

Multilingual Medical Reasoning for Question Answering with Large Language Models

Pietro Ferrazzi^{1,2}, Aitor Soroa³, Rodrigo Agerri³

¹Fondazione Bruno Kessler, Trento, Italy

²University of Padova, Italy

³HiTZ Center - Ixa, University of the Basque Country EHU

pferrazzi@fbk.eu, {a.soroa,rodrigo.agerri}@ehu.eus

Abstract

Large Language Models with reasoning capabilities have recently shown strong promise for medical Question Answering. Current approaches report competitive results on English benchmarks through knowledge distillation from large models, yet multilingual coverage remains limited. We extend this line of research by presenting a method to generate multilingual reasoning traces that enrich the distillation paradigm with factual information from Wikipedia. Using a retrieval-augmented generation pipeline over WikiProject Medicine, we produce 500k reasoning traces in English, Italian, and Spanish. The traces are generated to solve medical questions drawn from MedQA and MedMCQA, both of which we extend to Italian and Spanish. We evaluate our approach across in-domain and out-of-domain Medical QA benchmarks, demonstrating consistent performance gains in both few-shot in-context learning and supervised fine-tuning settings, achieving state-of-the-art results among 8B-parameter models across multiple model families. With the aim of fostering the development of multilingual clinical decision-support tools, we release the complete suite of resources: reasoning traces, translated datasets, Medical Wikipedia, and fine-tuned models.

1 Introduction

Medical reasoning with Large Language Models (LLMs) is a rapidly evolving area aimed at improving Natural Language Processing systems in clinical settings. Generating explicit reasoning traces, or chains of thought, has emerged as a promising approach to guide models toward more accurate answers (Wei et al., 2022). Moreover, the development of models explicitly trained to generate answers via complex, multi-step reasoning traces has marked a notable achievement for many domains (Plaat et al., 2025), while its direct application to medicine presents unique challenges:

the context is usually incomplete, concepts can be ambiguous, and there is a high degree of specialization. Such a scenario has fostered increased interest in medical reasoning, although a few important aspects remain underexplored. Recent approaches (Chen et al., 2025; Huang et al., 2025; Yu et al., 2025) have primarily relied on distilling reasoning traces from powerful models to train smaller systems, suffering from heavy dependence on synthetic data. Furthermore, multilingual reasoning has been largely neglected. Only Chen et al. (2025) proposes a bilingual approach, including English and Chinese, that relies solely on distillation from OpenAI’s models as a source of medical knowledge. While achieving good performance, this approach falls short in linking its reasoning to any sort of external medical knowledge.

In this paper, we propose a methodology to generate multilingual reasoning traces for multiple-choice medical question answering, based on the knowledge from the WikiProject Medicine¹. We define reasoning traces as answer-conditioned rationales that elaborate medical notions and knowledge in order to reach a final answer to a question. We evaluate such traces in various settings and demonstrate that training on them systematically yields state-of-the-art performance. Our contributions can be summarised as follows:

- We present the first dataset of medical reasoning traces for Italian, Spanish, and English with the potential to be extended to any language in Wikipedia.
- We provide comprehensive experiments that demonstrate the effectiveness of applying our reasoning traces to multilingual, multiple-choice medical QA via in-context learning and supervised fine-tuning.

¹https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine

MULTILINGUAL REASONING TRACES GENERATION

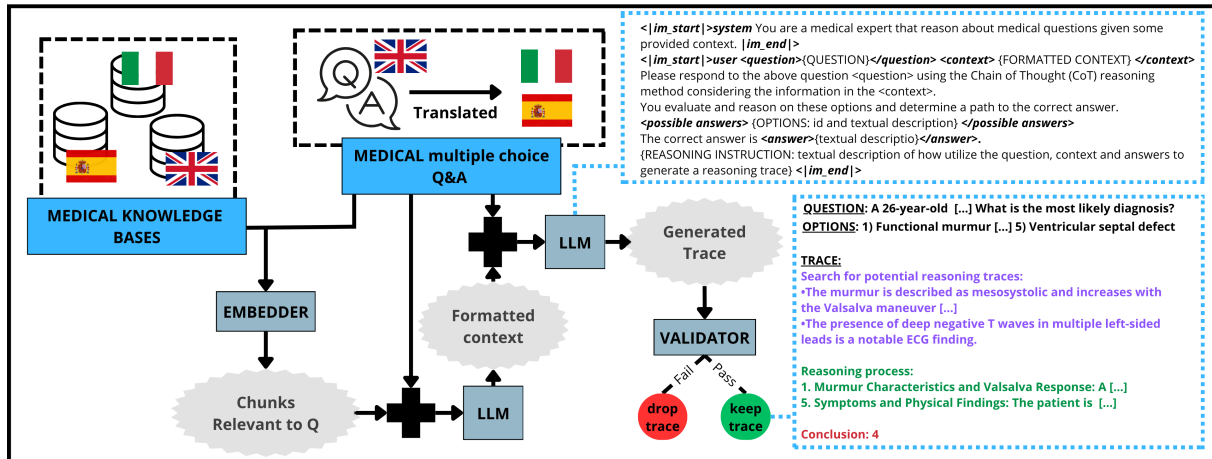


Figure 1: **Schema of the pipeline** to generate reasoning traces for multilingual multiple-choice medical question answering (QA). First, we create a Knowledge Base (KB) of medical information for each language. We extract relevant chunks from the KB for each pair of Question-Options (QO pair) in the source QA datasets, which we automatically port from English into Italian and Spanish. We prompt an LLM with the retrieved chunks and the QO pair for context formatting. Finally, we utilize the formatted context, the question, the options and the correct answer to generate a reasoning trace that answers the question itself. Answers that lead to the wrong conclusion are dropped, while the remaining form our reasoning traces dataset. The trace is composed of 3 parts. First (violet), there is a search for potential paths and useful knowledge required to answer the question, which are then used to elaborate on the provided options (green part). Finally (red), the final answer is reported as conclusion.

- We release a multilingual reasoning model that achieves state-of-the-art results in Italian, English, and Spanish medical QA.

In addition, we release two other datasets. First, a collection of all Wikipedia pages related to medicine in English, Italian, and Spanish. Second, a translated version of two common, English-based medical question answering datasets (MedMCQA and MedQA) in Italian and Spanish that we use to guide and test the reasoning traces generation pipeline. All resources are available through HuggingFace² and GitHub³.

2 Related work

Reasoning with LLMs. In the context of Large Language Models, *reasoning* refers to the ability to answer questions by complex, multi-step processes with intermediate steps (Zhang et al., 2025a), producing a chain of thought before providing the actual answer to the user. The first model explicitly designed for this purpose was OpenAI’s GPT-o1 (OpenAI et al., 2024), whose development process remains mostly undisclosed. Extensive work has

been done on replicating GPT-o1 training phases and performance (Qin et al., 2024; Huang et al., 2024; Zeng et al., 2024). As a result of such efforts, DeepSeek’s R1 (Guo et al., 2025) was among the first open-source models to achieve comparable results, similarly to QwQ (Team, 2025), marco-o1 (Zhao et al., 2024), and skywork-o1 (He et al., 2024). By achieving high performance in several benchmarks, these models have opened a line of research focused on the impact of reasoning in several domains, including math (Ahn et al., 2024), physics (Xu et al., 2025), and biology (Liu et al., 2025b).

Reasoning for the medical domain. Previous work showed how medical-oriented tuning of foundational LLMs can lead to major improvements (Luo et al., 2022; Wu et al., 2024), even surpassing human experts on some benchmarks (Singhal et al., 2025). These results highlighted that domain-specific alignment can be beneficial for medical tasks. More recently, these findings have led to the integration of domain-shift and reasoning capabilities. Huatuo (Chen et al., 2025) was among the first released models, trained via distillation on medical questions. This approach was improved via test-time scaling by Huang et al. (2025). Wu et al. (2025) introduced the grounding of the distillation process in factual medical knowledge (more details in Appendix A).

²huggingface.co/collections/NLP-FBK/multilingual-medical-reasoning

³github.com/ferrazzi Pietro/multilingual-med-reasoner

3 Methodology

We introduce a novel methodology for distilling reasoning traces that are explicitly grounded in Wikipedia medical knowledge. This involves four steps: *i*) selecting medical questions, *ii*) constructing a curated Knowledge Base (KB) of reliable medical text, *iii*) enriching each medical question with evidence retrieved from the KB using retrieval-augmented generation (Lewis et al., 2020), and *iv*) generating a step-by-step reasoning trace that leads to the correct answer among the given options with a high-capacity LLM.

3.1 Selection of Medical Questions

Datasets. The MultiMedQA benchmark introduced in the Med-PaLM paper (Singhal et al., 2023), including MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and MMLU clinical (Hendrycks et al., 2021) is the main attempt to standardize multiple-choice medical QA datasets. Following prior work (Wu et al., 2025; Chen et al., 2025), we focus our approach on MedMCQA and MedQA, which are sets of exams from India and the USA. Additionally, we include MedExpQA (Alonso et al., 2024), as it is currently the only manually validated multiple-choice medical QA dataset available for Italian, Spanish, and English. A summary of the selected datasets is provided in Appendix B (Table 6).

Dataset translation. MedMCQA and MedQA are English datasets. Given the multilingual objective of our work, we translate each Question-Options pair into Italian and Spanish using Qwen-2.5-72B (Qwen et al., 2025), prompted with a 5-shot example setup. To assess the quality of the translation, we employ back-translation: the obtained text is translated back into English and compared against the original. We compared the original and back-translated question-answer pairs by means of COMET, chrF (Popović, 2015), chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2020) metrics. Prior work has demonstrated that back-translation scores can serve as a useful proxy for translation quality and correlate with human judgments (Rapp, 2009; Zhuo et al., 2023). Furthermore, the widespread use of back-translation as a data augmentation and validation technique (Sennrich et al., 2016; Bourgeade et al., 2024; Sugiyama and Yoshinaga, 2019) supports its use for assessing translation quality in Italian and Spanish. As reported in Appendix B.2, back-translation quality is consistently

high across languages, achieving BERTScore of 96.6, COMET 88.9, chrF 78.6, and chrF++ 77.9.

As a further validation of the back-translation approach, we verify automatic metrics behavior on medical questions which translation has been manually revised (MedExpQA). We employ the method described above to translate English MedExpQA into Italian and Spanish. We compare the automatic translation to the manually revised translation using the same metrics employed for back-translation. The results (BERTscore 97, COMET 88, chrF 78, chrF++ 75) are consistent with what observed on MedQA and MedMCQA, supporting the hypothesis that these metrics indicate good alignment.

Since automatic metrics may not fully capture translation inaccuracies, we reserve the manually validated multilingual MedExpQA dataset exclusively for testing, guaranteeing out-of-domain evaluation on human-verified data.

3.2 Knowledge Base Generation

Our goal is to generate reasoning traces grounded in factual medical knowledge across the three languages. Hence, we require sources that comprehensively cover the range of medical specialties. Moreover, to ensure a fair comparison of results across languages, the underlying knowledge must be aligned across all three of them. Accordingly, the requirements for the KB are: *i*) reliability and diversity of the medical texts, and *ii*) parallel information availability across the three languages. To satisfy these conditions, and taking advantage of its open collaborative editing model and of *WikiProject Medicine*, we considered Wikipedia as our primary source of knowledge, based on Smith (2020).

Medical-Wikipedia creation. Although previous work on Medical QA has already tested Wikipedia as a medical knowledge source (Alonso et al., 2024; Xiong et al., 2024), our method to leverage Wikipedia is novel. We construct our medical knowledge base upon the WikiProject Medicine, a Wikipedia initiative that aims to collect all pages related to the medical domain. First, we collected links to all the English-relevant pages⁴. For each page, we extracted the main text, infobox content, and interlanguage links. Using the links, we retrieved the corresponding pages in Italian and Spanish. This procedure results in a multilingual Medical-Wikipedia dataset in English, Italian, and

⁴Pages listed at en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Lists_of_pages/Articles

	Number of pages (million words)		
	Source	De-duplicated	All 3 lang
en	33.859 (81)	29.306 (72)	8.577 (20)
it	11.045 (15)	8.492 (8)	8.492 (8)
es	12.886 (20)	10.124 (15)	8.513 (11)

Table 1: **Pages extracted from the Wikipedia Project** for the three languages (*Source*). *De-duplication* combines medical topics repeated in multiple links. Finally, the KB is restricted to pages available in all three languages, with slight variations in page counts, as concepts can be partitioned differently across pages in different languages.

Spanish, suitable for general NLP applications beyond medical question answering.

From Medical-Wikipedia to Knowledge Base. We employ the constructed dataset to build language-specific knowledge bases suitable for querying to retrieve context to support answering medical questions. To ensure a meaningful cross-lingual comparison, we enforce consistency across the three knowledge bases by retaining only pages available in all three languages. While this filtering step reduced the dataset size (particularly for English) it resulted in nearly parallel multilingual corpora. Although we cannot guarantee perfect alignment in page content and information density across languages, the resulting knowledge bases ensure that the same medical topics are present in all three languages. We segmented the collected pages into chunks to construct the target knowledge bases. Each section of a page is treated as a chunk. Sections exceeding 5,000 words were discarded, while sections containing fewer than 250 words were merged with the preceding ones. We filtered out irrelevant sections, such as “Bibliography” and “External links” (a comprehensive list is provided in Appendix C). An overview of the outcome of each step is presented in Table 1.

3.3 Retrieval Augmentation

Once the knowledge base is constructed, we enriched each medical question from MedQA and MedMCQA with the most relevant contextual information. Following the paradigm of Retrieval-Augmented Generation, we pre-pended the retrieved passages to the model input prior to prompting (Figure 1). We pre-compute the embedding of each chunk in the three language-specific knowledge bases, calculating their cosine similarity with each Question-Options pair. We select the

top five most similar chunks as context. Embeddings are computed using the top-performing model from the MTEB leaderboard (Enevoldsen et al., 2025) retrieval task at the time of writing (Qwen3-Embedding-8B by Zhang et al. (2025b)). To address the issue of potential redundancy and inclusion of irrelevant information in the context highlighted by Wang et al. (2024), we prompt an LLM (Qwen3-32B, Yang et al. (2025)) to rewrite the retrieved chunks. This step ensures consistency within the retrieved chunks, enhancing conciseness and avoiding duplication. The prompt template and more details on the results in Appendix B.1.

3.4 Generation of Reasoning Traces

We prompt an open-source reasoning model, Qwen3-32B, to generate reasoning traces, providing an input composed of four parts : *i*) the formatted contextual information retrieved from the KB, to provide factual knowledge on the medical topic; *ii*) the medical question that defines the problem *iii*) the answer options that constrain the space of reasoning paths; *iv*) the correct answer to guide the model toward selecting the most appropriate path. We enforce the use of the correct answer, motivated by our goal of generating the most accurate and informative traces possible. To prevent the model from collapsing into simply stating the correct answer, we design the prompt to encourage exploration of a space of potential answers before reaching the conclusion. The system prompt, with a structure based on the findings of Wu et al. (2025), is described in Appendix B.1. We verify the traces’ conclusions by extracting the answer identifier via regular expression matching and discarding the incorrect ones, as suggested by the “philosophy of verifiability” in Chen et al. (2025). Notably, this resulted in discarding less than 5% of the traces, meaning that our pipeline produces answers that usually lead to correct answers.

3.5 Multilingual Reasoning Traces Dataset

We obtain a dataset of more than 500k reasoning traces, each answering a different medical question given the options in one of the three languages. The Italian split consists of 166.257 traces from MedMCQA, 9.468 from MedQA; the Spanish split of 168.771 from MedMCQA, 9.584 from MedQA; the English split of 169.098 from MedMCQA, 9.520 from MedQA. An example of a generated trace can be seen in the Appendix, Figure 4.

model	MedExpQA			MedMCQA			MedQA			AVG
	IT	EN	ES	IT	EN	ES	IT	EN	ES	
Llama3-1B	26.3 +2.6	33.0 -1.2	27.3 +0.0	31.4 +0.6	40.2 +3.8	33.6 +2.6	27.4 +0.1	35.9 +2.5	32.0 +3.5	31.9 +1.6
Qwen3-1.7B	45.1 +8.1	51.9 +3.6	41.6 +4.7	51.5 +16.9	57.2 +19.8	53.4 +18.0	46.8 +11.3	57.2 +16.7	48.3 +13.4	50.3 +12.5
Gemma3-4B	47.8 -1.9	52.5 +11.8	52.5 -3.1	54.1 +11.4	58.7 +11.9	55.1 +12.3	50.7 +7.2	62.9 +13.4	53.5 +9.1	54.2 +8.0
MedGemma-4B	53.8 -2.7	61.8 +3.1	59.2 +5.6	56.5 +9.9	62.5 +10.7	57.9 +11.1	55.1 +4.6	67.2 +11.8	57.0 +7.1	59.0 +6.8
Llama3-8B	52.4 +3.2	69.4 +6.4	50.7 +2.8	55.7 +9.4	63.4 +6.0	58.0 +10.1	53.4 +3.0	70.9 +7.8	57.4 +7.6	59.0 +6.2
Qwen3-8B	64.7 +4.6	74.6 +8.6	64.8 +0.0	62.4 +9.3	67.4 +8.8	64.3 +10.4	67.5 +10.5	75.6 +8.5	66.5 +6.4	67.5 +7.5
Gemma3-27B	74.9 -7.0	79.4 -0.3	77.7 -4.5	67.3 +9.4	71.0 +8.2	68.4 +8.9	75.2 +6.0	79.8 +6.9	74.4 +5.5	74.2 +3.7
MedGemma-27B	76.8 -3.4	84.6 +4.5	81.9 -2.7	70.4 +8.7	76.0 +7.7	72.5 +9.8	82.4 +9.6	87.1 +9.5	81.5 +8.8	79.2 +5.8
Qwen3-32B	77.6 +0.2	80.6 -2.1	81.9 +0.8	69.6 +5.7	74.7 +5.8	70.9 +5.7	76.4 +4.1	84.3 +7.8	78.4 +4.5	77.1 +3.6
Llama3-70B	78.9 +4.9	83.8 +1.8	82.2 +5.2	71.1 +5.7	76.0 +4.2	73.3 +6.9	80.0 +14.7	85.5 +13.7	80.0 +13.7	79.0 +7.9
AVG δ	+0.9	+3.6	+0.9	+8.7	+8.7	+9.6	+7.1	+9.9	+8.0	+5.8
GPT-4o	87.0 -2.6	88.1 +2.9	89.9 +2.4	87.2 +1.1	91.2 +1.8	88.5 +0.8	75.0 -0.5	79.1 -1.4	76.0 -0.4	84.7 +0.5

Table 2: **Performances of models exposed to our traces via in-context learning** (2-shot examples, each being a question-trace-answer triplet). Accuracy is reported with deltas (δ) from baseline (2-shot examples without traces).

model	MedExpQA			MedMCQA			MedQA			AVG δ
	IT	EN	ES	IT	EN	ES	IT	EN	ES	
Llama3-8B	58.0 +9.0	65.4 +2.7	54.3 +6.4	46.8 +0.4	59.0 +1.5	50.6 +2.7	54.0 +2.3	71.6 +4.4	59.1 +6.2	57.6 +4.0
Qwen3-8B	62.2 +2.1	69.7 +3.7	69.7 +4.8	52.3 -0.7	60.9 +2.3	53.2 -0.6	61.7 +3.1	71.7 +5.4	65.0 +4.0	63.0 +2.7
AVG δ	+5.6	+3.2	+5.6	-0.1	+1.9	+1.0	+2.7	+4.9	+5.1	+3.3

Table 3: **Performances of models fine-tuned on our reasoning traces** (SFT). Accuracy is reported with deltas from the baseline (2-shot without our reasoning traces).

4 Traces Evaluation

We evaluate the usefulness of our generated reasoning traces by measuring how effectively they support Medical QA in both in-context learning (ICL) and supervised fine-tuning (SFT) settings. Our primary metric is downstream task accuracy, which directly reflects whether the traces fulfill their intended purpose. Nonetheless, in Section 5 we offer a manual evaluation of our newly generated reasoning traces by having 11 medical doctors analyze the correctness of a randomly chosen sample of *question*, *option*, *reasoning* triplets.

For both ICL (few-shot) and SFT experiments, examples are drawn from the train splits of MedQA and MedMCQA. Out-of-domain evaluation is conducted on the combined train, validation, and test splits of MedExpQA, guaranteeing out-of-distribution testing on human-validated multilingual data. An additional evaluation is also performed on the test splits of MedQA and MedMCQA (Appendix B, Table 6), guaranteeing that the models do not, at any time in the evaluation step, see the correct answers from the training split used

to generate the reasoning traces. Finally, we also report a comparison with previous state-of-the-art.

Baselines. To determine if our traces are helpful, we need to define a baseline to compare against. We adopt a few-shot evaluation setting, where each prompt includes two example questions with their options and correct answers, followed by the test question. To ensure a competitive baseline, the examples are selected through similarity search in an embedding space: for each test question, we retrieve the most similar training questions (along with their answers) to use as few-shot examples.

Statistical analysis We employ statistical tests to determine the significance of the differences in performance. For each comparison, we treat the combination of model, dataset, and language as an independent unit, representing a single instance of the treatment (i.e., exposure to the reasoning traces). We test the hypothesis that the treatment produces an effect greater than zero. Since multiple statistical tests are performed on the same data, we apply a Bonferroni correction to adjust the significance threshold.

4.1 In-Context Learning

We determine the impact of utilizing our traces via ICL. While baseline examples include only questions, options, and answers, our ICL setup also includes the reasoning traces in the 2 examples. The retrieval strategy remains identical across both settings, ensuring that any observed improvements can be solely attributed to the inclusion of our traces. We test several models from diverse families in their instructed versions and report the results in Table 2.

Results. We observe that our traces enable all models to generate more accurate answers compared to the baseline, as reported in Table 2. The average effect of our traces among datasets and languages is shown in Figure 2 (top-right). For medical questions in MedMCQA and MedQA, we observe an average increase in accuracy of +7 to +10 points across all languages. The out-of-distribution dataset (MedExpQA) is the one that benefits the least, with the case of Gemma-3-27B and Qwen-3-32B not getting any benefit. Nevertheless, the overall impact among models on this dataset is positive, with an average increase in accuracy of +1.8 points. To verify its statistical significance, we apply a paired t-test on the performance deltas between models trained with and without our traces, obtaining a p-value of 0.02, which falls below the standard significance threshold of $\alpha = 0.05$.

Disentangling reasoning from knowledge. When we expose models to our traces via ICL, we select the two examples that are the most similar to the target. These example traces therefore include medical knowledge that is required to answer the test question. By removing the similarity-based selection of the examples – and picking them randomly instead – we can estimate how much of the increase in performance is due to the reasoning capabilities acquired via ICL by itself, disentangling the effect of the similarity-based example traces.

For computational constraints, we re-run all ICL with 2 random examples on 9/10 models (excluding Llama-70B), MedQA (in-distribution), and the test set of MedExpQA (out-of-distribution)⁵. Randomly picked examples have a positive impact on 8/9 models (Llama-1B, the weakest model by a margin), with an average impact of +2.5 points (MedExpQA), and +4.7 (MedQA). This result highlights that reasoning capabilities acquired

via ICL are useful for the models, even when the traces they are exposed to are not selected to be the most relevant for the test question. Full results are reported in the Appendix (Table 10).

4.2 Supervised Fine-Tuning

To measure how our reasoning traces influence model learning, we fine-tuned models in a supervised setting. Due to computational constraints, we restricted our experiments to Llama3 and Qwen3 in their 8B versions. We trained each model on examples formatted as $\langle q \rangle \{question\} \langle /q \rangle \langle o \rangle \{options\} \langle /o \rangle \{trace\}$.

We follow Thapa et al. (2025) by using “reasoning intensive” examples for training as opposed to “knowledge intensive”, and Liu et al. (2025a), which suggests prioritizing MedQA sampling ratio over MedMCQA. Our final fine-tuning dataset comprises 5,837 traces from MedMCQA and 5,594 from MedQA per language, for a total of 34,293 examples across English, Italian, and Spanish.

We fine-tuned the models for 3 epochs on two H200 GPUs, with a per-device batch size of 32, AdamW optimizer with a learning rate of $5e-6$, a cosine scheduler, and a warm-up ratio of 0.1. Training on these settings took 1.5 hours per model.

Results. We observed that training on our traces benefits both Llama-3.1-8B-Instruct and Qwen3-8B, yielding an average accuracy improvement of +3.9 points when compared with the baseline (Table 3). The average impact is significantly higher than zero with a p-value < 0.005 . Interestingly, for *in-distribution* datasets (MedMCQA, MedQA), the gains are smaller than those seen in ICL settings. We attribute this to the nature of the traces selected as two-shot examples: they are highly relevant for the test question, acting as targeted guidance for instructed models that are capable of handling long prompts. In contrast, fine-tuning aims to generalize such knowledge, which turns out to be less effective than delivering the exact relevant information at inference time. We further analyze this behavior in the next section, showing that combining fine-tuned models with our traces in few-shot prompts leads to the best overall performance. On the other hand, the opposite is observed for *out-of-distribution* performance (MedExpQA), where SFT obtained a higher average increase among languages with respect to the baseline (+4.8 against +1.8 in ICL). This is reasonable, as SFT enables the model to internalize reasoning patterns and medical knowledge that generalize beyond the specific examples provided at

⁵While deltas for MedExpQA are comparable with what reported in Table 11, the absolute numbers are not, as here we run inference solely on the test set.

inference time. In contrast, ICL relies on retrieving highly relevant in-context examples, which may be less aligned with out-of-distribution questions.

4.3 Comparison with previous work

As described in Section 2, the main previous efforts to generate reasoning traces for medical question answering are Huatuo (Chen et al., 2025), MedReason (Wu et al., 2025), and m1-m23k (Huang et al., 2025). They released 19.704, 32.682, and 23.493 question-trace pairs (licence apache 2.0), respectively, on which they trained models using SFT and reinforcement learning techniques. We aim to compare our traces against those from previous work. First, we evaluate their effect in *i*) few-shot prompting and *ii*) supervised fine-tuning. Finally, *iii*) we evaluate each method under its strongest reported configuration. This final step goes beyond a trace-level comparison, demonstrating that our best model consistently outperforms any other alternatives.

We do not perform a comparison to the generation mechanisms presented in the literature, but to the data resources existing work released⁶. A summary of the characteristics of the traces we employ is reported in Table 4. Our traces’ length is comparable to MedReason (226 vs. 370 words on average) and substantially shorter than m1 (1123 tokens), while providing a much larger pool of ICL examples (179k vs. 32.7k for MedReason and 23.5k for m1). While m1 and Huatuo use GPT-4o for the generation, m1 and us use open-source models (DeepSeek R1, Qwen3-32B). All traces are generated from MedQA and MedMCQA. For SFT, we downsample to 34k examples, making our dataset comparable to previous monolingual approaches for English. As our traces are shorter, this corresponds to 7.8M words in English and 23M words across the three languages, comparable to m1 (26M words) and larger than Huatuo (16M words) and MedReason (12M).

Few-Shot. We test the effect of *MedReason* and *m1* traces at inference time on the same eleven models described in Section 4.1 and measure accuracy improvements, directly comparing against our traces. *Huatuo* traces could not be evaluated in this setting because they are framed as open-ended questions without multiple-choice options or

⁶We employ the reasoning traces as in the official releases in HuggingFace: UCSC-VLAA/MedReason, UCSC-VLAA/m23k-tokenized, and FreedomIntelligence/medical-o1-reasoning-SFT.

	ICL		SFT	
	n traces	avg length	n traces	tot words
MedReason	32.7 k	370	32.7 k	12.1M
m1	23.5 k	1123	23.5 k	26.4M
Huatuo-o1	-	-	40k	16.0M
OUR-en	179k	228	34k	7.8M
OUR-en+es+it	530k	226	102k	23.0M

Table 4: **Summary of the characteristics of each trace type.** Comparison of the number of traces and average trace length (words) for In-Context Learning (ICL) and the total number of traces and words for Supervised Fine-Tuning (SFT) for the different trace types. MedReason, m1, and Huatuo are English-only. Therefore, we report statistics on our approach for both English and multilingual traces. Huatuo is not employed via ICL.

unique correct answers. We observed that our traces provide the greatest average improvement among datasets and languages; overall impact in Figure 2 (top-right) (more details in Appendix D).

Fine-Tuning. We fine-tune the 8B versions of Llama3 and Qwen3 using the reasoning traces and official codebases provided by *MedReason*, *m1*, and *Huatuo*. We retained the hyperparameters specified in each work, training on two H200 GPUs for a total of approximately 18 hours. The resulting models are evaluated in our multilingual setting. Since the evaluated trace types are English-only, models trained on them tend to produce reasoning in English at inference time. To mitigate this, we explicitly prompted them to generate responses in the target language of each question, as shown in B.1. Downstream accuracy demonstrates that our traces are superior across all datasets and languages compared to other models (Figure 2, bottom-left; detailed results in D), with the sole exception of Huatuo traces applied to Qwen on MedMCQA and MedQA.

State-of-the-art performance. Finally, we compare the overall performance of the strongest models from prior work - which included both SFT and Reinforcement Learning - to our best fine-tuned model based on Qwen3-8B. The results aggregated by language and dataset are presented in Figure 2 (bottom-right), while detailed results can be found in the Appendix, Table 13. We find that our model in its basic configuration outperforms the best prior model, *Huatuo*, by an average of +2.9 points. In addition, improving our fine-tuned model with two of our traces as examples at inference time obtains the highest overall accuracy, surpassing *Huatuo* by +7 points. When considering only the English

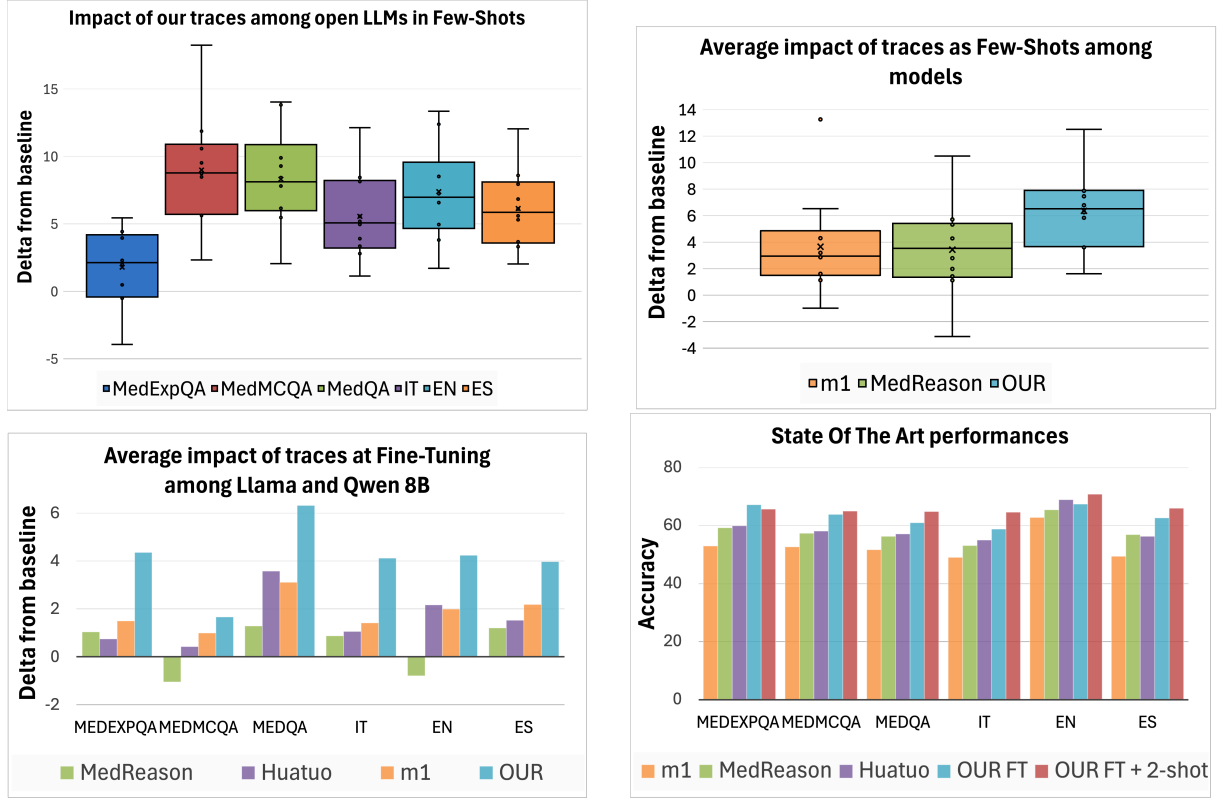


Figure 2: **Impact of our traces via In-Context Learning (ICL) and Supervised Fine-Tuning (SFT).** The **top-left** boxplots represent the improvement due to exposing the 10 open-source models to our traces as two-shot examples (ICL). The improvement is measured in terms of increased accuracy over the baseline (model prompted with two-shot, where each example is a question-answer pair without the reasoning trace). The **top-right** box-plots represent the improvement in accuracy among 10 open-source models by trace type when provided at inference time (two-shot) with respect to the baseline. Our traces are more effective than previously proposed ones. The **bottom-left** histogram shows the average impact of different types of traces when provided at fine-tuning time. The **bottom-right** histogram shows the performances of the best models proposed in the literature against our fine-tuned version of Qwen3-8B ("OUR FT") and the same model further enhanced with 2 of our traces ("OUR FT + 2-shot").

portion of the datasets, our improved model still outperforms *Huatuo* by +1.7 points.

4.4 Ablation on SFT

To better understand the sources of performance gains, we conduct an ablation study comparing three training settings: *i*) standard SFT on question-answer pairs, and *ii*) SFT on question-reasoning-answer triplets. To quantify the contribution of external medical knowledge to the reasoning traces, we generated *iii*) an additional set of reasoning traces without incorporating information from the *Medical-Wikipedia* by prompting Qwen3-32B without retrieving context from the knowledge base. This setup allows us to disentangle the effects of externally injected medical knowledge from those of reasoning capabilities distilled from Qwen3-32B alone. Due to the training computational cost, we constrain the analysis to one model

(Llama-3.1-8B-Instruct). Finally, our ablation study includes training the model individually on each language versus training it jointly on all languages simultaneously (see Appendix E for more details). As reported in Table 5, training on reasoning traces yields an average improvement of +2.7 accuracy points over standard SFT, while grounding traces in Medical-Wikipedia provides an additional +3.3 points gain, showcasing how structured reasoning and medical knowledge from Wikipedia help to improve performance.

5 Qualitative error analysis

Although aggregate accuracy metrics summarize model performance, they do not explain the underlying sources of error. To address this limitation, manual error analysis is used to identify systematic failure modes. First, we analyze a subset of 20 questions that our best-performing model (Qwen3-8B)

Configuration	MedExpQA			MedMCQA			MedQA			AVG	δ
	IT	EN	ES	IT	EN	ES	IT	EN	ES		
Llama8B+SFT	51.6	57.4	47.3	45.5	55.2	47.3	50.4	59.6	50.2	51.6	
Llama8B+SFT+reason	57.4	58.0	50.0	45.6	55.0	47.4	52.0	67.4	55.8	54.3	+2.7
Llama8B+SFT+RAG+reason	61.7	67.0	46.2	46.8	58.8	51.5	51.6	69.8	56.7	56.7	+5.1
Llama8B+SFT+RAG+reason+mutlilingual	58.0	65.4	54.3	46.8	59.0	50.6	54.0	71.6	59.1	57.6	+6.0

Table 5: Impact of different dimensions at training time. **Llama8B+SFT**: train on question-answer pairs. **Llama8B+SFT+reason**: train on question-reasoning-answer triplets, where the reasoning is distilled from Qwen32B without contextual information from Medical-Wikipedia. **Llama8B+SFT+RAG+reason**: train on question-reasoning-answer triplets for each language separately, where the reasoning is obtained by prompting Qwen32B with relevant contextual information retrieved from Medical-Wikipedia. **Llama8B+SFT+RAG+reason+multilingual** same as the previous one, but the model is trained with traces from the three languages at once.

trained on our traces answered incorrectly. We asked two medical doctors to determine which dimensions would be useful to analyze to help us understand the primary error sources. Based on this sample, they identified that issues were raised because of: *i*) incomplete or incorrect use of information provided in the question, *ii*) insufficient factual medical knowledge, and *iii*) limitations in applying logical reasoning based on that knowledge. Based on these findings, we constructed three questions to guide further analysis: *i*) *Does the answer take into account the useful elements present in the question?* *ii*) *Does the answer report medical knowledge mistakes?* *iii*) *Does the answer contain logical mistakes?*

To quantify the contribution of each error source, we randomly sampled an additional 75 question-option-reasoning triplets and reviewed them with the support of eleven medical doctors selected to cover different linguistic areas⁷. The physicians were asked to analyze triplets in their native language (English, Italian, or Spanish). For each triplet, we asked one doctor to analyse it in the light of the three defined questions. A careful analysis of the answers revealed that the primary performance bottleneck is the model’s limited medical knowledge: the most salient reasoning inconsistencies arise from difficulties in integrating relevant clinical information rather than from flawed reasoning per se. Moreover, overlooked information frequently concerns the interplay between patient characteristics and underlying medical conditions.

More specifically, this includes failure to integrate the full patient history, instead overemphasizing isolated details (Appendix F, Figure 8); difficulty in correctly applying standard diagnostic

protocols (Appendix F, Figure 6); and restricted capacity to exercise nuanced clinical judgment (Appendix F, Figure 7). Additional illustrative cases for each language are provided in Appendix F.

6 Conclusion

This paper introduces a method for generating multilingual medical reasoning traces based on knowledge extracted from the WikiProject Medicine. Our contributions are threefold: first, we present the first dataset of medical reasoning traces for English, Italian, and Spanish, generated by answering questions from MedQA and MedMCQA; second, comprehensive experiments show that exposure to these traces improves both in-context learning and supervised fine-tuning, consistently achieving state-of-the-art performance; third, we release a multilingual reasoning model alongside three new resources: a collection of 500k+ reasoning traces, translated versions of established medical QA benchmarks, and a manual evaluation by medical doctors assessing the reasoning quality of a randomly sampled subset of the generated traces.

Comparative evaluation against previous work confirms that our approach yields stronger accuracy in multilingual, multiple-choice medical question answering, with gains observed even on English-specific benchmarks. An ablation study quantifies the contributions of reasoning, factual knowledge injection, and multilingual training. A qualitative, expert-driven analysis of our best model further reveals that residual errors stem primarily from limitations in the model’s medical knowledge rather than from reasoning failures. Together, these findings underscore the value of knowledge-based multilingual reasoning traces as a scalable path toward more reliable clinical decision-support tools.

⁷Six from Italy (IT); two from Peru and one from Spain (ES); one from Canada and one from Vietnam (EN).

Limitations

Our work has some limitations. First, we focus on only three languages, namely, English, Italian, and Spanish, whereas multilingual medical reasoning could be extended to a much broader set of languages. Second, we rely on Wikipedia as our primary source of medical knowledge due to its cross-lingual parallelism, though more authoritative medical sources could further strengthen grounding. The use of Wikipedia may introduce Wikipedia-derived biases into the multilingual training, and the possible selection bias introduced by filtering traces solely based on the correctness of final answers. Furthermore, the English knowledge base is significantly larger than the others due to the nature of Wikipedia pages. Third, our fine-tuning experiments are restricted to 8B-parameter models because of computational constraints; scaling to larger models remains unexplored. Finally, we limit our study to multiple-choice medical question answering, leaving open-ended clinical reasoning tasks for future work.

Ethical considerations

Despite the promising results, our work presents some potential risks. First, reasoning traces may contain factual inaccuracies, and incomplete clinical considerations, even when grounded in external medical knowledge. This is particularly critical in healthcare, where incorrect recommendations could negatively affect clinical decision-making. Our manual analysis represents a first step to assess this dimension. Second, reasoning traces should not be interpreted as faithful explanations of the internal decision process of the model. While they may improve interpretability, they can also create a false sense of reliability or transparency.

Acknowledgments

We acknowledge the contributions of the following medical doctors who contributed to performing the manual error analysis described in Section 5: Marcello Albini, Bianca Bernante, Hong Anh Chau, Davide Dotto, Marta Dziech, Daniel García Guadix, Sara Malerba, Elena Maurin, Matteo Savio, José Luis Paredes Sosa, Luca Vavolo, Juan Luis Cuentas Zavala. This work has been partially supported by the Basque Government (Research group funding IT-1805-22), HumanAIze (AIA2025-163322-C61) funded by MICIU/AEI/10.13039/501100011033 and DeepThought (PID2024-159202OB-C21)

funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.

References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237.
- Inigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering. *Artificial Intelligence in Medicine*, 155:102938.
- Tom Bourgeade, Silvia Casola, Adel Mahmoud Wizan, and Cristina Bosco. 2024. Data augmentation through back-translation for stereotypes and irony detection. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 90–97, Pisa, Italy. CEUR Workshop Proceedings.
- Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.
- Junyang 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.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Jujie He, Tianwen Wei, Rui Yan, Jiakai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. 2024. Skywork-o1 open series.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask

- language understanding. In *International Conference on Learning Representations*.
- Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning in large language models. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.
- Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. [O1 replication journey – part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson?](#) Preprint, arXiv:2411.16489.
- 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*, 11:6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025a. [Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl](#). Preprint, arXiv:2505.17952.
- Yuyang Liu, Liuzhenghao Lv, Xiancheng Zhang, Li Yuan, and Yonghong Tian. 2025b. Bioprobench: Comprehensive dataset and benchmark in biological protocol understanding and reasoning.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, Suzhou, China. Association for Computational Linguistics.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *arXiv:2412.16720*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of Machine Learning Research*, volume 174, pages 248–260. PMLR.
- Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2025. [Multi-step reasoning with large language models, a survey](#). Preprint, arXiv:2407.11511.
- 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.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 replication journey: A strategic progress report - part 1. *CoRR*, abs/2410.18982.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). Preprint, arXiv:2412.15115.
- Reinhard Rapp. 2009. The backtranslation score: Automatic MT evaluation at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 133–136, Suntec, Singapore.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models.
- Denise A Smith. 2020. Situating wikipedia as a health information resource in various contexts: A scoping review. *PloS one*, 15(2):e0228786.
- Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China.
- Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and Yu Rong. 2025. [ReasonMed: A 370K multi-agent generated dataset for advancing medical reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26446–26467, Suzhou, China. Association for Computational Linguistics.
- Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.
- Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2025. [Disentangling reasoning and knowledge in medical large language models](#). *Preprint*, arXiv:2505.11462.
- Wenxuan Wang, Zizhan Ma, Meidan Ding, Shiyi Zheng, Shengyuan Liu, Jie Liu, Jiaming Ji, Wenting Chen, Xiang Li, Linlin Shen, and Yixuan Yuan. 2025. [Medical reasoning in the era of llms: A systematic review of enhancement techniques and applications](#). *Preprint*, arXiv:2508.00669.
- Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. [Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs](#). *Preprint*, arXiv:2504.00993.
- 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, Bangkok, Thailand. Association for Computational Linguistics.
- Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin ZHANG, Shizhe Diao, Can Yang, and Yang Wang. 2025. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. In *Forty-second International Conference on Machine Learning*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Hongzhou Yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2025. [FinemedLM-o1: Enhancing medical knowledge reasoning ability of LLM from supervised fine-tuning to test-time training](#). In *Second Conference on Language Modeling*.
- Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. [Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective](#). *Preprint*, arXiv:2412.14135.
- Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025a. From system 1 to system 2: A survey of reasoning large language models. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP:1–20.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. [Marco-o1: Towards open reasoning models for open-ended solutions](#). *Preprint*, arXiv:2411.14405.

Terry Yue Zhuo, Qionghai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337, Toronto, Canada.

A Related Work on Medical Reasoning

Chen et al. (2025) proposed *Huatuo*, among the first works to investigate the generation of reasoning traces for medical QA. *Huatuo* relies on automatic generation of verifiable problems from multiple-choice medical QA datasets. The generated traces, distilled from closed-source LLMs, are used to train models via supervised fine-tuning (SFT) and reinforcement learning (RL) in both Chinese and English. The authors found that models trained on those traces outperform their base counterparts on medical QA, even if the sole source of medical knowledge is a closed-source LLM.

Wu et al. (2025) advance beyond pure distillation by linking each question with factual medical knowledge. Such knowledge is derived from PrimeKG (Chandak et al., 2023), a knowledge graph of 17k diseases designed for medicine analysis in English. The model trained on those traces (*MedReason*) achieves slightly better performance than *Huatuo*, highlighting that medically grounded data can enhance reasoning capabilities for the English language.

Huang et al. (2025) further advanced the field focusing on the effect of test-time-scaling (Muenighoff et al., 2025). The authors train a model on a few thousand examples generated via distillation, highlighting that the quality of the data is what matters.

Other recent works follow the approach of generating traces and training on them via SFT or RL. Sun et al. (2025) build on distillation approaches with architectural adjustments for trace generation, while Yu et al. (2025) construct hundreds of thousands of synthetic medical instructions derived from Common Crawl. While innovative, this method relies heavily on LLM-generated content with limited human validation and uncertain underlying medical quality. Liu et al. (2025a) explore eliciting reasoning capabilities from models without distillation and avoiding intensive use of resources, similarly to Thapa et al. (2025), who present a method to optimize the use of existing traces. Finally, Wang et al. (2025) provide a comprehensive review of recent advances.

B Prompts and Translation

	Original Split			Our Split	
	train	val	test	train	test
MedExpQA	434	63	125	-	622
MedMCQA	182k	4k	6k	182k	4k
MedQA	10k	1.3k	1.3k	10k	1.3k

Table 6: Medical QA datasets distribution for both the original split and the one we utilize. We employed MedExpQA only for testing. Following prior work, we used the original training split of both MedQA and MedMCQA for the development of the reasoning traces, and performed the evaluation on the validation split.

B.1 Prompts

Prompt template for automatic translation. The automatic translation is performed in 5-shot. Here we report the prompt and the examples provided for each Language.

System prompt:

You are a helpful assistant that translates text from {source_language} to {target_language}. You must translate the text without adding any additional information or explanations. Never add any additional information or explanations.

User prompt:

Translate the following text to {target_language}. Return only the translated text.
text

Examples:

input: "Growth hormone has its effect on growth through?"
output it: "L'ormone della crescita esercita il suo effetto sulla crescita attraverso?"
output es: "La hormona del crecimiento tiene su efecto sobre el crecimiento a través de?"
input: "What is the main function of the liver?"
output it: "Qual è la funzione principale del fegato?"
output es: "Cuál es la función principal del hígado?"
input: "Thyroxine"
output it: "Tiroxina"
output es: "Tiroxina"
input: "IG1-1"
output it: "IG1-1"
output es: "IG1-1"
input: "Duodenal Switch"
output it: "Duodenal Switch"
output es: "Switch Duodenal"

Prompt template for context rewriting.

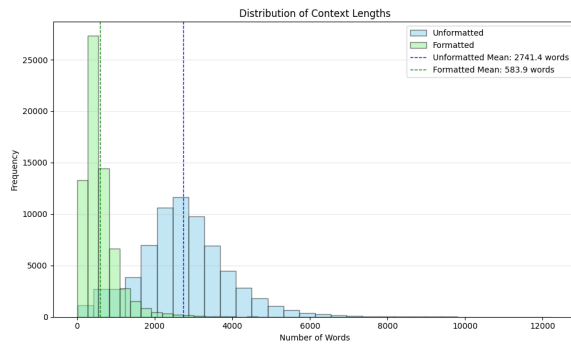


Figure 3: Length distribution of the retrieved context per question versus the LLM-formatted context. The formatted context is notably shorter than the original one.

You are a helpful assistant that rewrites medical text to be more clear and concise. Your task is to rearrange the text to improve its clarity and coherence. Return the rewritten text without saying anything else.

The formatting of the context produced the results reported in Figure 3.

Prompt template for reasoning trace generation.

System prompt:

You are a medical expert that reason about medical questions given some provided context.

User prompt:

```
<question> {question}</question>
<context>{context}</context>
Please
respond to the above question <question> using
the Chain of Thought (CoT) reasoning method,
considering the information in the <context>.
You evaluate
and reason on these options and determine a
path to the correct answer.<possible answers>
{options}
</possible answers>
The
correct answer is <answer>{answer}</answer>.
1. Explore the question. If any of the possible
answer paths seem useful, incorporate them
naturally as if you discovered them yourself.
2. Analyze and evaluate
the <possible answers> carefully, expanding
on the most relevant ones to construct
a logical, well-supported explanation.
3. Do not mention the existence of <context>
nor provided <answer> in your response.
4. Do not assume the
given <answer> is correct. Instead, determine
the answer solely based on your reasoning.
5.
If your final conclusion contradicts the given
<answer>, acknowledge potential discrepancies
(e.g., "Wait, there might be something
wrong) and refine your response accordingly.
6. The answer should
```

just report the answer's number/identifier

Output:

Search for potential reasoning
traces: (Find potentially useful reasoning
traces. It must be a concise list of traces)

Reasoning

process: (Step-by-step reasoning process)

Conclusion: (The final answer resulting from
your reasoning. Must be the identifier of the
correct option, e.g. "1", "2", "3", "4" o "5".)

Prompt for 2-shot inference (Baseline). System prompt:

You are an {language} AI
assistant that helps people find information.

User prompt:

You are an expert in the medical domain. Given
a medical <Question> and <Possible answers>,
your task is to select the correct answer.
You must return the answer identifier.
Do not provide any reasoning or explanation.
Question: {question}
Possible Answers: {options}

Prompt for 2-shot inference with reasoning traces.

System prompt:

You are an {language} AI
assistant that helps people find information.

For the user prompt, here we report the Italian version. We found that providing the output formatting instructions in the target language enforces the generation in such a language. The Spanish and English prompts are the minimal adaptation of the following:

You are an expert in the medical domain.
Given a medical <Question>
and <Possible answers>, your task is to reason
step by step as if you are independently
determining the possible reasoning
paths and deriving the correct answer.
1. Explore the question
and generate a reasoning path to answer it.
2. You
consider and reason on the <Possible answers>
to determine a path to the correct answer.
3. Analyze and evaluate
the <Possible answers> carefully, expanding
on the most relevant ones to construct
a logical, well-supported explanation.
4. Generate your own
reasoning approach based on your expertise.
Input:
Question: question
Possible Answers: options
Output:
Ricerca di potenziali tracce di ragionamento:
(trovi tracce di ragionamento potenzialmente
utili. Deve essere una lista concisa di tracce)
Processo di ragionamento:
(processo di ragionamento fatto step-by-step.)
Conclusione:

Lang	BERT	COMET	chrF	chrF++
Italian	96.5	88.7	78.0	76.3
Spanish	96.7	89.1	79.2	77.5

Table 7: Evaluation metrics for back-translation in Italian and Spanish.

(La risposta finale che deriva dal tuo reasoning. Deve essere il numero dell'opzione corretta, ad esempio "1", "2", "3", "4" o "5".)
Your answer must be in Italian.

Prompt for inference with SOTA models. Huatuo, MedReason, and m1 need to be prompted to generate the reasoning in the target language. Otherwise, they generate in English no matter what language the question is. Therefore, we designed these minimalist prompts.
Italian

Rispondi alla seguente domanda a scelta multipla in italiano:
question
{option_str}.
La risposta deve essere in italiano!

Spanish

Por favor, responda a la siguiente pregunta de opción múltiple en español:
{question}
{option_str}.
¡La respuesta debe estar en español!

B.2 Translation quality

In Table 7 we report aggregate results. In Tables 8 and 9, we report the scores for back-translation calculated for each dataset split. They are divided into “options” and “questions”, which are the two parts of the datasets we had to translate.

C Wikipedia Sections excluded by the KB

Here is the list of Wikipedia sections that have been removed from each page before creating the KB. The scope is to drop irrelevant and noisy sections.
it : [‘Note’, ‘Altri progetti’, ‘Voci correlate’, ‘Collegamenti esterni’, ‘Bibliografia’, ‘Nella cultura di massa’, ‘Storia’]

es: [‘Referencias’, ‘Referencia’, ‘Véase también’, ‘Bibliografía complementaria’, ‘Lectura adicional’, ‘Referencias y notas de pie’, ‘Enlaces externos’, ‘Historia’, ‘Bibliografía’, ‘Notas’, ‘En la cultura popular’, ‘Otras lecturas’, ‘Ver también’, ‘Bibliografía adicional’]

en: [‘References’, ‘Additional bibliography’, ‘References and footnotes’, ‘External links’, ‘See also’, ‘History’, ‘Notes’, ‘Bibliography’, ‘Cultural

references’, ‘Related articles’, ‘Related topics’, ‘Literature’, ‘Society and culture’, ‘Further reading’]

D Comparison between different traces types

Table 11 reports results on the impact of different types of traces on eleven open source models. Models are exposed to the traces via two-shot examples, where each example is composed of a question, the options, and the trace. The two examples are selected via similarity search in an embedding space, comparing the testing question-options pair against all the training question-options pairs.

Table 12 reports results on the impact of different types of traces on Qwen3-8B and Llama-3.1-8B-Instruct at fine-tuning-time.

E Multilingual vs monolingual SFT

Experiments demonstrated that integrating multilingual data during fine-tuning leads to improved accuracy on downstream tasks. To further examine this phenomenon, we analyze cross-lingual transfer effects throughout model training.

We train Qwen3-8B using reasoning traces from individual languages as well as from all three languages put together, following the training protocol outlined in Section 4.2. The results indicate that multilingual training consistently matches or surpasses the performance of single-language training. In contrast, models trained exclusively on one language exhibit worse accuracy when evaluated on other languages (see Figure 5). These findings underscore the positive impact of cross-lingual learning and emphasize the value of incorporating diverse languages into the training data.

F Error analysis

Here we report three examples of the qualitative analysis we performed, one for each language. Each example is provided together with the expert comment on it.

F.1 English

QUESTION:

Gustavo comes to the emergency room with skin lesions and general malaise of several days of evolution. He has psoriasiform lesions on the trunk with involvement of palms and soles. He also presents asymmetric non-suppurative joint inflammation and bilateral ocular redness, as well as erosions on the glans penis. In the subsequent anamnesis, Gustavo recognizes a risky sexual contact 20 days before. What is his diagnosis?

OPTIONS:

‘1’: ‘HIV infection.’, ‘2’: ‘Secondary syphilis.’, ‘3’: ‘Reiter’s syndrome.’, ‘4’: ‘Erythema multiforme.’, ‘5’: None

QUESTION A 26-year-old man with no history of interest, asymptomatic, presents on examination with a mesosystolic murmur in the aortic focus and left parasternal border that increases with the Valsalva maneuver. BP and peripheral pulses are normal and the ECG shows signs of marked left ventricular enlargement and very deep negative T waves in V3, V4, V5 and V6. What is the most likely diagnosis?	OPTIONS 1) Functional murmur 2) Aortic stenosis 3) Insufficiency 4) Hypertrophic cardiomyopathy 5) Ventricular septal defect
TRACE Search for potential reasoning traces: • The murmur is described as mesosystolic and increases with the Valsalva maneuver. • The ECG shows marked left ventricular enlargement and very deep negative T waves in V3–V6. • The patient is asymptomatic with normal BP and peripheral pulses. • The murmur is located in the aortic focus and left parasternal border. • Valsalva increases the murmur intensity, which is a key feature of certain conditions. • The presence of deep negative T waves in multiple left-sided leads is a notable ECG finding. Reasoning process: 1. Murmur Characteristics and Valsalva Response: A mesosystolic murmur that increases with the Valsalva maneuver is a key clue. This is a hallmark of hypertrophic cardiomyopathy (HOCM), as the Valsalva maneuver reduces preload, increasing the dynamic left ventricular outflow tract obstruction and intensifying the murmur. 2. ECG Findings: The ECG shows marked left ventricular enlargement and very deep negative T waves in V3–V6. These findings are consistent with hypertrophic cardiomyopathy, which is often associated with left ventricular hypertrophy and abnormal T waves. 3. Location of the Murmur: The murmur is described as being in the aortic focus and left parasternal border. While aortic stenosis is also heard in the aortic area, the Valsalva response and ECG findings are more consistent with HOCM. 4. Symptoms and Physical Findings: The patient is asymptomatic with normal BP and peripheral pulses, which is more typical of early or mild HOCM rather than severe aortic stenosis, which often presents with symptoms like syncope, angina, or dyspnea. 5. Differential Diagnosis: • Aortic stenosis typically presents with a crescendo-decrescendo murmur that peaks later in systole in severe cases, and the murmur is best heard at the right upper sternal border. The Valsalva maneuver does not typically increase the murmur in aortic stenosis. • Aortic insufficiency is a diastolic murmur, not a systolic one. • Ventricular septal defect is a holosystolic murmur, not a mesosystolic one, and is typically heard at the left lower sternal border. • Functional murmur is usually benign and not associated with ECG changes like left ventricular hypertrophy or deep T wave inversions. Conclusion: 4	

Figure 4: Example of a reasoning trace given a question and the options. The trace is composed of 3 parts. First (violet), there is a search for potential paths and useful knowledge required to answer the question, which are then used to elaborate on the provided options (green part). Finally (red), the final answer is reported as conclusion.

	Options				Questions				
Split	BERT	COMET	chrF	chrF++	BERT	COMET	chrF	chrF++	AVG
MedMCQA Test	96.9	92.6	80.3	78.5	95.5	88.3	76.9	74.5	85.4
MedMCQA Train	96.9	92.8	80.5	78.5	95.9	88.6	76.3	73.9	85.4
MedMCQA Val	96.8	92.0	78.6	76.8	96.0	88.0	75.7	73.4	84.7
MedQA Test	97.7	94.1	84.9	83.5	95.7	80.4	71.1	70.0	84.7
MedQA Train	97.7	94.1	84.7	83.3	95.7	79.9	71.0	69.8	84.5
MedQA Val	97.7	94.1	84.6	83.1	96.0	80.9	72.7	71.5	85.1
Micro AVG	97.0	93.0	82.0	81.0	96.0	84.0	74.0	72.0	

Table 8: Back-translation scores on Italian MedMCQA and MedQA (options and questions). Reported scores are BERTScore, COMET, chrF, and chrF++. The per-column average is weighted by the number of examples per dataset split.

CORRECT ANSWER:

3

MODEL ANSWER:

Search for potential reasoning traces: - Psoriasiform lesions on the trunk with involvement of palms and soles. - Asymmetric non-suppurative joint inflammation. - Bilateral ocular redness. - Erosions on the glans penis. - History of risky sexual contact 20 days before. - These features suggest a sexually transmitted infection with a reactive or disseminated phase. - Secondary syphilis is known to present with a rash, mucocutaneous lesions, and systemic symptoms. - Reiter's syndrome is a reactive arthritis that follows a genitourinary infection but does not typically present with psoriasiform lesions or mucocutaneous erosions. - Erythema multiforme is a hypersensitivity reaction and does not typically involve the glans penis or joints. - HIV infection may present with systemic symptoms but does not typically include the described mucocutaneous and ocular features.

Reasoning process: 1. Gustavo presents with a combination of skin lesions, joint inflammation, ocular redness, and

mucocutaneous erosions. 2. The psoriasiform rash on the trunk and palms/soles is a key feature of secondary syphilis. 3. The mucocutaneous erosion on the glans penis is also consistent with secondary syphilis. 4. The bilateral ocular redness and asymmetric joint inflammation are additional features of secondary syphilis. 5. The history of a risky sexual contact 20 days before is consistent with the incubation period of syphilis. 6. Reiter's syndrome is a reactive arthritis that follows a genitourinary infection but does not typically present with psoriasiform lesions or mucocutaneous erosions. 7. Erythema multiforme is a hypersensitivity reaction and does not typically involve the glans penis or joints. 8. HIV infection may present with systemic symptoms but does not typically include the described mucocutaneous and ocular features.

Conclusion: 2

Expert Comment. The model makes a critical medical knowledge error regarding Reiter's syndrome. It incorrectly states that Reiter's does

	Options				Questions				
Split	BERT	COMET	chrF	chrF++	BERT	COMET	chrF	chrF++	AVG
MedMCQA Test	97.3	93.7	82.3	80.5	95.6	88.9	79.0	76.5	86.7
MedMCQA Train	97.2	93.7	82.6	80.7	96.2	89.3	80.0	77.6	87.2
MedMCQA Val	97.2	93.1	80.8	79.1	96.2	88.6	78.1	75.8	86.1
MedQA Test	98.0	94.9	87.0	85.7	97.4	83.7	81.2	80.3	88.5
MedQA Train	98.0	94.7	86.6	85.3	97.5	83.9	82.1	81.1	88.7
MedQA Val	98.1	94.8	86.7	85.3	97.6	84.4	82.8	81.9	89.0
Micro AVG	98.0	94.0	84.0	83.0	97.0	86.0	80.0	79.0	

Table 9: Back-translation scores on Spanish MedMCQA and MedQA (options and questions). Reported scores are BERTScore, COMET, chrF, and chrF++. The per-column average is weighted by the number of examples per dataset split.

	MedExpQA			MedQA			AVG
model	EN	IT	ES	EN	IT	ES	
Llama-3.2-1B-Instruct	28.7 -4.8	20.2 -8.5	28.6 -2.2	31.7 -2.4	21.4 -8.8	25.3 -4.6	26.0 -5.2
Qwen3-1.7B	40.4 +6.9	33.9 +4.6	25.4 -0.7	50.4 +15.0	35.3 +3.1	37.8 +4.4	37.2 +5.6
gemma-3-4b-it	50.0 +0.0	39.6 -2.4	47.4 +9.1	51.3 +4.6	45.0 +3.9	46.7 +5.3	46.7 +3.4
medgemma-4b-it	57.4 +5.9	48.1 +3.4	48.6 +2.4	62.3 +8.1	48.1 +1.2	49.0 +0.9	52.3 +3.6
Llama-3.1-8B-Instruct	62.8 +5.9	44.8 -0.9	44.9 -3.0	66.1 +5.5	46.8 +0.0	49.2 +0.8	52.4 +1.4
Qwen3-8B	73.0 +9.7	59.0 +7.4	59.5 +1.5	69.1 +6.6	58.1 +5.6	59.2 +4.5	63.0 +5.9
gemma-3-27b-it	77.1 +4.8	74.3 +5.2	74.6 -1.5	77.7 +8.8	70.9 +6.7	70.6 +6.4	74.2 +5.1
medgemma-27b-text-it	81.4 +5.9	79.8 +6.4	78.9 +2.3	84.3 +12.5	78.2 +11.2	79.1 +11.1	80.3 +8.2
Qwen3-32B	80.9 +4.8	77.0 +3.6	77.3 +0.7	81.1 +6.5	74.6 +4.8	76.0 +5.6	77.2 +4.3
AVG δ	+4.3	+2.1	+1.0	+7.2	+3.1	+3.8	+3.6

Table 10: Performances of models exposed to our traces via in-context learning with **random sampling** (2-shot examples, each being a question-trace-answer triplet). Accuracy is reported with deltas (δ) from baseline (2-shot examples without traces).

not typically present with psoriasiform lesions or mucocutaneous erosions. In fact, these features are classic characteristics of the syndrome (known as keratoderma blennorrhagicum and circinate balanitis). This fundamental mistake led the model to incorrectly rule out the correct diagnosis.

F.2 Italian

QUESTION: Un uomo di 30 anni è caduto in piedi da un'altezza di 2 metri. All'arrivo al Pronto Soccorso era supino e aveva cateterizzato la vescica con urine chiare. Presenta un gonfiore al tallone sinistro, un leggero formicolio alla parte anteriore della coscia destra e allo scroto e disestesia a entrambi i piedi. Il paziente saluta il barelliere con una stretta di mano e una contemporanea pacca sulla spalla. La rotazione del MMII sulla barella afferrata dalle caviglie è indolore e la compressione del bacino è asintomatica. Il prossimo:

OPZIONI:

'1': 'Il paziente viene fatto sedere sulla barella in modo da poter esplorare il rachide cervicale.', '2': 'Fare riferimento alla

Rx per lo studio degli arti inferiori.', '3': 'Palpare l'addome e ruotarlo in blocco per palpare la colonna vertebrale toracolombare.', '4': 'Il paziente verrà sottoposto a un collare cervicale e inviato in radiologia per la diagnostica per immagini del rachide cervicale e degli arti inferiori.', '5': 'Verrà richiesta un'ecografia addominale.'

RISPOSTA CORRETTA:

3

RISPOSTA DEL MODELLO:

Ricerca di potenziali tracce di ragionamento: - Il paziente ha subito una caduta da 2 metri in piedi, con impatto potenzialmente traumatico. - All'arrivo al Pronto Soccorso presenta sintomi neurologici: formicolio alla coscia destra, scroto e disestesia ai piedi. - La stretta di mano e la pacca sulla spalla suggeriscono una funzionalità motoria conservata, ma non esclude una lesione spinale. - Il gonfiore al tallone sinistro potrebbe indicare una frattura o una contusione, ma non spiega i sintomi neurologici. - La manovra di rotazione del MMII e la compressione del bacino sono asintomatiche, suggerendo una possibile assenza di lesione spinale a livello lombare o sacrale. - I sintomi neurologici (formicolio, disestesia) suggeriscono una possibile lesione spinale a livello toracico o cervicale. - La

model	MedExpQA			MedMCQA			MedQA		
	IT	EN	ES	IT	EN	ES	IT	EN	ES
Llama-3.2-1B-Instruct	23.7	34.2	27.3	30.8	36.5	31.1	27.3	33.3	28.5
+ m23k	-2.9	-1.4	-8.1	-1.3	+0.4	-0.5	-0.5	+5.0	+0.2
+ MedReason	-12.5	-11.0	-8.1	-12.1	-7.1	-8.7	-10.4	-1.0	-8.8
+ OUR	+2.6	-1.2	+0.0	+0.6	+3.8	+2.6	+0.1	+2.5	+3.5
Qwen3-1.7B	37.0	48.3	37.0	34.5	37.4	35.4	35.5	40.5	34.9
+ m23k	+15.0	+9.3	+15.8	+9.1	+9.9	+9.2	+16.3	+16.7	+18.0
+ MedReason	-0.2	+4.5	+11.0	+7.5	+11.7	+8.0	+14.9	+19.3	+18.1
+ OUR	+8.1	+3.6	+4.7	+16.9	+19.8	+18.0	+11.3	+16.7	+13.4
gemma-3-4b-it	49.7	50.7	55.6	42.7	46.8	42.8	43.5	49.5	44.5
+ m23k	-4.1	-0.3	-7.6	+6.6	+5.1	+6.5	+6.5	+9.1	+6.9
+ MedReason	+5.5	+6.9	-5.2	+4.1	+5.5	+6.0	+10.8	+12.3	+11.8
+ OUR	-1.9	+11.8	-3.1	+11.4	+11.9	+12.3	+7.2	+13.4	+9.1
medgemma-4b-it	56.5	58.8	53.6	46.6	51.8	46.7	50.5	55.5	49.9
+ m23k	-6.1	+4.4	+5.6	+8.6	+7.9	+8.0	+8.8	+11.2	+10.2
+ MedReason	-3.7	-17.2	-8.8	-0.9	-8.1	-2.9	+4.8	-3.8	+4.7
+ OUR	-2.7	+3.1	+5.6	+9.9	+10.7	+11.1	+4.6	+11.8	+7.1
Llama-3.1-8B-Instruct	49.2	63.0	47.9	46.4	57.5	47.9	50.4	63.1	49.7
+ m23k	-2.8	+0.2	+4.9	+2.7	+2.1	+4.8	+2.4	+5.7	+5.7
+ MedReason	+4.4	-1.4	+12.9	+1.4	+1.1	+2.3	+7.3	+5.6	+8.6
+ OUR	+3.2	+6.4	+2.8	+9.4	+6.0	+10.1	+3.0	+7.8	+7.6
Qwen3-8B	60.1	66.0	64.8	53.0	58.6	53.8	57.0	67.1	60.1
+ m23k	-0.1	+2.0	-3.2	+5.5	+4.4	+4.5	+5.4	+3.8	+4.9
+ MedReason	+3.9	+5.2	+0.8	+3.7	+3.9	+4.6	+11.8	+10.3	+9.7
+ OUR	+4.6	+8.6	+0.0	+9.3	+8.8	+10.4	+10.5	+8.5	+6.4
gemma-3-27b-it	81.9	79.7	82.2	57.9	62.8	59.4	69.2	72.9	68.9
+ m23k	-5.9	-3.7	-7.8	+2.7	+5.0	+5.3	+7.1	+5.4	+6.2
+ MedReason	-7.5	-5.3	-6.2	+6.4	+7.5	+6.8	+4.2	+4.5	+3.7
+ OUR	-7.0	-0.3	-4.5	+6.0	+6.9	+5.5	+9.4	+8.2	+8.9
medgemma-27b-text-it	80.2	80.1	84.6	61.7	68.3	62.7	72.8	77.7	72.7
+ m23k	-1.8	-7.3	-6.2	+5.9	+2.2	+6.3	+5.7	+0.3	+4.9
+ MedReason	-5.8	-0.9	-8.6	+6.7	+9.5	+2.7	+1.1	+4.2	+1.2
+ OUR	-3.4	+4.5	-2.7	+9.6	+9.5	+8.8	+8.7	+7.7	+9.8
Qwen3-32B	77.4	82.7	81.1	63.9	69.0	65.1	72.3	76.5	73.9
+ m23k	-4.6	+2.1	-1.9	+3.3	+3.9	+2.8	+3.6	+2.5	+2.9
+ MedReason	+1.0	-4.3	+2.1	+4.9	+7.0	+6.2	+1.8	+1.9	+0.6
+ OUR	+0.2	-2.1	+0.8	+4.1	+7.8	+4.5	+5.7	+5.8	+5.7
Llama-3.3-70B-Instruct	74.0	82.0	77.0	65.3	71.8	66.4	65.3	71.8	66.4
+ m23k	+3.6	+0.4	-4.2	+11.8	+10.6	+10.4	+2.4	+1.1	+2.5
+ MedReason	+5.2	+0.4	+7.0	+14.1	+14.8	+14.8	+2.6	+2.3	+3.1
+ OUR	+4.9	+1.8	+5.2	+14.7	+13.7	+13.7	+5.7	+4.2	+6.9

Table 11: Performances of baseline models with deltas produced by exposure to different types of reasoning traces at inference time. Baselines are models prompted via 2-shot examples, each example being a question-options-answer triplet, while the deltas are obtained by prompting the same models via 2-shot examples, each example being a triplet question-options-trace triplet.

manovra di rotazione del MMII e la compressione del bacino non sono sufficienti per escludere una lesione spinale. - È necessario un esame neurologico più approfondito, iniziando con la valutazione del rachide cervicale.

****Processo di ragionamento:**** 1. Il paziente ha subito una caduta da 2 metri in piedi, un evento che può causare lesioni spinali, specialmente se l'atterraggio non è stato attutito. 2. I sintomi neurologici (formicolio, disestesia) indicano una possibile lesione spinale, in particolare a livello toracico o cervicale. 3. La stretta di mano e la pacca sulla spalla suggeriscono una funzionalità motoria conservata, ma non escludono una lesione spinale. 4. La manovra di rotazione del MMII e la compressione del bacino sono asintomatiche, suggerendo

una possibile assenza di lesione spinale a livello lombare o sacrale. 5. Per escludere una lesione spinale a livello cervicale, è necessario effettuare una valutazione neurologica iniziale, iniziando con la valutazione del rachide cervicale. 6. Il paziente deve essere posizionato in modo da permettere una valutazione neurologica completa, iniziando con il rachide cervicale. 7. L'opzione 1 propone di far sedere il paziente per esplorare il rachide cervicale, il che è coerente con la valutazione iniziale richiesta. 8. Le altre opzioni non sono appropriate per la valutazione iniziale di un paziente con sospetta lesione spinale.

****Conclusione:**** 1

Expert Comment. Presenta alcuni errori di

Model	MedExpQA			MedMCQA			MedQA		
	IT	EN	ES	IT	EN	ES	IT	EN	ES
Llama-3.1-8B-Instruct	48.9	62.8	47.9	46.4	57.5	47.9	50.4	63.1	49.7
+ m23k FT	+4.0	+4.3	+10.3	+6.8	+3.7	+3.1	+5.9	+10.0	+12.0
+ MedReason FT	+3.6	-0.7	+5.5	-1.1	-3.9	-0.6	+0.2	+0.5	+5.2
+ Huatuo FT	-5.2	+0.1	+0.0	-2.3	-0.3	-0.8	-2.0	+3.8	+1.3
+ OUR FT	+9.1	+2.6	+6.4	+0.4	+1.5	+2.7	+3.6	+8.5	+9.4
Qwen3-8B	60.1	66.0	64.9	53.0	58.6	53.8	57.0	67.1	60.1
+ m23k FT	-1.7	-2.1	-5.7	-3.0	-2.0	-2.7	-3.4	-1.9	-3.9
+ MedReason FT	+0.8	-1.2	-1.9	+0.1	+0.0	-0.7	+1.6	+0.5	-0.2
+ Huatuo FT	+4.2	+3.2	+2.1	+3.4	+0.3	+2.2	+8.2	+6.0	+4.2
+ OUR FT	+2.1	+3.7	+4.8	-0.7	+2.3	-0.6	+4.7	+4.6	+4.9

Table 12: Performances of baseline models with deltas produced by exposure to different types of reasoning traces at fine-tuning time. Baselines are models prompted via 2-shot examples, each example being a question-options-answer triplet, while the deltas are obtained by fine-tuning the same models on the dataset of traces.

		MedExpQA			MedMCQA			MedQA			AVG
family	size	IT	EN	ES	IT	EN	ES	IT	EN	ES	
HuatuoGPT	8B	56.4	67.0	56.3	51.1	64.1	54.4	57.4	75.6	58.3	60.1
MedReason	8B	55.1	64.3	58.4	49.5	61.1	50.3	54.8	70.9	62.0	58.5
m1	7B	48.6	62.2	48.1	47.7	59.3	48.6	50.8	66.9	51.5	53.7
Our SFT	8B	62.2	69.7	69.7	52.3	60.9	53.2	61.7	71.7	65.0	63.0
Our SFT+ 2-shot	8B	64.0	68.1	65.0	61.9	67.6	63.8	68.1	76.7	69.1	67.1

Table 13: Performance (accuracy) of state of the art 8B models on MedExpQA, MedMCQA, and MedQA in Italian, English and Spanish. Our proposed SFT model achieves the best performances in the 2-shot settings.

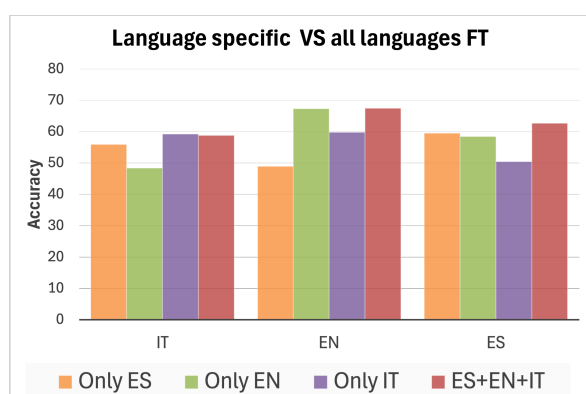


Figure 5: The histogram reports on the impact of fine-tuning on single-language ('Only ES/EN/IT') versus training on all the languages ('ES+EN+IT')

conoscenza medica: Concentra la valutazione sul rachide cervicale senza senza valutare l'esame neurologico complessivo

Contiene alcuni errori logici: Non considera la possibilità che la lesione spinale si aggravi facendo mettere seduto io paziente prima di aver controllato tutti i distretti

F.3 Spanish

PREGUNTA:

En un paciente con enfermedad pulmonar obstructiva crónica (EPOC) leve en fase estable se evidencia poliglobulia e insuficiencia respiratoria. En la radiografía de tórax no se aprecian alteraciones reseñables. ¿Cuál de los siguientes procedimientos consideraría realizar en primer lugar para descartar la coexistencia de otras enfermedades que puedan justificar los hallazgos descritos?:

OPCIONES:

'1': 'Poligrafía respiratoria.', '2': 'Ecocardiograma.', '3': 'Tomografía computarizada torácica.', '4': 'Gammagrafía pulmonar.', '5': 'Punción de médula ósea.'

RESPUESTA CORRECTA:

1

RESPUESTA DEL MODELO:

Búsqueda de posibles muestras de razonamiento: - La poliglobulia en un paciente con EPOC puede deberse a hipoxia

crónica, pero también puede ser secundaria a otras condiciones como mieloproliferativas o tumores. - La insuficiencia respiratoria en un paciente con EPOC leve sugiere que podría haber una causa subyacente no evaluada. - La radiografía de tórax no muestra alteraciones, lo cual no excluye patología pulmonar o sistémica. - La gammagrafía pulmonar puede ayudar a evaluar la perfusión y la ventilación pulmonar, lo cual es útil para descartar neumonía, embolia pulmonar o neoplasias. - La tomografía computarizada torácica es más sensible que la radiografía, pero no es el primer paso en un paciente con EPOC estable. - La ecocardiografía puede ser útil para evaluar hipertensión pulmonar, pero no es el primer procedimiento para descartar otras causas de poliglobulia. - La poligrafía respiratoria es útil para evaluar la función pulmonar, pero no aporta información sobre causas sistémicas de poliglobulia. - La punción de médula ósea es invasiva y no es el primer paso para descartar causas no hematológicas de poliglobulia.

Proceso de razonamiento: 1. El paciente tiene EPOC leve en fase estable, pero presenta poliglobulia e insuficiencia respiratoria, lo cual sugiere una causa subyacente no evaluada. 2. La poliglobulia puede deberse a hipoxia crónica, pero también puede ser secundaria a otras condiciones como mieloproliferativas, tumores o infecciones. 3. La gammagrafía pulmonar permite evaluar la perfusión y la ventilación pulmonar, lo cual es útil para descartar neoplasias, embolia pulmonar o neumonía. 4. La tomografía computarizada torácica es más sensible, pero no es el primer procedimiento a realizar en un paciente con EPOC estable. 5. La ecocardiografía puede ser útil para evaluar hipertensión pulmonar, pero no es el primer paso para descartar otras causas de poliglobulia. 6. La poligrafía respiratoria es útil para evaluar la función pulmonar, pero no aporta información sobre causas sistémicas de poliglobulia. 7. La punción de médula ósea es invasiva y no es el primer paso para descartar causas no hematológicas de poliglobulia. 8. Por lo tanto, la gammagrafía pulmonar es el procedimiento más adecuado para descartar otras enfermedades que puedan justificar los hallazgos descritos.

Conclusión: 4 **Expert Comment.** La respuesta no tiene en cuenta todos los elementos útiles: Información clínica del paciente, usualmente se repetiría el hemograma para ver si sigue teniendo policitemia.

QUESTION Gustavo comes to the emergency room with skin lesions and general malaise of several days of evolution. He has psoriasiform lesions on the trunk with involvement of palms and soles. He also presents asymmetric non-suppurative joint inflammation and bilateral ocular redness as well as erosions on the glans penis. In the subsequent anamnesis Gustavo recognizes a risky sexual contact 20 days before. What is his diagnosis?	OPTIONS 1) HIV infection 2) Secondary syphilis 3) Reiter's syndrome. 4) Erythema multiforme 5) None
MODEL ANSWER Search for potential reasoning traces: - Psoriasiform lesions on the trunk with involvement of palms and soles. - Asymmetric non-suppurative joint inflammation. - Bilateral ocular redness. - Erosions on the glans penis. - History of risky sexual contact 20 days before. - These features suggest a sexually transmitted infection with a reactive or disseminated phase. - Secondary syphilis is known to present with a rash, mucocutaneous lesions, and systemic symptoms. - Reiter's syndrome is a reactive arthritis that follows a genitourinary infection but does not typically present with psoriasiform lesions or mucocutaneous erosions. - Erythema multiforme is a hypersensitivity reaction and does not typically involve the glans penis or joints. - HIV infection may present with systemic symptoms but does not typically include the described mucocutaneous and ocular features. Reasoning process: 1. Gustavo presents with a combination of skin lesions, joint inflammation, ocular redness, and mucocutaneous erosions. 2. The psoriasiform rash on the trunk and palms/soles is a key feature of secondary syphilis. 3. The mucocutaneous erosion on the glans penis is also consistent with secondary syphilis. 4. The bilateral ocular redness and asymmetric joint inflammation are additional features of secondary syphilis. 5. The history of a risky sexual contact 20 days before is consistent with the incubation period of syphilis. 6. Reiter's syndrome is a reactive arthritis that follows a genitourinary infection but does not typically present with psoriasiform lesions or mucocutaneous erosions. 7. Erythema multiforme is a hypersensitivity reaction and does not typically involve the glans penis or joints. 8. HIV infection may present with systemic symptoms but does not typically include the described mucocutaneous and ocular features. Conclusion: 2	EXPERT ANALYSIS The model makes a critical medical knowledge error regarding Reiter's syndrome. It incorrectly states that Reiter's does not typically present with psoriasiform lesions or mucocutaneous erosions. In fact, these features are classic characteristics of the syndrome (known as keratoderma blennorrhagicum and circinate balanitis). This fundamental mistake led the model to incorrectly rule out the correct diagnosis."

Figure 6: Example of the expert analysis on a model answer which presents a critical clinical knowledge mistake.

QUESTION Indicate the clinical situation in relation to hepatitis B virus infection in a 5-year-old patient from Nigeria, with normal physical examination and the following serology for hepatitis B: HBsAg + / ANTI-HBs - / HBeAg - / ANTI-HBe + / ANTI-HBc IgM - / ANTI-HBc IgG + / DNA HBV +:	OPTIONS 1) Acute infection. 2) Chronic infection. 3) Vaccinated patient. 4) Asymptomatic carrier. 5) None
MODEL ANSWER Search for potential reasoning traces: - HBsAg positivity indicates current infection with the hepatitis B virus. - Anti-HBs negativity rules out immunity from vaccination or prior infection. - HBeAg negativity and anti-HBe positivity suggest a shift from active replication to a less infectious state. - Anti-HBc IgM negativity and IgG positivity indicate a past infection rather than an acute one. - HBV DNA positivity confirms active viral replication. - The combination of HBsAg +, anti-HBs -, HBeAg -, anti-HBe +, anti-HBc IgG +, and HBV DNA + is consistent with **chronic hepatitis B infection**. Reasoning process: 1. The patient is HBsAg positive, indicating current infection with the hepatitis B virus. 2. Anti-HBs is negative, ruling out immunity from vaccination or prior infection. 3. HBeAg is negative, and anti-HBe is positive, suggesting a shift from active viral replication to a less infectious state. 4. Anti-HBc IgM is negative, and IgG is positive, indicating a **past infection** rather than an acute one. 5. HBV DNA is positive, confirming active viral replication. 6. The combination of HBsAg +, anti-HBs -, HBeAg -, anti-HBe +, anti-HBc IgG +, and HBV DNA + is consistent with **chronic hepatitis B infection**. 7. The patient is asymptomatic, which is common in chronic hepatitis B, especially in the inactive carrier phase. 8. The presence of HBV DNA confirms that the virus is still replicating, even if the patient is asymptomatic. Conclusion: 2	EXPERT ANALYSIS The model correctly identifies the condition as a form of chronic hepatitis B and even mentions the "inactive carrier phase" in its reasoning. However, it fails to select the most specific and correct option, "Asymptomatic carrier" (4), and instead chooses the broader, less precise answer "Chronic infection" (2). This is a mistake in clinical conclusion.

Figure 7: Example of the expert analysis on a model answer which presents a mistake in linking the provided medical information with the most appropriate answer.

QUESTION A 42-year-old female patient reports pain with inflammatory features and swelling in both wrists, 2nd and 3rd metacarpophalangeal and proximal interphalangeal joints bilaterally and left ankle of 4 months of evolution accompanied by morning stiffness of more than one hour duration. Hand X-ray shows an erosion in the styloid process of the ulna in the right carpus. Laboratory tests showed Hb: 10 g/dL with ESR of 45 mm in the first hour, CRP 16 mg/L, rheumatoid factor 160 IU/ML. After 6 months of treatment with indomethacin and methotrexate, the patient persists with pain and swelling of both carpal, morning stiffness lasting 30 minutes and a CBC showing an ESR 30 mm in the first hour and a CRP 9 mg/dL. Regarding the attitude to take, which of the following is true:	OPTIONS 1) Suspend the prescribed treatment due to lack of response and initiate prednisone at high doses for symptom control only.', 2) Maintain the therapeutic attitude taken since we have only been on it for 6 months and it would be necessary to wait a minimum of 9 months to evaluate therapeutic response. 3) If there is no medical contraindication, consider adding an anti-TNF alpha to the treatment. 4) Start a second disease-modifying drug as soon as possible, since it would not be possible to start treatment with biologic therapy alone after methotrexate. 5) Consider starting treatment with anti-CD20 therapy associated with methotrexate.'
MODEL ANSWER Search for potential reasoning traces: - The patient has a 4-month history of symmetric joint involvement, morning stiffness, and inflammatory features. - Laboratory findings include anemia, elevated ESR and CRP, and a positive rheumatoid factor. - X-ray shows erosions in the carpus, consistent with rheumatoid arthritis (RA). - The patient has been on indomethacin and methotrexate for 6 months but still has persistent symptoms. - The treatment duration for methotrexate is typically longer than 6 months before a full response is expected. - Methotrexate is a first-line DMARD for RA and is often used as monotherapy or in combination. - If a patient does not respond adequately to methotrexate after 6 months, adding another DMARD or a biologic agent is appropriate. - Anti-TNF alpha agents are commonly used as second-line therapy in RA when methotrexate is insufficient. - Starting a biologic agent alone is not recommended without prior methotrexate. - Anti-CD20 therapy is not typically used in RA and is more relevant in lymphoproliferative disorders. Reasoning process: 1. The patient's clinical presentation, including symmetric joint involvement, morning stiffness, and inflammatory markers, is consistent with rheumatoid arthritis. 2. The X-ray findings of erosions in the carpus further support the diagnosis of RA. 3. The patient has been on methotrexate for 6 months, which is a standard duration for evaluating its effectiveness. 4. Despite treatment, the patient still has persistent symptoms, indicating that the current regimen is not sufficient. 5. Methotrexate is a first-line DMARD for RA and is often used as monotherapy or in combination with other DMARDs. 6. If a patient does not respond adequately to methotrexate after 6 months, adding another DMARD or a biologic agent is appropriate. 7. Anti-TNF alpha agents are commonly used as second-line therapy in RA when methotrexate is insufficient. 8. Starting a biologic agent alone is not recommended without prior methotrexate. 9. Anti-CD20 therapy is not typically used in RA and is more relevant in lymphoproliferative disorders. 10. Therefore, the most appropriate next step is to add an anti-TNF alpha agent to the current treatment regimen. Conclusion: 3	EXPERT ANALYSIS The model's clinical judgment is incorrect. It failed to recognize the patient's partial response to methotrexate (improvement in stiffness and inflammatory markers) and hastily concluded the treatment was insufficient. The correct approach, given the partial response, is to wait longer for the full effect of methotrexate, making option 2 the correct answer and the model's conclusion (option 3) a medical error.

Figure 8: Example of the expert analysis on a model answer which presents misuse of the patient information.