En-CoT, which generates reasoning in English, performed better by leveraging the model's stronger reasoning abilities in that language, although the final response was not produced in the target language (Figure 1b). This pattern was especially pronounced in the medical domain compared with areas such as mathematics and code. It also persisted in supervised settings. Fine-tuning models to generate native-language reasoning directly yielded limited gains. In contrast, English-guided supervision methods generally performed better, particularly when explicitly pivoting to English for intermediate reasoning, as demonstrated by PLUG (Figure 1c) (Zhang et al., 2024b).

Third, while PLUG stood out as the strongest baseline in our study, its limited efficiency poses a significant barrier to practical deployment. PLUG operates by first translating the input question into English, generating step-by-step reasoning in English, and then producing the final answer in the target language. Although effective, this process often results in more than twice the number of generated tokens during inference. Furthermore, PLUG performance decreases when extended to a larger multilingual setting. These limitations are particularly problematic in the medical domain, where real-world deployment often requires low-latency inference, resource-constrained environments, and support for diverse languages to ensure accessibility and equity in patient care.

Given the observed trade-offs between performance and efficiency in English-guided approaches like PLUG, we propose ComPLUG (compact pivot-language-guided generation for multilingual medical reasoning), a lightweight alternative that preserves the benefits of English reasoning while addressing its practical limitations (Figure 1d). Rather than generating explicit English reasoning during inference, ComPLUG encodes intermediate reasoning into continuous reasoning tokens. The model then generates the final answer directly in the target language, eliminating the need for verbose intermediate outputs. ComPLUG achieves accuracy comparable to PLUG, while reducing reasoning token length by 65.80% and inference time by 62.59%. Taken together, our study provides a comprehensive analysis of multilingual medical reasoning in LLMs and introduces a practical solution that balances performance and efficiency across diverse languages and clinical contexts.

The remainder of this paper is organized as follows. Section 2 reviews prior work on multilingual reasoning and medical LLMs. Section 3 describes the experimental settings and baseline methods considered in this study, providing the foundation for our subsequent evaluation. Section 4 presents our proposed ComPLUG framework, which builds on insights from these baselines. Sections 5 and 6 report experimental results and offer in-depth analyses, respectively, and Section 7 concludes the paper.

2 Related Work

2.1 Multilingual Large Language Models

Multilingual LLMs have advanced rapidly, evolving from early corpus-balanced pre-training to instruction-tuned architectures with broad language coverage. Early models such as XGLM (Lin

et al., 2021) (30 languages) and BLOOM (Le Scao et al., 2023) (46 languages) relied on balanced multilingual corpora to enable zero-shot and few-shot inference across languages, narrowing the gap between high- and low-resource settings. Subsequent instruction-tuned models, including Aya-101 (Üstün et al., 2024)—built on mT5 (Xue et al., 2021) and trained across 101 languages—and Aya-23 (Aryabumi et al., 2024), which prioritizes depth over breadth for selected languages, demonstrated strong cross-lingual performance on both generation and understanding tasks. Mistral NeMo (Sreenivas et al., 2024) supports over 100 languages and and is optimized for long-context reasoning and multi-turn dialogue.

Efforts to develop multilingual medical LLMs have only recently emerged. Unlike general multilingual LLMs, current medical models cover only a small set of widely spoken languages, reflecting the scarcity of publicly available clinical data. MMed-Llama 3 (Qiu et al., 2024), trained on the 25.5B-token MMedC corpus spanning six languages (English, Spanish, French, Russian, Chinese, Japanese), introduces MMedBench for rationale-based evaluation; its 8B variant rivals GPT-4 on multilingual medical QA. Apollo (Wang et al., 2024) builds the ApolloCorpora dataset and XMedBench benchmark, releasing 0.5B–7B parameter models that surpass language-specific variants via multilingual training or weight averaging, with notable gains in English, Chinese, French, Spanish, Arabic, and Hindi. Together, these represent the most advanced open-source multilingual medical LLMs to date and serve as the primary baselines for our study.

2.2 Multilingual Reasoning in Medicine

Chain-of-thought, translation-based, and other cross-lingual prompting or fine-tuning strategies have proven effective for multilingual reasoning in general domains, particularly in mathematical and coding tasks (Shi et al., 2023; Qin et al., 2023; Zhu et al., 2024b; Yong et al., 2025). In the medical domain, however, research remains overwhelmingly English-centric, and multilingual reasoning has been largely unexplored. Most medical LLMs, such as Meerkat (Kim et al., 2025) and UltraMedical (Zhang et al., 2024a), are trained primarily on English data and tested on English QA benchmarks (e.g., USMLE-style datasets), with minimal evaluation in other languages. Qiu et al. (2024) and Wang et al. (2024) evaluated recent LLMs on medical QA datasets in languages such as Chinese, French, and Spanish, but did not compare prompting or fine-tuning methods for multilingual reasoning, nor did they include state-of-the-art medical LLMs. Consequently, their ability to perform medical reasoning and QA beyond English remains largely unknown. Our work fills this gap by (1) broadly evaluating existing LLMs, prompting methods, and fine-tuning strategies in five non-English languages, and (2) introducing ComPLUG, a method that balances English-guided reasoning with multilingual efficiency.

2.3 Efficiency in Model Reasoning

Explicit CoT enhances reasoning accuracy but incurs substantial computational overhead, particularly as the reasoning traces become excessively long. Prior efforts to mitigate these costs have leveraged knowledge distillation (Yang et al., 2024; Chen et al., 2024; Shen et al., 2025) and latent-representation compression (Deng et al., 2023; Xia et al., 2025), which shorten reasoning traces with minimal performance loss. Building on these ideas, we tackle the additional overhead introduced by PLUG (Zhang et al., 2024b), which requires both English translations and explicit reasoning traces. To our knowledge, this constitutes the first attempt to reduce such costs in pivoting-based reasoning and the first application of this strategy to the medical domain.

3 Experimental Setup and Baseline Methods

We conducted zero-shot evaluations of general-purpose multilingual and medical-specialized LLMs, and additionally assessed supervised multilingual models. Full details of hyperparameters, computational resources, and other experimental settings are provided in Appendix A. Details of the baseline LLMs, prompting strategies, and supervised training protocols are provided in Appendices F and G.

3.1 Datasets

We used five QA benchmarks: FrenchMedMCQA for French (Labrak et al., 2022), MedQA-SWE for Swedish (Hertzberg and Lokrantz, 2024), CMExam for Chinese (Liu et al., 2023), IgakuQA for Japanese (Kasai et al., 2023), and KorMedMCQA for Korean (Kweon et al., 2024). These benchmarks are modeled after each country's medical licensing exams and are written in their native

languages. We selected medical exam–style benchmarks because they contain reasoning-intensive questions and are widely used as testbeds for evaluating LLM performance (Nori et al., 2023; Kim et al., 2025).

In addition, we include two medical conversation datasets to complement the QA benchmarks: MedDialog EN for single turn dialogue and MediTOD for multi turn dialogue. We add this dialogue component to evaluate the model's conversational ability beyond QA.

3.2 Evaluation Metrics

We evaluated model performance using two complementary metrics: (1) accuracy and (2) language adherence ratio. Accuracy served as the primary measure of correctness. However, a truly multilingual model should respond in the language of the query; a correct answer in another language undermines practical utility, making accuracy alone insufficient. To address this, we introduce the language adherence ratio, which quantifies whether model outputs remain in the specified language. Responses are tokenized with the Llama-3 tokenizer to form tri-grams, each of which is classified using a fasttext language identification model (Joulin et al., 2016a,b). For each sample, we compute the proportion of tri-grams identified as belonging to the target language, and the final language-adherence score is obtained by averaging this ratio across all samples. For the dialogue experiments, we report BLEU, ROUGE L, and BERTScore as generation metrics.

4 Proposed Method

As discussed in Section A.2, PLUG explicitly generates English outputs as intermediate steps, which introduces additional overhead. To address this limitation, we propose compressing the English pivoting process into a continuous latent representation rather than generating all intermediate tokens explicitly, as detailed in the following section.

4.1 ComPLUG

Figure 2 illustrates the training pipeline, which consists of the following three stages: (1) PLUG fine-tuning, (2) reasoning-token distillation, and (3) question-token distillation. This process gradually replaces explicit intermediate text with learned latent tokens while preserving answer distributions through self-distillation (Yang et al., 2024; Chen et al., 2024; Shen et al., 2025). Unlike PLUG, our model is trained to predict $p(a^l|q^l;\mathbf{z};\mathbf{r})$, directly generating the target-language response without first producing q^{en} or a^{en}, thereby significantly reducing generation time. Here, $\mathbf{z} = \langle z_0, \ldots, z_{l-1}\rangle$ and $\mathbf{r} = \langle r_0, \ldots, r_{k-1}\rangle$ are latent token sequences of lengths l and k, which can be kept small to minimize memory overhead.

4.1.1 PLUG Fine-tuning

This first stage serves as a warm-up phase, applying the same training objective as described in PLUG. Let $\mathbf{y} = [y_1, \ldots, y_T] = [q^{\text{en}}, a^{\text{en}}, a^l]$ denote the concatenated output sequence. Specifically, we minimize the standard autoregressive negative log-likelihood as follows:

$$\mathcal{L}_1 = -\sum_{t=1}^{T} \log p_{\theta_1}\big(y_t \mid y_{<t}, q^{\ell}\big). \tag{1}$$

The resulting model, θ_1, is then used as both teacher and student in subsequent stages.

4.1.2 Reasoning-token Distillation

The teacher model (frozen θ_1 from Stage 1) is given the full context $[q^l, q^{\text{en}}, a^{\text{en}}]$, while the student model, initialized with θ_1, receives a compressed context $[q^l, q^{\text{en}}, \mathbf{r}_n]$, where the English CoT response is replaced with latent representations. $\mathbf{r}_n$ consists of k uniquely initialized special tokens, repeated n times. We keep k small to reduce the burden of learning new tokens (e.g., $k = 15$), and instead repeat them n times so that the model can handle varying input lengths during training and n can be flexibly set by the user at inference time. The value of n is sampled based on the input question length as follows:

$$\begin{aligned} \mathbf{r}_n &= \underbrace{\langle r_0, \ldots, r_{k-1}\rangle \cdots \langle r_0, \ldots, r_{k-1}\rangle}_{n \text{ repeats}}, \\ n &= \left\lceil \gamma\, |q^l| \right\rceil, \quad \gamma \sim \mathcal{U}(\gamma_{\min}, \gamma_{\max}). \end{aligned} \tag{2}$$

Here, γ is a scaling factor sampled from $\mathcal{U}(\gamma_{\min}, \gamma_{\max})$, allowing the latent sequence length to vary with the input length during training.

The student model is trained to model the teacher's distribution while preserving exact response correctness:

$$\begin{aligned} \mathcal{L}_2 = \alpha\, \mathrm{KL}\big(p_{\theta_1}(\mathbf{y}|q^l, q^{\text{en}}, a^{\text{en}}) \,\big\|\, p_{\theta_2}(\mathbf{y}|q^l, q^{\text{en}}, \mathbf{r}_n)\big) \\ +\ (1-\alpha)\ \mathrm{CE}\big(\mathbf{y},\ p_{\theta_2}(\cdot|q^l, q^{\text{en}}, \mathbf{r}_n)\big), \end{aligned} \tag{3}$$

where KL denotes the Kullback–Leibler divergence and α is a hyperparameter.

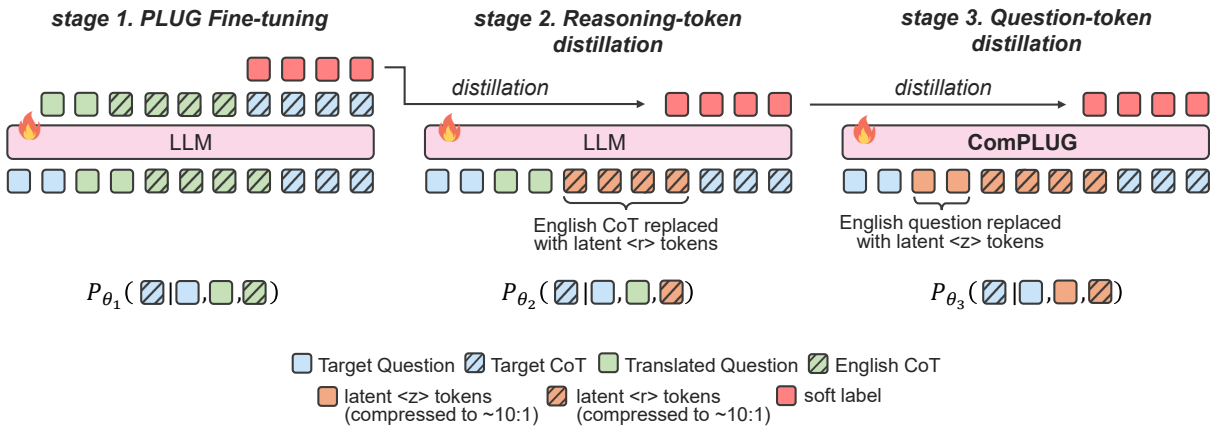

Figure 2: Training pipeline of ComPLUG based on self-distillation, comprising three stages: PLUG fine-tuning, reasoning-token distillation, and question-token distillation.

4.1.3 Question-token Distillation

Using the Stage 2 model as the teacher, we now compress the English pivot question. The input for the teacher model is $\langle q^l, q^{\text{en}}, \mathbf{r}_n \rangle$, whereas the student ($\theta_3$), initialized with θ_2, receives only $\langle q^l, \mathbf{z}_n, \mathbf{r}_n \rangle$ as follows:

$$\mathbf{z}_n = \underbrace{\langle z_0, \ldots, z_{l-1} \rangle \cdots \langle z_0, \ldots, z_{l-1} \rangle}_{n \text{ repeats}}. \tag{4}$$

The loss is computed as follows:

$$\mathcal{L}_3 = \alpha \, \mathrm{KL}\big(P_{\theta_2}(\mathbf{y} \mid q^l, q^{\text{en}}, \mathbf{r}_n) \,\big\|\, P_{\theta_3}(\mathbf{y} \mid q^l, \mathbf{z}_n, \mathbf{r}_n)\big) + (1-\alpha) \, \mathrm{CE}\big(\mathbf{y}, \; P_{\theta_3}(\cdot \mid q^l, \mathbf{z}_n, \mathbf{r}_n)\big). \tag{5}$$

4.2 Translation Model

LLM-based general-domain translators have achieved strong performance, but reliable medical-domain translation models remain unavailable. To fill this gap, we construct a domain-specialized translator by fine-tuning the open-source multilingual model, GemmaX2-28 9B (Cui et al., 2025a) to translate English content into target languages. We aggregate publicly available English–target-language parallel corpora covering 25 languages and rebalance the distribution to reduce extreme language imbalance. The resulting model accurately preserves clinical terminology while enabling the construction of high-quality multilingual training data. Detailed descriptions of the training corpora, balancing strategy, and both quantitative and qualitative evaluation results are provided in Appendix B.

5 Results

5.1 Zero-shot Performance

Table 1 summarizes the zero-shot performance of all baseline models on medical licensing exam–style benchmarks across five languages. First, existing multilingual LLMs pre-trained on large multilingual corpora remain insufficient for multilingual medical reasoning. General-domain multilingual models such as Llama-3.1 8B, Mistral-Nemo 12B, and Aya-101 13B achieve only moderate zero-shot accuracy (50.21%, 52.87%, and 51.75% in Native-CoT, respectively), and the smaller Llama-3.2 3B reaches just 38.24%. MMed-Llama-3 8B, a model continuously pre-trained on a medical multilingual corpus, performs even worse (40.81% in Native-CoT), with particularly low accuracy on low-resource languages such as Korean (24.83% in Native-CoT). This pattern suggests that medical-domain pre-training alone provides limited benefits for multilingual reasoning in medicine. In contrast, medical instruction-tuned English-centric models such as UltraMedical 8B and Meerkat 8B deliver the strongest zero-shot performance (55.84% and 59.75% in Native-CoT, respectively), highlighting the importance of domain-specific instruction tuning for reasoning.

Language-adherence scores in Table 1 provide additional perspective on model behavior. English outputs exhibit consistently high adherence, whereas responses in other target languages are markedly lower, reflecting tendency of models to generate English even when prompted in the target language. This behavior is evident in multilingual models such as Mistral-Nemo 12B and

Model (Zero-shot)	Prompting Method	Adh. (%)	Fr (Acc)	Sv (Acc)	Zh (Acc)	Ja (Acc)	Ko (Acc)	Average
Llama-3.1 8B (Dubey et al., 2024)	En-CoT	94.20	55.78	59.35	57.21	49.08	36.82	51.65
	Native-CoT	88.14	57.63	57.45	53.02	46.19	36.78	50.21
Llama-3.2 3B (Dubey et al., 2024)	En-CoT	93.55	47.80	56.68	49.55	34.01	23.51	42.31
	Native-CoT	84.67	45.17	44.09	46.50	32.69	22.76	38.24
Mistral-Nemo 12B (Sreenivas et al., 2024)	En-CoT	95.09	64.17	52.20	52.09	55.93	49.33	54.74
	Native-CoT	23.33	58.57	61.60	47.98	50.44	45.75	52.87
Aya-101 13B (Üstün et al., 2024)	En-CoT	92.39	54.91	50.63	54.36	48.04	51.28	51.84
	Native-CoT	66.20	52.21	57.94	49.06	51.95	47.57	51.75
MMed-Llama-3 8B† (Qiu et al., 2024)	En-CoT	92.16	54.31	34.57	54.60	49.03	27.82	44.07
	Native-CoT	87.71	53.27	34.18	50.28	41.50	24.83	40.81
UltraMedical 8B† (Zhang et al., 2024a)	En-CoT	43.82	54.67	58.49	59.16	55.26	53.18	56.15
	Native-CoT	67.57	60.12	59.69	56.64	51.50	51.26	55.84
Meerkat 8B† (Kim et al., 2025)	En-CoT	95.21	65.80	69.96	56.52	58.93	56.86	61.61
	Native-CoT	11.09	65.42	69.12	54.61	55.56	54.02	59.75

Table 1: Zero-shot accuracy on medical licensing exam–style benchmarks across French (Fr), Swedish (Sv), Chinese (Zh), Japanese (Ja), and Korean (Ko). *Adh.* denotes the language-adherence ratio, and † indicates a medical-specialized model.

Model (Supervised)	Adh. (%)	Fr (Acc)	Sv (Acc)	Zh (Acc)	Ja (Acc)	Ko (Acc)	Average
Zero-shot Baseline (Llama-3.1 8B + Native-CoT)	88.14	57.63	57.45	53.02	46.19	36.78	50.21
Native-CoT	88.70	61.06	62.89	46.28	40.19	32.87	48.66
MAPO⋆ (She et al., 2024)	76.41	66.51	64.37	49.45	45.04	46.16	54.31
LIDR⋆ (Yang et al., 2025)	82.45	65.23	68.92	54.60	51.32	49.49	57.91
PLUG⋆ (Zhang et al., 2024b)	88.79	68.54	71.70	52.02	59.38	63.91	63.11
PLUG (15 langs)	88.17	69.33	71.26	50.47	53.28	63.02	61.47
ComPLUG⋆	77.68	69.40	68.94	53.07	59.28	61.85	62.51
ComPLUG (15 langs)	79.24	72.10	70.56	52.37	58.88	63.51	63.48

Table 2: Accuracy of supervised models on medical licensing exam–style benchmarks across French (Fr), Swedish (Sv), Chinese (Zh), Japanese (Ja), and Korean (Ko). All models are trained on the Llama-3.1 8B backbone. ⋆ indicates models jointly trained on the five target languages, and *15 langs* models are trained jointly on 15 languages (details in the Appendix B). Native-CoT is trained separately for each language. *Adh.* denotes the language-adherence ratio.

Aya-101 13B, which achieve only 23.33% and 66.20% adherence in Native-CoT, respectively, despite their extensive multilingual pre-training. It also appears in instruct-tuned medical models like UltraMedical 8B and Meerkat 8B, reaching 67.57% and 11.09% adherence in Native-CoT, respectively. Interestingly, UltraMedical 8B shows atypically low adherence in English (43.82% in En-CoT), as it frequently produces target-language responses even when explicitly instructed to answer in English. These findings indicate achieving reliable multilingual adherence remains challenging.

Additionally, LLMs generally reason more effectively in English. Across all models, En-CoT outperforms Native-CoT on average; for example, Mistral Nemo achieves 54.74% average accuracy with En-CoT versus 52.87% with Native-CoT, while MMed-Llama-3 shows a larger gap (44.07% vs. 40.81%). This pattern reflects the models' stronger reasoning ability in English and is consistent with prior findings in other domains. We hypothesize that this effect is even more pronounced in the medical domain, which we confirm through additional comparisons with mathematics, code, and commonsense reasoning benchmarks in Appendix C, where the English advantage is indeed stronger in medical tasks. These observations reinforce that multilingual medical reasoning remains challenging, and that English-pivoted reasoning is currently the most reliable zero-shot strategy.

5.2 Supervised Performance

Table 2 presents the performance of supervised baselines. Native-CoT, which trains the model to

generate reasoning directly in the target language without any English guidance, performs poorly, reaching only 48.66% average accuracy. Its performance is slightly lower than even the zero-shot baseline (50.21%), underscoring the limited effectiveness of directly training models to produce native-language reasoning. In contrast, methods that leverage English-guided supervision—PLUG, MAPO, and LIDR—show substantial improvements, achieving 63.11%, 54.31%, and 57.91% accuracy, respectively. Among these, PLUG, which explicitly generates English reasoning before producing the target-language answer, achieves the highest accuracy but suffers from efficiency limitations. Our proposed ComPLUG reaches 62.51% average accuracy, closely matching PLUG while providing substantial efficiency benefits, as further discussed in Section 6.2.

Additionally, ComPLUG effectively learn across multiple languages within a single model. When the training is extended to a 15-language joint setting, PLUG exhibits a performance drop from 63.11% to 61.47%, whereas ComPLUG improves to 63.48%, surpassing PLUG. We hypothesize that PLUG's performance drop arises from its explicit repetition of English pivot reasoning, which may overfit in multi-language joint training. In contrast, ComPLUG compresses the pivot reasoning into a latent space, reducing redundancy and allowing the model to better leverage the diversity of 15-language supervision. Details of language-scaling experiments are in Appendix D.

Supervised fine-tuning also substantially improves language adherence while maintaining high reasoning accuracy. For example, Meerkat 8B in Native-CoT achieves 59.75% accuracy but only 11.09% language adherence (in Table 1), indicating that most responses are produced in English. In contrast, our ComPLUG (15 langs) model attains 63.48% accuracy with 79.14% adherence, striking a favorable balance between reasoning quality and faithful target-language generation.

For completeness, we also evaluate doctor-turn generation on two dialogue corpora. Multi-turn MediTOD (Saley et al., 2024) in the NLG-only setting and single-turn MedDialog-EN (Zeng et al., 2020) are summarized in Appendix E. We follow the datasets and metric protocol described in Section 3.1 and 3.2, and we report BERTScore (F1), BLEU, and ROUGE-L by language.

Model	Fr	Sv	Zh	Ja	Ko	Avg.
ComPLUG (1-lang)	70.52	68.18	52.04	57.39	61.45	61.93
w/o Distillation	64.02	61.96	50.94	53.17	52.28	56.47

Table 3: Effect of self-distillation.

6 Analysis

In this section, we present three analyses: an ablation study, an efficiency evaluation, and attention visualizations. Additional results, including qualitative reasoning examples and the effect of varying the latent-token repetition n, are provided in the Appendix F.

6.1 Effect of Self-distillation

Table 3 shows the effect of self-distillation. In this analysis, the model was trained independently for each language, as joint multilingual training can introduce cross-language interactions that obscure the effect of individual components. Removing self-distillation (ComPLUG w/o Distillation), which omits the teacher–student KL objective in Stages 2–3 and trains only with cross-entropy, reduces average accuracy by 5.46% (absolute), confirming that the KL regularizer is crucial for transferring the semantics of the compressed English pivot.

6.2 Efficiency

To evaluate accuracy–cost trade-offs, we consider two additional baselines for reasoning compression. First, *Token-Budget* (Han et al., 2024) is a training-free variant of PLUG that limits output length by prepending a length directive to the prompt, allowing the model to truncate its own chain of thought. Second, *Token-Skip* (Xia et al., 2025) trains the model to identify and skip low-perplexity tokens during reasoning, effectively retaining only the tokens most critical for producing the answer. As shown in Figure 3, Token-Budget and Token-Skip both achieve substantial reductions in generated tokens compared with PLUG but incur large drops in accuracy. By contrast, the 5-language ComPLUG model averages 397 tokens per output, a 65.80% reduction from PLUG's 1,161 tokens, while maintaining over 62% accuracy. It achieves a 10% absolute gain in accuracy over both compression baselines, while generating even fewer tokens. Throughput measurements using vLLM (Kwon et al., 2023) on an NVIDIA A100×4 show that PLUG reaches 12.48

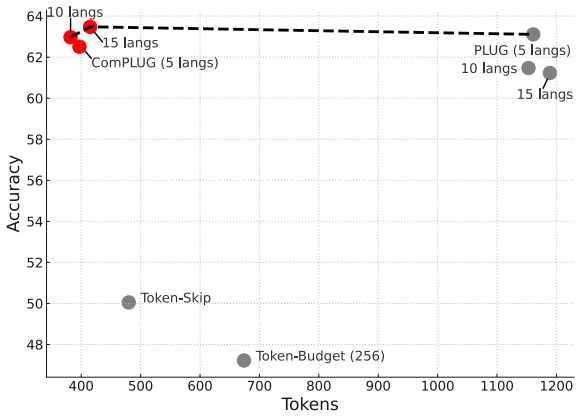

Figure 3: Accuracy–cost trade-off of PLUG, ComPLUG, and compression baselines, Token-Budget and Token-Skip.

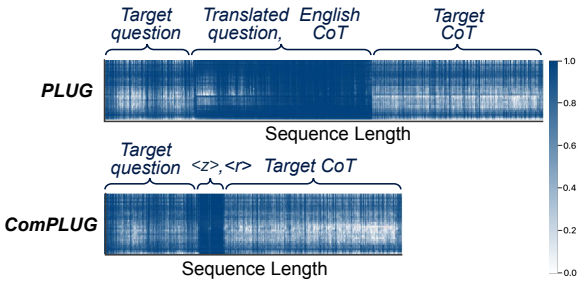

Figure 4: Self-attention heat maps for PLUG (top) and ComPLUG (bottom). Darker blue indicates higher activation on English tokens.

samples/s, whereas ComPLUG achieves 33.36 samples/s, corresponding to a 62.59% reduction in per-sample inference time. These results demonstrate that self-distilled reasoning compression in ComPLUG offers a substantially more favorable accuracy–cost trade-off for budget-constrained deployment.

6.3 Attention Patterns

Previous studies show English often serves as an internal pivot language in multilingual LLMs, with intermediate layers translating reasoning into English (Zhao et al., 2024). Our analysis reveals a complementary *sequence*-axis phenomenon (Figure 4): autoregressive attention patterns dynamically "language switch." English-biased attention dominates for English parts (q^{en} and a^{en}), shifting to the target language at q^l and a^l boundaries. This shows that multilingual reasoning is not only layer-dependent but also temporally modulated across the generated sequence, suggesting that the model internally alternates between languages as it reasons and generates the final answer.

7 Conclusion

We conducted a comprehensive investigation of multilingual medical reasoning, offering an early systematic perspective on model behavior in this setting. Evaluation across five languages shows that current models remain insufficient for reliable performance: they lack domain specialization or often revert to English responses, reflecting low language adherence. This limitation highlights an important barrier to broader deployment, where both reasoning quality and language fidelity are essential. Our analysis also reveals that English-pivoted reasoning consistently improves accuracy, but incurs substantial inference overhead that constrains practical adoption. Our ComPLUG mitigates this challenge by compressing English-guided reasoning into latent representations, retaining the benefits of pivot-language reasoning without generating verbose intermediate outputs. This design reduces token usage by 65.80% and cuts inference time by over 62.59% while maintaining strong multilingual performance. By systematically characterizing these limitations and introducing an efficient solution, this work lays the groundwork for advancing multilingual LLMs toward resource-aware applications in the medical domain.

8 Limitations

First, all evaluations rely on automatic metrics such as accuracy, BLEU, ROUGE, and BERTScore. No expert-based human evaluation has yet been conducted to assess the clinical validity, factual soundness, or linguistic appropriateness of the model outputs. Future work should include domain-expert assessment to verify the safety and interpretability of generated responses.

Second, although we carefully control data quality and evaluation settings, our analysis does not incorporate external validation in real clinical or multilingual user environments. Further studies involving real-world medical professionals or multilingual clinical datasets will be required to confirm the model's practical reliability and fairness.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, pages 6621–6642. PMLR.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025a. Multilingual machine translation with open large language models at practical scale: An empirical study. *Preprint*, arXiv:2502.02481.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025b. Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study. *Preprint*, arXiv:2502.02481.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*.

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

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2024. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*.

Niclas Hertzberg and Anna Lokrantz. 2024. Medqaswe-a clinical question & answer dataset for swedish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186.

ISO Iso. 2002. 639-1: 2002–codes for the representation of names of languages–part 1: Alpha-2 code. *International Organization for Standardization, Geneva, Switzerland*.

Shaoxiong Ji and Pinzhen Chen. 2025. How many languages make good multilingual instruction tuning? a case study on bloom. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2575–2581.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, and 1 others. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240.

Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations. *Preprint*, arXiv:2403.01469.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickaël 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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Boxiang Liu and Liang Huang. 2021. Paramed: a parallel corpus for english–chinese translation in the biomedical domain. *BMC Medical Informatics and Decision Making*, 21.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam – a comprehensive chinese medical exam dataset. *Preprint*, arXiv:2306.03030.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Vishal Saley, Goonjan Saha, Rocktim Das, Dinesh Raghu, and 1 others. 2024. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Llm pruning and distillation in practice: The minitron approach. *CoRR*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the ntcir-13 medweb task. *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-13)*.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, and 1 others. 2024. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations*.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, and 1 others. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

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

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. Plug: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *CoRR*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024a. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *arXiv preprint arXiv:2405.01345*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. In *ACL (Findings)*.

A Experimental Settings

All experiments were run on 8 × NVIDIA A100 (80 GB) GPUs using `PyTorch 2.2` and `DeepSpeed 0.14.1`. We trained for 3 epochs with the AdamW optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01), a peak learning rate of 2×10^{-5}, linear warm-up over 10% of updates, gradient accumulation 32, and mixed-precision (`bf16`) training. Unless noted otherwise, inference uses beam 4, temperature 0.0, top-p 0.95, and a max sequence length of 2048.

A.1 LLM Baselines

Experimental Setup All zero-shot evaluations were conducted on a single NVIDIA A100 GPU. The models were run with a maximum sequence length of 4096 tokens, a batch size of 1, and a top-p value of 0.95. Temperature was set to 0.6 for all generations. For the general-domain multilingual models, we used the latest instruction-tuned variants available. **Llama-3.1-8B-Instruct** and **Llama-3.2-3B-Instruct** were deployed via the official Hugging Face repositories. **Mistral-Nemo-Instruct-2407** was sourced from its official release and utilized its native prompt template. **Aya-101**, built on mT5, was also accessed through its Hugging Face model card. The medical-domain models were fine-tuned on top of existing Llama architectures. **Meerkat-8B** was initialized from Llama-3-8B and **UltraMedical-8B** from Llama-3.1-8B. **MMed-Llama-3-8B** was derived from Llama-3-8B and continuously pre-trained on a multilingual medical corpus. All models were evaluated with standard zero-shot prompting, and no in-context examples were provided.

Native CoT Inference Prompt This prompt elicits a step by step reasoning process in the same language as the input question. It directs the model to perform an analytic summary and then to conclude with a single option.

> The following is a multiple-choice question about medical knowledge.
> Solve the question in a step-by-step fashion, in the language of the given question.
>
> Final answer format:
> The answer is {option_id}.

English CoT Inference Prompt This prompt asks the model to perform reasoning in English regardless of the question's original language. It

requires a detailed summary of the available information followed by a step by step solution.

> The following is a multiple-choice question about medical knowledge.
> Solve the question in a step-by-step fashion in English.
>
> Final answer format:
> The answer is {option_id}.

A.2 Supervised Models

Our comparative analysis included a CoT fine-tuning baseline and advanced training methodologies such as PLUG, MAPO, and LIDR. All supervised models were trained on the Llama-3.1-8B-Instruct backbone. The training setup was designed to be realistic, using only English instruction-response pairs from the MedQA dataset, which were then translated into the target language using a domain-specialized medical translator. A total of 9,308 instruction-response pairs were utilized, with responses being GPT-4 generated CoT explanations.

Native-CoT This baseline represents standard supervised fine-tuning directly in the target language. The model is trained to generate reasoning and answers in the target language l by modeling the conditional probability $p(a^l|q^l)$, where q^l is the translated instruction and a^l is the corresponding CoT response.

> The following is a multiple-choice question about medical knowledge.
> Solve the question in a step-by-step fashion, starting by summarizing the available information.
>
> Final answer format:
> {answer_phrase_in_target_lang}

PLUG This model is conditioned on the target-language instruction q^l and generates a sequence consisting of the English instruction q^{en}, the English response a^{en}, and finally the target-language response a^l. The model is trained to generate the sequence $p([q^{\text{en}}; a^{\text{en}}; a^l]|q^l)$.

> The following is a multiple-choice question about medical knowledge.
> Please interpret the question in English, and then provide the final answer in {answer_lang}.
> Solve the question in a step-by-step fashion, starting by summarizing the available information.
> Output a single option from the given options as the final answer.
> You are strongly required to follow the specified output format; conclude your response with the phrase {conclude_phrase}.

MAPO This method aligns target-language CoT with English counterparts to mitigate cross-lingual reasoning inconsistencies. For each prompt, the model generates two reasoning traces, one in the target language and one in English. An off-the-shelf translator estimates symmetric translation probability to measure semantic alignment, and the trace with higher alignment is used as the preferred sample for Direct Preference Optimization (DPO).

LIDR LIDR leverages the inherent language bias of instruction-tuned LLMs towards English by generating responses in both English and the non-dominant language for a given input. The English output is established as the preferred sample, while the non-dominant-language output is the dispreferred. The model is trained with a DPO + NLL objective to increase the likelihood of the preferred response while penalizing the dispreferred one, without requiring an external translation model.

A.3 Reasoning Compression Baselines

In this section, we describe the methodological settings outlined in Section 6.2.

Token-Budget Token-Budget follows the "length-directive" recipe: we prepend *"use less than 256 tokens"* to the PLUG prompt and let the model trim its own chain of thought. This prompt encourages the model to think in a step by step fashion while obeying a strict token limit for reasoning. It specifies the maximum number of tokens to be used and requires the selection of a single answer from the provided options. The model then outputs the answer following a precise concluding phrase in the target language.

The following is a multiple-choice question about medical knowledge.
Let us think step by step and use fewer than {restrict_token_num} tokens.
You must choose only one answer from the given options.
Follow the specified output format precisely and conclude with {conclude_phrase_in_target_lang}.

Question: {question}

Answer:

Token-Skip Token-Skip keeps the original prompt but sparsifies the generated chain of thought by systematically removing every k-th token. This prompt asks the model to first perform reasoning in English and then to deliver the answer in the language of the input question. It frames the task as a step by step analysis and enforces selection of exactly one answer.

You are a helpful assistant.
The following is a multiple-choice question about medical knowledge.
Think step by step in English first, then answer with the language of the given question.
You must choose only one answer from the given options.
Final answer format: The answer is {option_id}.

User: {question}

Assistant:

B Translation

We curated translation pairs using datasets from Appendix B (refer to Table 4 for dataset statistics). 100K instances were used to finetune the base GemmaX2-28-9B-v0.1 model (Cui et al., 2025b), and the remaining 10K formed an unseen evaluation set.

Medical Parallel Corpora for Training Translation Model To construct a domain-specific training set for training the translation model, we aggregated publicly available medical parallel corpora: EMEA (Tiedemann, 2012), WMT20 (Barrault et al., 2020), ParaMed (Liu and Huang, 2021), NTCIR-13 MedWeb (Wakamiya et al., 2017), and AIHUB. [1] We evaluate MT with BLEU (Papineni et al., 2002), chrF (Popović, 2015), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005), COMET (Rei et al., 2020) score.

Language	# instances	# sentences	Source
Swedish (sv)	18,780	100,004	EMEA
Czech (cs)	19,093	100,004	EMEA
Danish (da)	19,229	100,004	EMEA
German (de)	19,581	100,003	EMEA
Greek (el)	19,575	100,002	EMEA
Spanish (es)	18,966	100,000	EMEA
Estonian (et)	18,907	100,000	EMEA
Finnish (fi)	19,057	100,003	EMEA
French (fr)	19,074	100,003	EMEA
Hungarian (hu)	19,361	100,002	EMEA
Italian (it)	19,216	100,003	EMEA
Lithuanian (lt)	19,492	100,002	EMEA
Latvian (lv)	19,114	100,000	EMEA
Maltese (mt)	18,789	100,002	EMEA
Dutch (nl)	19,071	100,004	EMEA
Polish (pl)	19,056	100,002	EMEA
Portuguese (pt)	18,864	100,003	EMEA
Romanian (ro)	18,887	100,000	EMEA
Slovak (sk)	18,999	100,006	EMEA
Slovenian (sl)	19,038	100,004	EMEA
Bulgarian (bg)	19,179	100,002	EMEA
Russian (ru)	6,029	46,782	WMT20
Chinese	62,127	62,127	Paramed
Japanese	$1,920 \times 5$	$1,920 \times 5$	NTCIR-13
Korean (ko)	100,000	100,000	AIHub

Table 4: medical parallel data

Results Table 5 shows that our finetuned model, GemmaX2-28-9B-v0.1 (Cui et al., 2025b) + FT (Ours), achieves higher scores on most evaluation metrics for both Korean and French. After training, the model was evaluated with five standard MT metrics (BLEU, ChrF, TER, METEOR, and COMET).

On Korean it improves the BLEU score from 15.96 to 24.86 and reduces TER by nearly 14 points relative to the vanilla GemmaX2 model; on French it raises BLEU from 41.41 to 48.22 and lowers TER by over 4 points. Across both languages our system matches or surpasses GPT-4o on every metric, except that GPT-4o scores slightly higher on ChrF for French, demonstrating that targeted fine-tuning with a modest amount of quality data can exceed the performance gap with a much larger proprietary model.

Qualitative example We conducted a qualitative evaluation to assess the clinical terminology preservation capabilities of our domain-specialized

[1] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=111

Model	Target	BLEU↑	ChrF↑	TER↓	METEOR↑	COMET↑
GPT−4o	Ko	13.20	41.37	78.64	31.78	92.02
	Fr	42.52	**70.56**	48.22	59.06	85.63
GemmaX2−28−9B−v0.1	Ko	15.96	42.07	73.46	33.88	91.15
	Fr	41.41	68.38	51.44	61.13	85.28
GemmaX2−28−9B−v0.1 + FT (Ours)	Ko	**24.86**	**50.95**	**59.39**	**45.12**	**92.45**
	Fr	**48.22**	70.31	**46.98**	**67.52**	**85.74**

Table 5: Machine-translation performance on Korean (Ko) and French (Fr) test sets. Higher is better for BLEU, ChrF, METEOR, and COMET (expressed here as percentages), while lower is better for TER. Best scores for each language/metric are bold.

Question:
A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98\% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient? (A) Ampicillin (B) Ceftriaxone (C) Doxycycline (D) Nitrofurantoin

Response:
The patient is presenting with symptoms consistent with a urinary tract infection (UTI), which are common during pregnancy due to physiological changes. Her symptoms include burning upon urination, and she has no signs of systemic infection, such as fever or costovertebral angle tenderness (which would suggest pyelonephritis).

When choosing a treatment for UTI in a pregnant woman, it is important to consider the safety of the medication for both the mother and the developing fetus. Therefore, the choice of antibiotic should be one that is known to be safe during pregnancy.

(A) Doxycycline is contraindicated in pregnancy due to the risk of adverse effects on fetal teeth and bone development.
(B) Nitrofurantoin is considered safe for use during most of the pregnancy but is usually avoided at the time of labor and delivery due to the risk of neonatal hemolysis.
(C) Ceftriaxone is a cephalosporin antibiotic that is generally considered safe in pregnancy.
(D) Ampicillin is also considered safe during pregnancy and is commonly used for UTIs in pregnant women.
Considering the safety profile and the common practice of choosing antibiotics for UTIs in pregnancy, the best choice for this patient would likely be either nitrofurantoin or ampicillin. However, since nitrofurantoin should be avoided at the end of pregnancy and this patient is only at 22 weeks gestation, it would be an appropriate choice. Further into the pregnancy, closer to term, ampicillin might be preferred.

Therefore, the answer is (B) Nitrofurantoin.

(a) English

Question:
임신 22주차인 23살 여성이 배뇨 시 작열감을 호소하며 병원을 찾았습니다. 증상은 1일 전부터 시작되어 물을 더 많이 마시고 크랜베리 추출물을 복용해도 악화되고 있다고 합니다. 그 외에는 건강하다고 느끼며 임신 관련하여 의사의 정기 검진을 받고 있습니다. 체온은 36.5°C(97.7°F), 혈압은 122/77 mmHg, 맥박은 분당 80회, 호흡은 분당 19회, 실내 공기 중 산소 포화도는 98%입니다. 신체 검진 결과 요각 통증은 없으며 임신한 자궁이 관찰됩니다. 다음 중 이 환자에게 가장 적합한 치료법은 무엇입니까? (A) 독시사이클린 (B) 니트로푸란토인 (C) 세프트리악손 (D) 암피실린

Response:
환자는 임신 중에 생리학적 변화로 인해 흔히 발생하는 요로감염(UTI)과 일치하는 증상을 보이고 있습니다. 그녀의 증상으로는 배뇨 시 작열감이 있으며, 열이나 요각 통증(신우신염을 시사할 수 있음)과 같은 전신 감염의 징후는 보이지 않습니다.

임산부에서 UTI 치료제를 선택할 때는 모체와 태아 모두에게 약의 안전성을 고려하는 것이 중요합니다. 따라서 선택된 항생제는 임신 중에 안전한 것으로 알려져 있어야 합니다.

(A) 독시사이클린은 태아의 치아와 뼈 발달에 미치는 부정적인 영향으로 인해 임신 중 금기입니다.
(B) 니트로푸란토인은 임신 대부분 기간 동안 안전하게 사용할 수 있지만, 신생아 용혈증 위험으로 인해 분만 시기에는 보통 피합니다.
(C) 세프트리악손은 일반적으로 임신 중 안전한 것으로 간주되는 세팔로스포린계 항생제입니다.
(D) 암피실린도 임신 중 안전한 것으로 간주되며, 임신 중 UTI를 위해 흔히 사용됩니다.

안전성 프로필과 임신 중 UTI의 항생제 선택에 대한 일반적인 관행을 고려할 때, 이 환자에게 가장 좋은 선택은 니트로푸란토인 또는 암피실린이 될 가능성이 높습니다. 그러나 니트로푸란토인은 임신 말기에는 피해야 하므로 이 환자가 임신 22주차인 점을 고려하면 적절한 선택이 됩니다. 임신이 더 진행되어 출산이 가까워지면 암피실린이 선호될 수 있습니다.

따라서 답은 (B) 니트로푸란토인입니다.

(b) GPT-4o

Question:
임신 22주인 23세 임신부가 소변을 볼 때 타는 듯한 통증이 있다고 내원했다. 증상은 하루 전부터 시작되었으며 물을 더 마시고 크랜베리 추출물을 복용했음에도 증상이 악화되고 있다고 한다. 그 외에는 건강하고 임신으로 인해 주치의의 추적관찰을 받고 있다. 체온은 97.7°F(36.5°C), 혈압은 122/77 mmHg, 맥박은 80/분, 호흡수는 19/분, 산소포화도는 실내공기에서 98%이다. 신체검사에서 늑골-척추각 압통은 없고 임신부의 자궁이 만져진다. 이 환자에게 가장 적절한 치료는 무엇인가? (A) 독시사이클린 (B) 니트로퓨란토인 (C) 세프트리아손 (D) 암피실린

Response:
이 환자는 임신 중 생리적 변화로 인해 흔히 발생하는 요로감염(UTI)과 일치하는 증상을 보이고 있다. 이 환자의 증상에는 배뇨 시 타는 듯한 통증이 포함되며, 발열이나 늑골각 압통(신우신염을 시사할 수 있는)과 같은 전신 감염 징후는 없다.

임신부의 요로감염 치료제를 선택할 때는 산모와 태아 발달에 대한 약물의 안전성을 고려하는 것이 중요하다. 따라서 항생제 선택은 임신 중 안전한 것으로 알려진 것으로 해야 한다.

(A) 독시사이클린은 태아 치아와 뼈 발달에 대한 부작용 위험으로 인해 임신 중 금기이다.
(B) 니트로퓨란토인은 임신 대부분의 기간 동안 사용이 안전한 것으로 간주되지만, 신생아 용혈 위험으로 인해 분만 시에는 대개 피한다.
(C) 세프트리악손은 임신 중 일반적으로 안전한 것으로 간주되는 세팔로스포린 계열 항생제이다.
(D) 암피실린도 임신 중 안전한 것으로 간주되며, 임신부의 요로감염에 흔히 사용된다.

안전성 프로파일과 임신 중 요로감염에 대한 항생제 선택의 일반적인 관행을 고려할 때, 이 환자에게 가장 적합한 선택은 니트로퓨란토인 또는 암피실린 중 하나일 것이다. 그러나 니트로퓨란토인은 임신 말기에는 피해야 하므로, 이 환자는 임신 22주에 불과하기 때문에 적절한 선택이 될 것이다. 임신 말기로 갈수록 암피실린을 선호할 수 있다.

따라서, 이 문제의 정답은 (B) 니트로퓨란토인입니다.

(c) Translator

Figure 5: Translation comparison for the clinical term “costovertebral angle tenderness.” Subfigure (a) shows the original English source; subfigure (b) shows GPT-4o’s incorrect rendering “lumbar angle pain”; and subfigure (c) shows our domain-specialized translator’s accurate preservation of “costovertebral angle tenderness.

translator compared to GPT-4o. A notable case demonstrates the superior accuracy of our translator in handling critical medical terminology. In translating “costovertebral angle tenderness," GPT-4o produces 요각 통증 (lumbar angle pain), which represents a significant clinical mistranslation that could lead to diagnostic confusion, as this term incorrectly suggests lower back pain rather than the specific anatomical landmark used to assess kidney pathology. In contrast, our translator accurately renders this as 늑골-척추각 압통 (costovertebral angle tenderness), precisely preserving the clinical meaning by correctly identifying the anatomical structure formed between the ribs and vertebrae. This demonstrates that our domain-specialized fine-tuning approach successfully maintains the precision of clinical terminology that is essential for accurate medical communication, whereas general-purpose models may introduce dangerous semantic shifts in specialized medical contexts.

C English CoT Effectiveness Across Domains

Multilingual reasoning in the medical domain remains relatively underexplored, leaving the performance of current models in this area largely unknown. To begin addressing this gap, we evaluate the zero-shot and fine-tuned performance of Llama-3.1-8B-Instruct, Llama-3.2-3B, and Mistral-NeMo-12B on a diverse set of medical QA datasets spanning multiple languages. We also investigate whether the medical domain presents additional challenges for multilingual reasoning compared to other domains.

Datasets For comparison, we include three additional domains–math, code, and commonsense–using MGSM (Shi et al., 2023), HumanEval-XL (Peng et al., 2024), and X-CSQA (Zhu et al., 2024a), respectively. All three datasets were originally constructed in English and translated into other languages using machine translation. For the medical domain, we translated the test set of MedQA (Jin et al., 2021) into 11 target languages using GPT-4o. Accuracy is used as the evaluation metric across all domains and languages.

Results Appendix C shows that the gap between En-CoT and Native-CoT performance is most pronounced in the medical domain. Across 11 languages, the average gap in the medical domain is 6.3 points, compared to 3.2, 1.8, and 3.6 points in the math, code, and commonsense domains, respectively. Notably, the performance gap is even larger for low-resource languages such as Arabic and Swahili, with differences of 10.8 and 13.2 points, respectively, highlighting the pressing need for robust multilingual support in medical AI systems. We also evaluated two lighter instruction tuned models, Llama-3.2-3B and Mistral-NeMo-12B, using the same protocol. Full per domain and per language results appear in Table 6.

D Scaling to Fifteen Languages

To verify the language generalisation ability of ComPLUG, we evaluate its performance as the training corpus widens from one to fifteen languages. Figure 3 presents the outcome of this scaling experiment. Following the protocol of Ji and Chen (2025), we append ten additional languages in two successive alphabetical blocks of five, ordered by their ISO 639-1 codes (Iso, 2002) as BG, CS, DA, DE, EL, then ES, ET, FI, HU, and IT, thereby introducing controlled randomness while avoiding selection bias.

PLUG delivers the strongest single-language result (66.34%), yet its average score drops by nearly five points when the model is forced to share capacity across fifteen languages. Error analysis suggests that interference between typologically distant scripts (e.g. French vs. Japanese) degrades the model’s ability to exploit the English pivot efficiently.

By contrast, ComPLUG starts slightly lower (61.92%) but improves monotonically as additional languages are introduced, reaching 63.48% at the fifteen-language setting. The latent-pivot design absorbs cross-lingual conflicts that otherwise manifest as spurious English tokens in PLUG outputs (§ 2.2), allowing ComPLUG to benefit from the expanded multilingual signal without incurring the quadratic token overhead of explicit question–reasoning generation. Taken together, these results confirm that compressed latent pivots scale more gracefully than text-level pivots, yielding a +1.9-point advantage over PLUG in the fully multilingual scenario while still generating just one-third as many tokens at inference time.

E Dialogue Generation

We complement our QA evaluation with doctor-turn generation on two dialogue corpora. For MediTOD we follow the multi-turn setting and evaluate

Model	Domain	Dataset	Prompting Method	Target Language											AVG
				En	Fr	De	Es	It	Du	Ru	Zh	Ja	Ar	Sw	
Llama-3.1-8B	Math	MGSM	En-CoT	68.9	56.0	59.7	64.2	-	-	57.9	50.1	52.7	-	38.5	54.2
			Native-CoT	68.9	51.9	54.1	60.3	-	-	58.1	49.4	46.0	-	37.0	51.0
			△	0.0	4.1	5.6	3.9	-	-	-0.2	0.7	6.7	-	1.5	3.2
	Code	HumanEval-XL (Python, pass@1)	En-CoT	52.1	51.5	54.3	48.6	49.2	48.6	51.2	52.4	51.4	43.2	-	50.0
			Native-CoT	52.1	49.5	49.7	47.2	45.6	45.2	53.5	54.0	49.6	39.7	-	48.2
			△	0.0	2.0	4.6	1.4	3.6	3.4	-2.3	-1.6	1.8	3.5	-	1.8
	Commonsense	X-CSQA	En-CoT	62.7	57.0	54.8	57.3	52.5	49.2	51.3	56.3	51.7	45.1	38.5	51.4
			Native-CoT	62.7	53.9	51.5	56.0	50.2	46.3	49.6	50.2	45.9	39.5	34.4	47.8
			△	0.0	3.1	3.3	1.3	2.3	2.9	1.7	6.1	5.8	5.6	4.1	3.6
	Medicine	MedQA	En-CoT	62.8	50.8	57.3	56.3	53.4	53.1	54.4	54.8	50.9	48.6	47.5	52.7
			Native-CoT	62.8	46.0	54.1	49.8	50.3	51.0	50.8	47.7	42.1	37.8	34.3	46.4
			△	0.0	4.8	3.2	6.5	3.1	2.1	3.6	6.9	8.8	10.8	13.2	6.3
Llama-3.2-3B	Math	MGSM	En-CoT	59.1	46.7	50.1	52.4	–	–	49.8	45.0	44.5	–	30.9	47.3
			Native-CoT	59.1	42.5	45.5	48.1	–	–	48.6	42.3	39.0	–	28.6	44.2
			△	0.0	4.2	4.6	4.3	–	–	1.2	2.7	5.5	–	2.3	3.5
	Code	HumanEval-XL	En-CoT	33.8	34.2	32.4	29.5	29.0	28.1	34.7	33.2	31.6	26.5	–	31.3
			Native-CoT	33.8	30.9	29.8	27.0	27.7	26.4	35.2	30.5	29.5	23.4	–	29.4
			△	0.0	3.3	2.6	1.6	2.3	1.7	-0.5	2.7	2.1	3.1	–	2.1
	Commonsense	X-CSQA	En-CoT	54.4	45.1	47.9	50.4	47.6	44.0	46.8	48.5	45.7	39.0	38.2	46.2
			Native-CoT	54.4	42.3	43.8	46.6	44.5	42.1	42.9	43.4	41.1	33.2	34.7	42.6
			△	0.0	2.8	4.1	3.8	2.9	1.9	3.9	5.1	4.6	5.8	3.5	3.8
	Medicine	MedQA	En-CoT	51.5	40.9	47.5	46.3	43.0	43.8	44.2	44.7	40.8	38.6	37.1	43.5
			Native-CoT	51.5	36.8	44.3	42.0	40.7	41.8	40.5	37.9	32.0	30.4	30.3	37.6
			△	0.0	4.8	4.2	3.5	3.7	2.0	3.7	6.8	7.2	8.2	6.8	5.1
Mistral-NeMo-12B	Math	MGSM	En-CoT	58.4	54.9	53.5	55.1	–	–	53.7	53.0	51.6	–	43.4	53.0
			Native-CoT	58.4	51.0	49.9	52.5	–	–	50.8	48.6	46.9	–	39.9	48.5
			△	0.0	3.9	3.6	2.6	–	–	2.9	4.4	4.7	–	3.5	3.7
	Code	HumanEval-XL	En-CoT	50.9	49.4	47.4	48.2	46.8	47.5	45.7	45.6	45.4	40.9	–	46.8
			Native-CoT	50.9	48.2	46.1	45.7	46.5	47.9	43.0	42.6	44.0	38.0	–	44.7
			△	0.0	0.8	1.3	2.5	0.3	-0.4	2.7	3.0	1.4	1.9	–	1.5
	Commonsense	X-CSQA	En-CoT	67.2	63.1	62.9	64.3	61.5	60.3	59.6	61.9	61.3	56.6	38.0	59.7
			Native-CoT	67.2	58.2	59.7	62.2	58.5	56.4	56.5	56.3	55.9	52.9	34.9	55.2
			△	0.0	4.9	3.2	2.1	3.0	3.9	3.1	5.2	5.4	3.7	3.1	3.8
	Medicine	MedQA	En-CoT	61.4	56.7	57.4	58.6	56.6	55.6	56.8	56.0	52.9	48.1	46.8	55.2
			Native-CoT	61.4	52.7	52.6	55.4	52.6	53.1	52.4	51.0	46.4	42.4	40.5	49.9
			△	0.0	4.0	4.8	2.8	4.0	2.5	4.4	5.0	6.5	5.7	6.3	4.6

Table 6: Multilingual performance (accuracy) of Llama-3.1-8B, Llama-3.2-3B, and Mistral-NeMo-12B across four reasoning domains. English data was automatically translated into each target language. For the math, code, and commonsense domains, we used previously proposed translations. For the medical domain, we translated the MedQA dataset using GPT-4o. △ indicates the performance gap between CoT reasoning in English (En-CoT) and CoT in the target language (Native-CoT). For each model and target language, the largest performance gap (△) across domains is marked in red. En: English; Fr: French; De: German; Es: Spanish; It: Italian; Du: Dutch; Ru: Russian; Zh: Chinese; Ja: Japanese; Ar: Arabic; Sw: Swahili. AVG: Average score calculated across available target languages. Performance is not reported for languages where the corresponding dataset is unavailable.

NLG only with gold actions given the dialogue history. For MedDialog-EN we construct single-turn pairs where the input is the last patient turn and the target is the next doctor turn, using conversation-level splits to prevent leakage. Preprocessing standardizes punctuation and clinical units. We report BERTScore F1, BLEU-4, and ROUGE-L per language for French, Swedish, Chinese, Japanese, and Korean, keeping decoding and tokenization consistent with the main experimental protocol.

Across models the results show consistent gains from supervised instruction tuning over prompt-only baselines, with ComPLUG matching or surpassing prior supervised systems on several languages while remaining competitive elsewhere. On MediTOD, PLUG and ComPLUG families substantially improve BLEU-4 and ROUGE-L relative to zero-shot and Native-CoT, while BERTScore F1 reflects the semantic adequacy benefits of our approach under paraphrastic variation. On MedDialog-EN, where prior work often reports prompt-only numbers, our supervised setting clarifies the gap between in-context prompting and fine-tuning. Overall the multilingual trends align with Section 5.2, indicating that the proposed compression strategy preserves linguistic adequacy in generation and transfers across languages without degrading the core QA improvements reported in the main text. See Table 7 for MediTOD and Table 8 for MedDialog-EN.

F Additional Analysis

Together with the ablation, efficiency, and attention analyses in Section 6, we report further comple-

Model	Fr	Sv	Zh	Ja	Ko
Zero-shot	0.698 / 20.07 / 0.2867	0.748 / 19.56 / 0.2900	0.621 / 14.48 / 0.2295	0.582 / 10.01 / 0.1883	0.522 / 7.89 / 0.1384
Native-CoT	0.723 / 19.35 / 0.2767	0.722 / 15.88 / 0.2358	0.627 / 11.73 / 0.1861	0.596 / 8.47 / 0.1596	0.578 / 7.56 / 0.1329
MAPO	0.788 / 26.62 / 0.3798	0.780 / 24.36 / 0.3609	0.724 / 17.49 / 0.2770	0.684 / 12.89 / 0.2423	0.559 / 11.30 / 0.1981
LIDR	0.762 / 27.05 / 0.3861	0.858 / 26.14 / 0.3874	0.733 / 21.01 / 0.3329	0.722 / 12.83 / 0.2414	0.674 / 12.03 / 0.2111
PLUG	0.862 / 31.25 / 0.4448	0.912 / 29.99 / 0.4431	0.857 / 22.68 / 0.3582	0.744 / 16.17 / 0.3032	0.713 / 14.74 / 0.2578
PLUG (15 langs)	0.878 / 35.54 / 0.5077	0.881 / 30.76 / 0.4563	0.833 / 22.78 / 0.3611	0.746 / 16.86 / 0.3173	0.697 / 13.62 / 0.2390
ComPLUG	0.904 / 30.14 / 0.4293	0.900 / 28.82 / 0.4262	0.812 / 20.24 / 0.3199	0.764 / 15.95 / 0.2993	0.660 / 12.72 / 0.2227
ComPLUG (15 langs)	0.912 / 30.35 / 0.4326	0.870 / 25.55 / 0.3781	0.838 / 21.17 / 0.3349	0.787 / 14.80 / 0.2779	0.693 / 12.44 / 0.2178

Table 7: MediTOD in-domain NLG by language. Each cell shows BERTScore / BLEU / ROUGE-L.

Model	Fr	Sv	Zh	Ja	Ko
Zero-shot	0.561 / 13.64 / 0.2745	0.506 / 12.64 / 0.2368	0.438 / 11.05 / 0.2022	0.380 / 8.00 / 0.1568	0.385 / 6.27 / 0.1085
Native-CoT	0.561 / 15.69 / 0.2748	0.552 / 12.25 / 0.2588	0.427 / 10.69 / 0.1971	0.412 / 8.09 / 0.1705	0.399 / 7.18 / 0.1125
MAPO	0.647 / 21.57 / 0.3848	0.657 / 18.77 / 0.3740	0.547 / 15.92 / 0.3071	0.503 / 12.30 / 0.2530	0.493 / 10.72 / 0.1687
LIDR	0.750 / 21.90 / 0.4185	0.766 / 20.12 / 0.4086	0.595 / 19.11 / 0.3131	0.553 / 12.23 / 0.2602	0.516 / 11.40 / 0.1657
PLUG	0.919 / 25.27 / 0.5036	0.885 / 23.05 / 0.4637	0.711 / 20.60 / 0.3676	0.660 / 15.39 / 0.3052	0.596 / 13.95 / 0.1879
PLUG (15 langs)	0.930 / 28.82 / 0.5134	0.878 / 23.72 / 0.4634	0.759 / 20.75 / 0.3948	0.702 / 16.09 / 0.3273	0.607 / 12.93 / 0.1929
ComPLUG	0.899 / 24.42 / 0.4699	0.883 / 22.20 / 0.4412	0.699 / 18.41 / 0.3448	0.609 / 15.21 / 0.2688	0.608 / 12.06 / 0.1830
ComPLUG (15 langs)	0.885 / 24.54 / 0.4633	0.909 / 19.64 / 0.4548	0.716 / 19.23 / 0.3532	0.649 / 14.09 / 0.2868	0.629 / 11.77 / 0.1893

Table 8: MedDialog-EN single-turn NLG by language. Each cell shows BERTScore / BLEU-4 / ROUGE-L.

mentary analyses.

Effect of varying the latent-token repetition In Table 9, we investigate how the latent token repetition factor n affects ComPLUG accuracy across five languages. As n increases from one to three, average accuracy rises from 49.63% to 60.32%, indicating that moderate repetition of latent tokens enhances the model's ability to encode pivot reasoning in English. However, further increasing n to four and five causes average accuracy to drop to 58.16% and 56.44%, respectively. This degradation coincides with a rapid increase in training loss, suggesting that excessive repetition induces instability in the optimization process and leads to an unstable model. These results demonstrate that although latent token repetition can effectively compress reasoning traces, there exists an optimal repetition range beyond which training dynamics deteriorate.

Repetition	fr	sv	zh	ja	ko	Avg
1	61.89	53.61	44.53	50.24	37.88	49.63
2	64.72	62.03	45.36	49.31	49.21	54.13
3	68.07	68.15	51.64	57.47	56.26	60.32
4	66.93	62.85	46.31	53.52	61.20	58.16
5	70.19	65.20	39.15	52.75	54.91	56.44

Table 9: ComPLUG accuracy (%) across five languages with varying latent-token repetition factor n, including the mean accuracy.

ComPLUG reasoning example We performed qualitative evaluation to assess the clinical reasoning quality of ComPLUG compared to baselines. The example demonstrates how ComPLUG generates coherent medical reasoning in Chinese melanoma prognosis factors. The model correctly identifies that invasion depth (Breslow thickness) is the most critical prognostic factor for melanoma, explaining how deeper invasion correlates with worse outcomes while systematically ruling out other options like margin irregularity, age presentation, and lesion changes. The response maintains clinical accuracy and appropriate medical terminology while producing the answer entirely in the target language, showcasing ComPLUG's ability to preserve both reasoning quality and language adherence without generating verbose English intermediate outputs.

Input:
一位 43 岁女性患者因前臂皮肤病变异常到您的诊所就诊。患者担心患黑色素瘤，因为她的母亲死于黑色素瘤。您认为病变需要活检，以进一步评估是否为黑色素瘤。患者担心患恶性肿瘤。黑色素瘤的最重要预后因素是什么？ (A) 非典型细胞的浸润深度 (B) 边缘的不规则水平 (C) 就诊时的年龄 (D) 病变随时间的变化 <|q_0|> ... <|r_0|> ...

Output:
一位43岁女性担心自己的皮肤病变可能是黑色素瘤，尤其是考虑到她有黑色素瘤的家族史。如果病变是恶性的，她对自己的预后担心是情理之中的。黑色素瘤的最重要预后因素是恶性细胞的侵犯深度，这一深度由Breslow厚度描述。Breslow厚度是从表皮最上层测量至最深侵犯点，它与预后直接相关：黑色素瘤细胞侵犯的深度越深，预后越差。边缘不规则的程度、就诊时的年龄和病变随时间的变化可能与恶性风险或病变性质有一定关联，但这些与预后的关联程度均不如侵犯深度。 因此，答案是 (A) 非典型细胞的浸润深度

Base LLM	Method	Fr	Sv	Zh	Ja	Ko	Average
Llama-3.1-8B	Native-CoT	61.06	62.89	46.28	40.19	32.87	48.66
	ComPLUG	72.18	**73.40**	55.79	63.36	**63.96**	**65.74**
Mistral-NeMo	Native-CoT	58.57	61.60	47.98	50.44	45.75	52.87
	ComPLUG	**72.48**	71.91	**57.05**	**64.62**	61.10	65.43
Llama-3.2-3B	Native-CoT	45.17	44.09	46.50	32.69	22.76	38.24
	ComPLUG	57.19	60.17	52.36	42.50	35.96	49.64

Table 10: Comparison of Native-CoT and ComPLUG across five languages.

ComPLUG performance comparison across base LLMs We further analyzed ComPLUG's generalizability across different backbone architectures by evaluating its performance on five non-English medical QA benchmarks (French, Swedish, Chinese, Japanese and Korean) using three diverse models (Llama-3.1-8B, Mistral-NeMo and Llama-3.2-3B), as reported in Table 10. ComPLUG consistently outperforms Native-CoT across all architectures, yielding average improvements of 17.08, 12.56 and 11.40 percentage points for the three models respectively. The most pronounced gains appear in lower-resource languages, notably Japanese and Korean, demonstrating that latent English reasoning can be harnessed to enhance target-language inference without compromising output fluency. These results confirm ComPLUG as an effective compact pivot-language strategy for strengthening multilingual medical reasoning across diverse model architectures.