ENHANCING MEDITRON CAPABILITIES WITH SYNTHETIC AND REASONING DATASETS[*]

Master's Thesis

Xavier Theimer-Lienhard
Student
LiGHT
EPFL
`xavier.theimer-lienhard@epfl.ch`

Mary-Anne Hartley, Martin Jaggi
Supervision
LiGHT, MLO
EPFL
`mary-anne.hartley@epfl.ch`, `martin.jaggi@epfl.ch`

27th March 2025

---

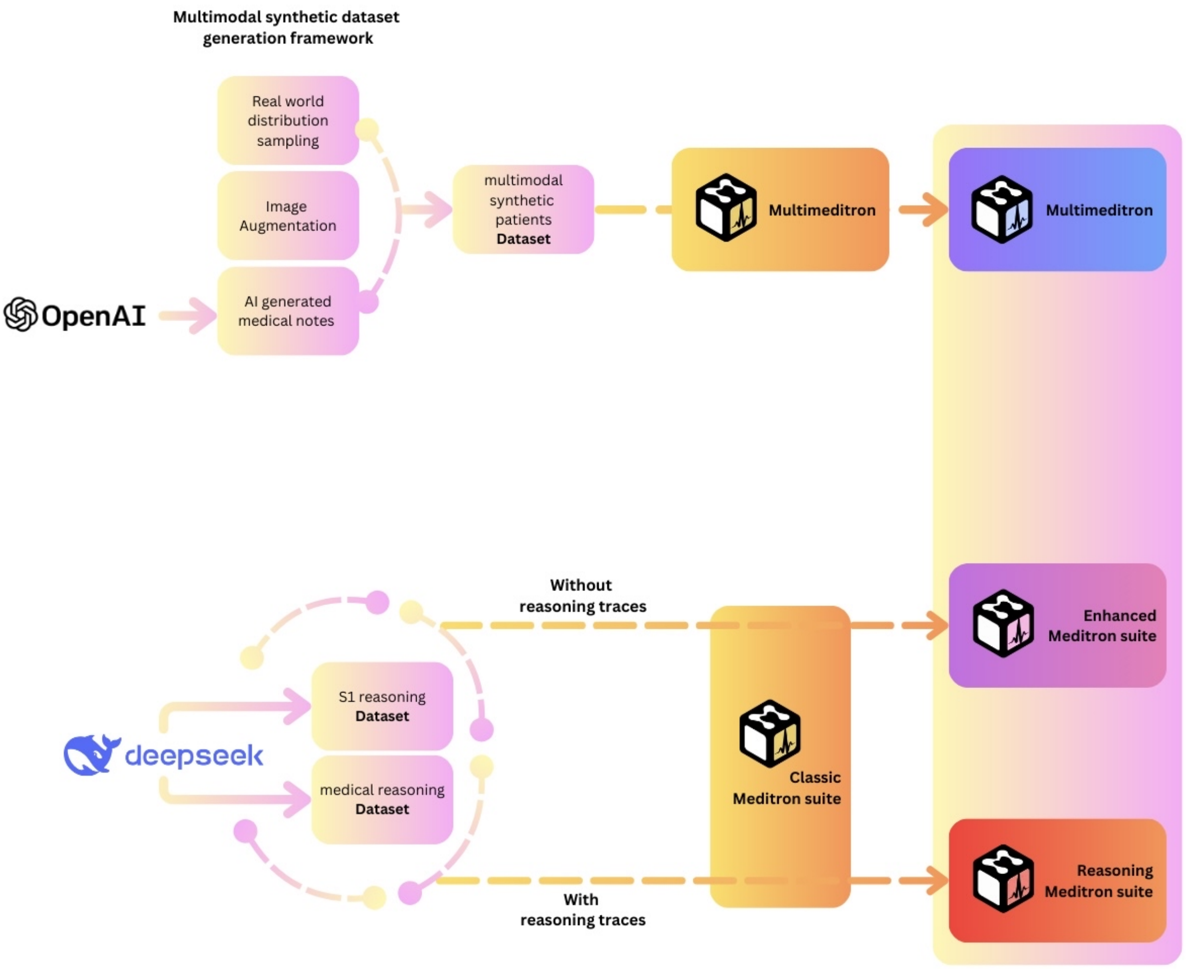
[*] *Citation*: **Xavier Theimer-Lienhard**; **Mary-Anne Hartley, Martin Jaggi** (Supervisor). *Enhancing Meditron capabilities with synthetic and reasoning datasets*. Mar 2025. Master's Thesis. Laboratory for Intelligent Global Health and Humanitarian Response Technologies (LiGHT), EPFL.

ABSTRACT

Medical language models face persistent challenges: privacy restrictions limit the availability of real patient data, and lack of clear interpretability makes it difficult to trust automated clinical decisions. Although existing models demonstrate strong factual capabilities, they often lack transparent reasoning steps that are vital for safe and explainable healthcare applications. To address this we developed a synthetic framework that produces privacy-preserving clinical records with textual patient files and relevant medical images. Next, we explored whether training on specialized reasoning datasets could improve the Meditron models' step-by-step analytical processes. We compared performance across different variants trained with or without explicit reasoning traces. Models finetuned with reasoning data showed improved performance on multiple benchmarks, and omitting explicit reasoning traces led to further gains, highlighting a complex interplay between reasoning visibility and accuracy. Our findings demonstrate that synthetic datasets can supplement scarce medical data, while structured reasoning enhances interpretability and accuracy. Future work will refine the balance between transparency and performance, extend the framework to more modalities, and evaluate real-world applications.

Keywords synthetic data · medical language models · reasoning · interpretability · multimodal

GRAPHICAL ABSTRACT



Enhancing Meditron suites with reasoning and synthetic datasets

1 Introduction

1.1 Problem overview and research gap

Medical datasets are notoriously difficult to obtain due to strict privacy regulations, yet robust medical language models depend on extensive, high-quality patient data. To address the scarcity of medical datasets, we propose a new framework that facilitates the generation of multimodal synthetic patient records. By leveraging this approach, we aim to produce highly realistic datasets while preserving patient confidentiality, thereby reducing the gap in reliable medical multimodal data.

Although language models exhibit impressive capabilities, interpretability remains a major challenge. In a medical context, it is not only the accuracy that matters but also the reasoning behind each decision. A recent advance in this domain is the emergence of reasoning models, which are trained to think before answering. These models produce a transparent reasoning trace, offering valuable insight into how and why the model arrives at its conclusions, making it significantly easier to interpret decisions in clinical settings. Despite these breakthroughs, medical professionals still need models that are both accurate and transparent, especially for critical patient-care decisions. To fulfill this need, we build upon the existing Meditron suite – an established family of open-source medical language models – by fine-tuning them specifically for enhanced medical reasoning. By incorporating the concept of thinking-before-answering, our enhanced Meditron models will provide detailed, interpretable reasoning paths that inspire greater trust in AI-assisted diagnostics and treatment recommendations, ultimately benefiting healthcare practitioners and their patients.

The two research directions, overcoming data scarcity and introducing better reasoning capabilities, will collectively address the broader challenge of developing trustworthy medical AI.

1.2 Literature review

Synthetic data Recent advances in deep learning have revolutionized synthetic data generation. Generative Adversarial Networks (GANs) have demonstrated remarkable capability in producing realistic medical images across various modalities, including X-rays, CT scans, and MRIs [1]. Similarly, Variational Autoencoders (VAEs) have shown promise in generating tabular and time-series medical data [2]. These approaches have significantly improved the quality and utility of synthetic medical data, but still struggle with maintaining consistent clinical relationships between patient characteristics, symptoms, and diagnoses. In the domain of clinical text generation, the SynTEG framework [3] generated synthetic clinical notes and patient records and attempted to address consistency challenges by incorporating medical knowledge graphs to constrain text generation, but still suffered from inconsistencies in complex multi-disease scenarios.

Synthetic generation has also been used in the Meditron project for creating instruction tuning datasets. These synthetic datasets were fully text-based, but they were augmented to be multiturn conversations and where specific to the medical domain, as to enhance the capabilities of Meditron. This effort also states that the most important factor for synthetic generation is diversity. For example they used resources to adequately model the world's population, they also used a diverse set of initial prompts.

Despite these advances, significant challenges remain in generating high-fidelity synthetic medical datasets that: (1) maintain realistic correlations between patient demographics and disease prevalence, (2) ensure clinical coherence across the entire patient record, (3) provide flexible control over data distributions to address specific research needs, and (4) incorporate multiple medical imaging modalities that align precisely with the synthetic clinical case. Our work aims to address these limitations through a modular framework that leverages recent advances in generative AI while incorporating medical domain knowledge to ensure clinical validity and relevance.

Medical Language Models and Reasoning Capabilities Large Language Models (LLMs) have emerged as powerful tools in the medical domain, with specialized models demonstrating significant potential for clinical applications. Recent developments in biomedical language models have shown increasing capabilities in medical knowledge representation, diagnosis assistance, and medical literature understanding, for example, MedPaLM [4], ChatGPT [5], or BioGPT [6]. The Meditron family of models [7] represents a significant advancement in open-source medical LLMs, exhibiting competitive performance across various medical benchmarks while maintaining transparency in development. These models have been fine-tuned on curated medical datasets using established foundation models like Llama and Qwen as bases, showing that domain adaptation can effectively transfer general reasoning capabilities to specific medical contexts.

Despite progress in factual medical knowledge, reasoning capabilities remain a critical challenge for medical LLMs. Complex clinical scenarios often require transparent step-by-step analysis that goes beyond simple pattern matching or fact retrieval [8][9]. It has been demonstrated that simply prompting models to "think step by step", also described as Chain-of-Thought can significantly improve reasoning performance without additional training [10][11]. To go further, fine-tuning approaches have been explored and consequently reasoning model like DeepSeek-R1 [12] have demonstrated state-of-the-art performance on reasoning benchmarks. The Simple test-time scaling approach [13] demonstrated that a small, high-quality reasoning distillation dataset from those reasoning models could significantly improve model performance on reasoning tasks without extensive computational resources.

Despite these advances in reasoning capabilities, there is still a lot of research to be done on reasoning and medical capabilities. Our works aims to understand what type of reasoning finetuning will advance best the medical capabilities of LLMs through finetuning and evaluation of the existing Meditron suite of models.

1.3 Research objectives

1. Explore multimodal synthetic patient data
    - (a) Create a framework for multimodal synthetic patient data generation
    - (b) Get clinician evaluation and feedback on the generated synthetic data
2. Explore reasoning finetuning of Meditron models
    - (a) Assess if reasoning capabilities can be obtained on Meditron models with simple finetuning
    - (b) Compare the impact of medical vs. non-medical reasoning finetuning
    - (c) Determine whether improved reasoning capabilities are detrimental to medical performances
    - (d) Determine whether explicit reasoning traces in training data improve or hinder model performance

1.4 Contributions

This research makes several key contributions to the field of medical AI. (1) We introduce a novel framework for generating high-quality multimodal synthetic medical datasets that preserve patient privacy while ensuring clinical realism, addressing a critical bottleneck in medical AI development. (2) We explore reasoning enhancement strategies for medical language models. (3) We develop and release the Reasoning Meditron suite—enhanced versions of the original Meditron models with improved reasoning capabilities while preserving medical expertise. Through these contributions, we advance the model capabilities necessary for building more transparent, interpretable, and trustworthy AI systems for healthcare applications.

2 Framework for multimodal synthetic medical dataset generation

2.1 Objectives

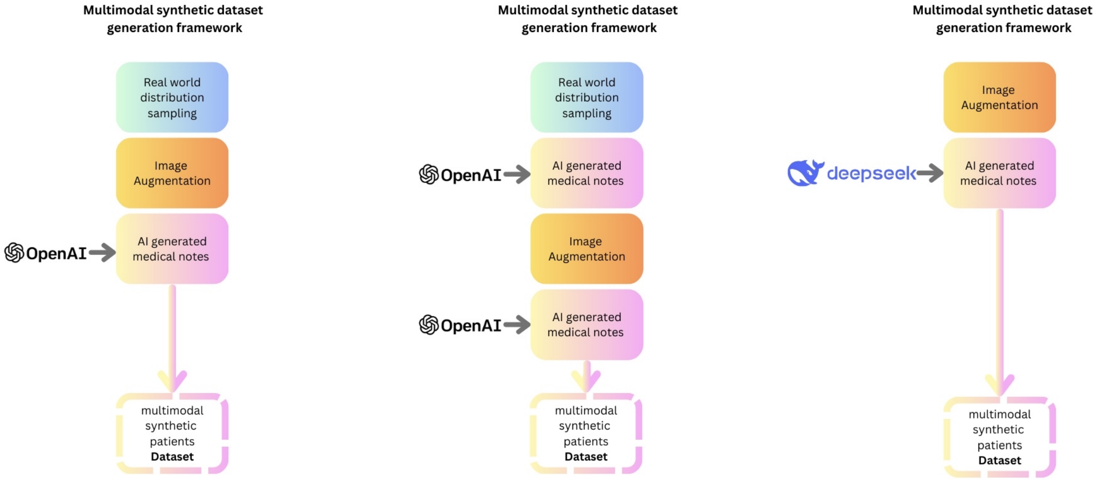

Multimodal synthetic dataset generation framework
Real world distribution sampling
Image Augmentation
OpenAI
AI generated medical notes
multimodal synthetic patients Dataset
Multimodal synthetic dataset generation framework
Real world distribution sampling
OpenAI
AI generated medical notes
Image Augmentation
OpenAI
AI generated medical notes
multimodal synthetic patients Dataset
Multimodal synthetic dataset generation framework
Image Augmentation
deepseek
AI generated medical notes
multimodal synthetic patients Dataset

Figure 2.1: This graph shows 3 different configurations using our synthetic framework. We build our pipeline with modularity and easy tailoring in mind.

Motivation Multimodal patient data provides comprehensive overview of a patient. Modalities range from electrocardiograms (ECG), magnetic resonance imaging (MRI), X-rays to blood work and so many others. When combined with text information, they facilitate representation of health and disease. By creating a synthetic framework for the generation of this type of data, we aim to leverage all the qualities of multimodal patient data. We also want to overcome the limitations of existing multimodal patient data. Multimodal patient data is rare because first of strong patient privacy regulations and second because multimodality is a rare component in the existing medical datasets. Existing datasets do not offer the flexibility of a synthetic framework. The framework could be used to simulate scenarios that are rare or underrepresented in existing datasets

Objectives Having stated why we want to create a framework for multimodal patient data generation, we still need to explain what are our main focuses when we built the framework. Our two objectives are : (1) We want to create a framework that is modular, so that the process itself of synthetic generation can me modified at will; flexible, so that it is able to represent any type of modality or situation; and that generates representative and coherent synthetic patient data, so that it can be as useful as existing patient datasets. (2) We want to iteratively design and refine the framework based on the feedback from clinicians. Engaging with domain experts is central to our approach, as their insights and feedback help ensure that the synthetic data aligns with real world clinical realities.

2.2 Framework design

We design the framework in a block architecture, as it is the most adapted to our flexibility and modularity focus. We create three blocks and leave space for more in our design. The three blocks are : (1) a sampling block as to create representative dataset that align with real distributions, (2) a generation block that leverages the emergent capabilities of LLMs to write patient information and fill the blanks, (3) a modality retrieval block to make the data multimodal.

Sampling block The sampling block allows to sample from any given distribution in input. Examples of distributions that could be used are disease prevalence distribution, age distribution per region, sex distribution, among others. We also allow for multiple type of statistical distribution, and we use weighted sampling.

Generation block The generation block is a simple local LLM or API use which allow for the user to select any prompt and system prompt. We provide the prompts that have been used in the repository. The block can also take as input any context that has been generated by any previous block, including another generation block. We also used DSPY in the hopes of finetuning the prompts but without results [14]. This block aims to generate patient histories or patient records of their medical process.

Modality retrieval block The modality retrieval block can retrieve from any modality dataset that the user gives the framework. The retrieval process uses the description of the modalities to select the best fitting modality to the rest of the context produced by the framework. It does so by embedding all modalities descriptions and the context and then uses a similarity metric to select the best one. This allows to select the most coherent modalities.

2.3 Implementation details

Our framework is dependent on important libraries and APIs. For the generation block, we first used the DSPY framework but finally reverted to a simpler project that allows use for most models with API keys. For the modality block, we use the OpenAIEmbeddings to create the embeddings of the documents, and we use the Langchain library to store them and perform retrieval.

2.4 Evaluation

We generate 20 examples for each configuration using our framework. And then present the results for evaluation and feedback to clinicians. Clinicians then give their feedback in a discussion with us and help us build a better framework.

Configuration used for clinician evaluation We use two different configurations for clinician evaluation : In the first one we have (1) A sampling block that samples randomly on the sex, (2) A sampling block that samples the age of the patient, (3) A generation block that writes the reason for which the patient presents themselves to the hospital, that is the patient history and the patient problematic, and (4) a modality retrieval block. In the second configuration we have (1) A modality retrieval block that samples randomly since there is no context, (2) A generation block that writes the patient history based on the the modality, and (3) A generation block that writes the patient problematic. Note that both configurations use as a few shot a clinician based example extracted from MIMIC [15].

2.5 Results

(1) We release a framework for multimodal synthetic patient data generation We create a framework that in design is modular, flexible and generates representative and coherent synthetic data. The code for the synthetic framework is available on the Synthetic dataset repository, under the Open Meditron group: GitHub repository for synthetic data. We provide an example of a generated patient record in Appendix A.5.

(2) Clinician evaluation We gathered feedback from clinicians in an iterative manner along the steps fo creating the framework. As the building of the framework was iterative, the first feedback on preliminary framework was mostly negative and highlighted the need for better prompting in the generation block and better formatting of the output. In the next iterations, they also highlighted that patient notes have a specific format that is very straight to the point and to format the synthetic data in this specific way would augment the representativeness and coherence of the data. Another feedback was that there was an over representation fo rare disease, which was inherent with the way we sampled the diseases. At the end of the iterative process, the prompts still needed some reformatting but the clinicians noted progress, notably in the patient history formatting. They also noted that the generated patient data was more coherent. Finally it has been remarked by the clinicians that the modalities tended to clash with the patient information mentioned in the rest of the context, and this specific feedback we have not been able to address except by making the framework in such a configuration that the first block is the modality retrieval.

2.6 Discussion

(1) We release a framework for multimodal synthetic patient data generation Our framework represents a step toward addressing the scarcity of multimodal medical datasets while respecting patient privacy concerns. While not without limitations, it demonstrates the potential for synthetic data to supplement real-world datasets in medical AI development and evaluation. As we continue to refine this approach, the balance between representativity, coherence, and efficiency will remain central to our efforts, guided by ongoing collaboration with clinical experts who can validate the utility of the generated data for real-world applications.

(2) We uncover challenges with the help of clinician feedback But we also see improvement in coherence and representativity. Through multiple iterations and clinician feedback, we observed significant improvements in both the coherence and representativity of our synthetic data. The framework's modular design allowed for rapid adjustments to the generation process, particularly in reformatting patient histories to align with clinical documentation standards. This iterative refinement process demonstrates the importance of domain expert involvement in synthetic data generation. The clinician feedback highlighted an important limitation: modalities often clashed with the patient information mentioned in the rest of the context. Our attempt to address this by reconfiguring the framework to prioritize modality retrieval as the first block represents only a partial solution. Future work could explore other approaches for modality-text alignment.

2.7 Future work

Integration with Medical Processes A comprehensive synthetic data framework should ideally model the entire patient journey through the healthcare system. Our current implementation focuses primarily on specific clinical snapshots rather than the complete medical process. To address this limitation, future iterations could incorporate temporal dynamics, treatment pathways, and follow-up care patterns. This would require creating dependencies between sequential synthetic data points and modeling the decision-making processes that guide clinical care, potentially through reinforcement learning approaches or process mining techniques applied to existing medical workflows.

Emergent Capabilities of Multimodal Models An important consideration is whether specialized frameworks like ours will remain necessary as multimodal foundation models continue to advance. Recent developments suggest that large multimodal models are developing emergent capabilities that might eventually eliminate the need for complex synthetic data generation pipelines. These models can increasingly generate coherent cross-modal content without explicit alignment mechanisms. However, the medical domain's specialized nature, with its strict requirements for accuracy and its unique vocabularies and relationships, may continue to necessitate domain-specific approaches even as general-purpose models improve.

Potential Applications Despite these challenges, our framework offers several valuable applications. It could be used to augment existing datasets for rare conditions, simulate clinical scenarios for medical education, test clinical decision support systems with synthetic patient cohorts, or evaluate the robustness of diagnostic algorithms across diverse patient populations. The framework could also serve as a testbed for exploring ethical issues in medical AI by explicitly modeling various bias scenarios and evaluating their downstream effects.

Efficiency Considerations The current implementation raises questions about computational efficiency, particularly when scaling to large synthetic cohorts. The reliance on LLM inference and embedding-based retrieval introduces computational bottlenecks that may limit practical applications. More efficient approaches might include distilled models specifically trained for medical data generation, cached retrieval patterns, or hierarchical generation strategies that reuse computation across similar synthetic cases.

3 Exploring reasoning finetuning of Meditron models

3.1 Study design

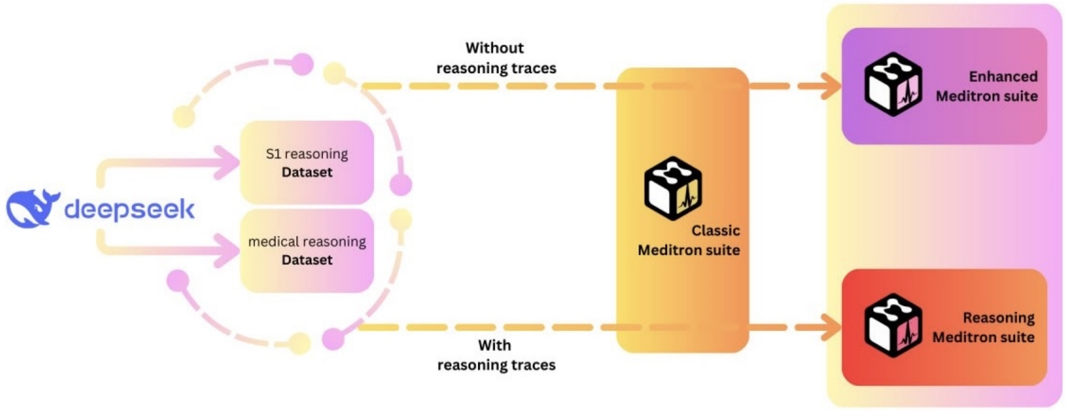


Figure 3.1: Exploring reasoning finetuning of Meditron models. This graph shows the reasoning training and its 4 different experiments : general, medical, with and without reasoning traces.

Large language models have demonstrated considerable strengths in medical knowledge, yet they often encounter difficulties when confronted with complex reasoning tasks that require transparent, step-by-step analysis. This limitation poses a significant challenge in clinical decision support systems, where trust and accountability depend on interpretable rationales. To address this gap, we initially hypothesized that finetuning with reasoning datasets could enhance reasoning capabilities without compromising medical expertise. To explore this hypothesis, we formulated three primary objectives: **(1)** assess if reasoning capabilities can be obtained on Meditron models with simple finetuning, **(2)** assess the relative impact of medical reasoning versus general reasoning datasets, and **(3)** examine if reasoning finetuning is detrimental to medical performances and knowledge. However, in the first experimental run, we inadvertently omitted the explicit reasoning traces from the training dataset. Contrary to our expectations, the absence of these traces led to improved outcomes on several benchmarks. This motivated us to introduce a fourth aim: **(4)** evaluate the efficacy of finetuning with explicit reasoning traces as compared to models trained without such traces. Through these four objectives, our study seeks to determine not only whether enhanced reasoning can be achieved with simple supervised finetuning using distilled traces of reasoning models but also how best to integrate reasoning strategies in models specialized for medicine.

3.2 Experimental setup

We conducted a systematic ablation study, generating six finetuned variants for each base model (so 12 finetuned models total). These variants were created by manipulating two key factors: (1) the inclusion or exclusion of explicit reasoning traces during training, and (2) the domain focus of the reasoning datasets (medical-specific versus general-purpose versus combining them). This design allowed us to compare effectively datasets and methods.

Base models We select two models from the Meditron suite of medical LLMs, Meditron-8b and Medicouenne-7b, because we believe that evaluating more models would give diminishing returns for our objectives. We chose them based on their results, their success in the open-source community and their reduced model size. **(1)** Medicouenne-7b: The Medicouenne-7b model is the Qwen-2.5-7B-Instruct foundation model finetuned on the Meditron dataset, offering enhanced performance for medical language tasks. **(2)** Meditron-8b: The Meditron-8b model is the Llama-3.1-8B-Instruct foundation model also finetuned on the Meditron dataset, offering enhanced performance for medical language tasks.

3.3 Datasets

We compare the impact of medical and non medical reasoning datasets. All datasets are distilled from Deepseek-R1, a large reasoning model with non-encrypted reasoning traces (which is not the case for the reasoning models of OpenAI). Datasets are remarkable for their very small size which is a choice we make for fast evaluation and based on the findings of the Simple test-time scaling paper[13]. We download the first dataset on hugging-face and upload the second one for reproducibility.

S1.1 Simple test-time scaling Dataset The S1.1 Simple test-time scaling Dataset [13] is a high quality distillation reasoning dataset. We chose that dataset because it is a state of the art dataset that has been used to create reasoning models using only 1000 distillation samples [13]. For this reason it is a very small dataset, of only 1,000 samples, which have been filtered from a larger dataset of over 60,000 samples. The dataset samples have been distilled from a large reasoning model (Deepseek-R1). The dataset includes mathematical problem-solving, logical deduction, commonsense reasoning, and analytical thinking. The paper which introduces the dataset explains in detail how the filtering process was made, to ensure accuracy, coherence, and consistency in the reasoning tasks presented, and shows that it gets better performances than the original 60,000 samples dataset.

Medical-QA-Reasoning Dataset The Medical-QA-Reasoning Dataset consists of 4,000 medical reasoning samples. We create this dataset because of there is no medical QA reasoning dataset openly available. We created the dataset by sampling randomly questions from the following medical benchmarks: MedQA [16], MedMCQA [17] and PubMedQA [18], and then generating answers using the reasoning model DeepSeek-R1 [19]. This created an unfiltered medical distillation reasoning dataset that we then filtered to keep only the correct answers using the truth labels from the medical benchmarks.

Combined Dataset We choose to also combine the S1.1 dataset and the Medical-QA-Reasoning Dataset. This will allow us to study the combined impact of both dataset. The combination process is a simple shuffle aggregation.

Datasets without reasoning traces For all three datasets we create a version without reasoning traces. We drop the reasoning traces in order to achieve our third objective, (3) evaluate the efficacy of finetuning with explicit reasoning traces as compared to models trained without such traces. For reference, reasoning models generate their answer in two phases, the reasoning phase and the answer phase, that they indicate with specialized tokens. We drop the reasoning model reasoning phase part.

Preprocessing To ensure consistency, we format each sample using a standardized chat template and special tokens to indicate start and end of the "thinking", as follows:

```
user: {{question}}
assistant: <start>thinking {{deepseek_reasoning}} <start>answer {{deepseek_attempt}}
```

3.4 Training

As stated earlier, we created 12 finetuned models. For each of these finetunings we used the same training configuration, which was based on the one outlined in the Simple test-time scaling paper [13].

Hyperparameters The finetuning process was conducted with a learning rate of $1e^{-5}$, ensuring a balance between convergence speed and stability. A micro batch size of 1 was used, with gradient accumulation steps set to 1, effectively increasing the batch size for stable updates. The training was performed using the AdamW optimizer, with β_1 set to 0.9, β_2 to 0.95, and a warm-up ratio of 0.05 to gradually introduce learning and prevent instability in the early stages of training. To further regularize the model, a weight decay of $1e^{-4}$ was applied. The model was trained with mixed precision (bfloat16), allowing for optimized memory efficiency and faster computations while maintaining numerical stability.

Computation infrastructure The models were trained on a cluster node equipped with 8 NVIDIA H100 GPUs from the RCP cluster at EPFL, leveraging a high-performance computing setup. The Meditron protocol was used as the training framework—this is a custom framework built on Axolotl, designed to streamline the finetuning process. The training pipeline was further accelerated using DeepSpeed, which enabled efficient distributed training. The model was trained on a cluster node equipped with 8 NVIDIA H100 GPUs.

Carbon impact Each training has a duration of approximately 15 minutes. The very short duration of those trainings is the consequence of the small size of the datasets, the small size of models and the parallel distribution on 8 GPUs. When calculating the CO2 impact of the study trainings, we amount to 0.672kgCO2. Knowing that a flight of 3 hour 35 minute is equivalent to 350kgCO2, we calculate that we are consuming 1000 times less CO2 than a 3h45 flight. This is also thanks to the high volume of hydroelectric power based electricity in Switzerland, which makes its electrical grid very CO2 efficient.

3.5 Evaluation

We evaluate our models on 3 types of evaluations. (1) Logits evaluation on medical benchmarks. This is the most common evaluation type but is also unfit for reasoning evaluation. (2) Generation based medical benchmarks. The same benchmarks but we let the model generate their answer to produce their reasoning. (3) Non-medical reasoning benchmark. To evaluate the reasoning performances of our models.

Logits evaluation on medical benchmarks To assess the clinical performance of our finetuned models, we evaluated them on four commonly used medical question-answering benchmarks: MedQA[16], PubMedQA[18], MedMCQA[17], and MMLU-Medical (a medical-only subset of MMLU) [20]. For each multiple-choice question, the model outputs a probability distribution over all possible answers. We select the answer with the highest logit score as the model's prediction. This approach directly measures how well the model's internal representations differentiate between correct and incorrect responses, but this approach is also not fit to evaluate reasoning capability as a consequence.

Generation based evaluation on medical benchmarks Instead of directly comparing logits, we prompt the model in a generative manner and allow it to produce a free-text answer. Although more computationally intensive, this approach captures real-world usage scenarios where the model must reason openly before concluding. It also tests whether explicit reasoning steps improvements translate into better final predictions. To identify the chosen option by the model, we parse the response using Regex. We evaluate models on the same four medical benchmarks : MedQA[16], PubMedQA[18], MedMCQA[17], and MMLU-Medical (a medical-only subset of MMLU) [20].

Reasoning benchmark On top of evaluating the finetuned models on medical benchmarks, we also evaluate the models on a reasoning benchmark: MMLU STEM (the subset of MMLU that is Science, Technology, Engineering and Mathematics (STEM) related) [20]. MMLU is a multitask benchmark that necessitates extensive world knowledge and problem solving abilities. We chose MMLU because it is commonly used to evaluate reasoning models [19][13]. And we chose the STEM subset because it is a subset that necessitates more reasoning that the others, as can be shown by the low performances of Meditron-8b (41.49% accuracy on MMLU STEM) and Medicouenne-7b (57.91% on MMLU STEM).

3.6 Results

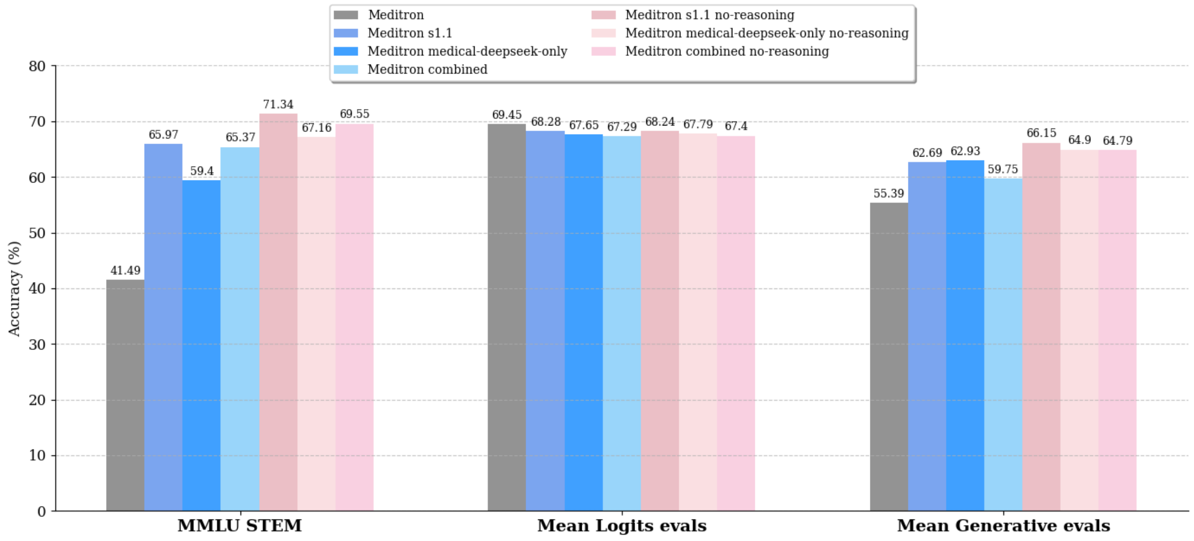


Figure 3.2: Benchmark performance of Meditron variants. The same graph for Medicouenne variants is in the Appendix A.1. medical-deepseek-only stands for the Medical-QA-reasoning dataset.

(1) The newly trained models are reasoning models The models results on MMLU STEM show a significant improvement over the base model, see table 3.1. The improvement for Meditron ranges from 20 to 30 points, while the improvement for Medicouenne ranges from 14 to 21 points. The models trained with no reasoning traces perform better or on par with the models trained with the traces on the reasoning benchmark. The average of results for

Meditron trainings with traces is 63.58, while the average without traces is 69.35, and the average of results for Medicouenne training is 75.03 with traces and 74.43 without. When comparing the datasets impact on MMLU STEM, we can see that all finetuned models outperform the base model by more than 15 points. We also present extracts of the generative evaluations in the Appendix A.2, A.3 and A.4, who show that the trained models indeed have a reasoning phase before their answer. Those results show that the reasoning finetunings have been successful, as the models now perform much better on the reasoning benchmark.

Model	**MMLU STEM**	**Mean Log**	**Mean Gen**
⬜ Medicouenne	57.91	67.02	61.49
🟦 Medicouenne s1.1	75.52	67.52	61.53
🟦 Medicouenne medical-deepseek-only	72.84	67.38	60.37
🟦 Medicouenne combined	76.72	66.63	62.83
🟥 Medicouenne s1.1 no-reasoning	**78.51**	68.15	61.52
🟥 Medicouenne medical-deepseek-only no-reasoning	73.73	**68.5**	64.08
🟥 Medicouenne combined no-reasoning	71.04	68.04	**65.92**
🟦 Medicouenne with reasoning Avg	75.03	67.18	61.57
🟥 Medicouenne without reasoning Avg	74.42	68.23	63.84
⬜ Meditron	41.49	**69.45**	55.39
🟦 Meditron s1.1	65.97	68.28	62.69
🟦 Meditron medical-deepseek-only	59.4	67.65	62.93
🟦 Meditron combined	65.37	67.29	59.75
🟥 Meditron s1.1 no-reasoning	**71.34**	68.24	**66.15**
🟥 Meditron medical-deepseek-only no-reasoning	67.16	67.79	64.90
🟥 Meditron combined no-reasoning	69.55	67.40	64.79
🟦 Meditron with reasoning Avg	63.58	67.74	61.79
🟥 Meditron without reasoning Avg	69.35	67.81	65.28

Table 3.1: Performance comparison of Meditron and Medicouenne finetunings across evaluation benchmarks. Log stands for Logits based evaluation on medical benchmarks. Gen stands for Generative based evaluation on medical benchmarks. medical-deepseek-only stands for the Medical-QA-reasoning dataset. All values are accuracies (%).

Dataset	**Meditron**	**Medicouenne**
S1.1 dataset	**64.42**	61.53
Medical reasoning dataset	63.92	62.23
Combined dataset	62.27	**64.38**

Table 3.2: Aggregated results of Meditron and Medicouenne finetunings across Generation based medical benchmarks. We aggregate independently of whether we finetune with or without reasoning traces. All values are accuracies (%).

(2) Combined dataset and non-medical reasoning dataset perform best On generation based medical benchmarks, we aggregate results independently of whether we train with the traces or not to evaluate the impact of the datasets, see table 3.2. We observe that for Medicouenne models, the best performing dataset is the combined dataset, followed by the medical reasoning dataset and then the S1.1 dataset. Inversely for Meditron, the best performing dataset is the S1.1 dataset followed by the medical reasoning dataset and then the combined dataset. When comparing the datasets impact on MMLU STEM, we can see that models trained with the S1.1 datasets perform better than those trained with the medical-QA-reasoning dataset or the combination of both. Then the model trained on the combined dataset perform close to the best models and finally, the models trained on the medical-QA-reasoning dataset have the least good results. This is expected as the S1.1 dataset is a higher quality reasoning dataset, and we expect the models trained on it to have better results on a reasoning benchmark.

(3) Reasoning finetunings improve medical performances All trainings improve accuracy significantly on generation based benchmarks. On average, the gain on accuracies ranges from 5 to 11 points for Meditron and from -1 to 5 points for Medicouenne. The only task that the newly finetuned models struggle on is PubmedQA and this only for Medicouenne. If we remove PubmedQA results from the average then the gain for Medicouenne range from 3 to 7 points. Those results are better than expected as the improvement on the medical benchmarks is really significant. We observe that results on logits based medical benchmarks are as expected similar or worse. See Appendix A.1 for detailed results table of logits evaluation. Trainings with reasoning traces do not improve results significantly for Medicouenne and reduce accuracy slightly for Meditron, while trainings with reasoning traces have a more noticeable impact but still smaller than 2 accuracy points on average for Medicouenne and also reduce slightly accuracy for Meditron. This is expected as reasoning finetuning should not improve results on logits evaluation as it is an evaluation that does not depend on generation and reasoning. Moreover we hypothesize that Meditron results being consistently reduced compared to the base model that Llama is a model better trained on medical knowledge compared to Qwen from the start. When comparing datasets results, we can see that the S1.1

dataset has the best impact on the logits evaluations, with or without reasoning traces. We estimate that this could be explained by the higher quality and novelty of the S1.1 dataset.

(4) Training without reasoning traces leads to better performances When evaluating on generation based medical benchmarks, we observe that the trainings without reasoning traces outperform on average the trainings with reasoning traces. If we average the mean results we obtain 61.79 for Meditron models trained with traces and 65.28 for Meditron models trained without traces. Similarly, we obtain 61.58 for Medicouenne models trained with traces and 63.84 for Medicouenne models trained without traces. We nuance this finding by also noting that Medicouenne trained with reasoning traces on the combined dataset is better than Medicouenne trained without reasoning traces on the S1.1 dataset. The best training is the training without reasoning traces on the combined dataset for Medicouenne and the training without reasoning traces on the S1.1 dataset for Meditron.

Model	MedMCQA Gen	MedQA Gen	PubMedQA Gen	MMLU Med Gen
■ Medicouenne	52.52	56.56	**57.2**	79.67
🟦 Medicouenne s1.1	53.86	65.75	45.2	81.3
🟦 Medicouenne medical-deepseek-only	56.51	65.75	44.4	74.8
🟦 Medicouenne combined	55.37	66.85	48.6	80.49
🟥 Medicouenne s1.1 no-reasoning	56.3	62.29	46.2	81.3
🟥 Medicouenne medical-deepseek-only no-reasoning	**56.73**	62.06	53.8	**83.74**
🟥 Medicouenne combined no-reasoning	56.22	**69.96**	56.2	81.3
■ Meditron Ins	53.55	59.47	37.00	71.54
🟦 Meditron Ins s1.1	57.97	67.56	49.60	75.61
🟦 Meditron Ins medical-deepseek-only	55.99	**67.87**	49.00	78.86
🟦 Meditron Ins combined	53.98	65.51	49.60	69.92
🟥 Meditron Ins s1.1 no-reasoning	60.32	66.93	54.40	**82.93**
🟥 Meditron Ins medical-deepseek-only no-reasoning	**60.89**	64.10	**59.00**	75.61
🟥 Meditron Ins combined no-reasoning	59.86	64.26	57.00	78.05

Table 3.3: Performance comparison of Meditron and Medicouenne finetunings across Generative based medical benchmarks. Gen stands for Generative based evaluation on medical benchmarks. medical-deepseek-only stands for the Medical-QA-reasoning dataset. All values are accuracies (%).

3.7 Discussion

(1) Reasoning capabilities can be learned effectively and economically with simple finetuning Our results are in complete accordance with the results of the Simple test-time scaling paper [13]. On top of these findings, we also find that removing the traces lead to improved results. But as a consequence of these improved results, the finetuned models also do not behave as reasoning models since they are not trained as such. Exploiting both results, we find a way to improve our standard models and a way to transform our models into reasoning models.

(2) Combined dataset and non-medical reasoning datasets perform best We have shown that training with the combined dataset leads to the best performances on medical benchmarks for Medicouenne. For Meditron it is more nuanced as training with only the non-medical dataset leads to slightly better performances. If we compare the impact of the medical reasoning dataset, for a dataset of a much lower quality, as it as not been filtered other than for correctness and is a very small sample, it is clear that its impact is almost as good as the S1.1 dataset. This also hints in favor of combined medical datasets utility and efficiency.

(3) Reasoning finetunings improve medical performances We observe that all training lead to improved performances on medical benchmarks, except for PubmedQA on Medicouenne trainings. This shows big promise in making open source models even more accurate. We believe that reasoning models will give access to better datasets thanks to distillation and to better models thanks to reasoning capabilities.

(4) Training without reasoning traces leads to better performances On most if not all evaluations, training without the reasoning traces give better or similar results compared to training with the reasoning traces. This is unexpected and is interesting because without the reasoning traces, the datasets are only very good answers to questions that have been thought of but where the reasoning to the final answer is not present. This can maybe be explained by the incorrectness and noisiness of the reasoning traces. Reasoning traces are commonly very long sequences of text where the model is gonna check on its own answer multiple times and repeat itself, and as such is maybe not good to use for distillation.

3.8 Future work

The medical-QA reasoning dataset is promising But needs to be enhanced, in the same manner as the S1.1 dataset has been. Our next effort will be into creating a better Medical-QA-reasoning dataset and evaluate its impact. To create a better version, we plan to select questions in the training subsets of Medical benchmarks that the Meditron model struggle on rather than simply sampling at random in the training subsets. We think the outcomes look very promising since the results of the current Medical-QA-reasoning dataset are already great.

Our study should be expanded to bigger models We have limited our study to 2 small models : Medicouenne-7b and Meditron-8b, we are curious of whether the impact is the same for bigger models in 14b, 32b and 70b, the highest performing Meditron model.

We would like to evaluate the fluency of the new reasoning models When looking at the newly finetuned variants, we think the results fluency is better than expected, for models of this size, see Appendix A.2, A.3 and A.4. We would like to conduct an evaluation to have human annotators evaluate the fluency and logical reasoning of the answers of the finetuned models compared to the base model.

4 Conclusion

Our research addressed dual challenges in medical AI: data scarcity and reasoning transparency. We demonstrated that synthetic data generation can produce privacy-preserving clinical records while maintaining realism, offering a solution to data limitations. Our reasoning experiments revealed that fine-tuning Meditron models with small, high-quality reasoning datasets significantly enhanced performance on both medical and general reasoning tasks. Interestingly, models trained without explicit reasoning traces often outperformed those with traces, highlighting a complex interplay between transparency and accuracy.

The enhanced Meditron suite developed through this work offers improved medical reasoning while maintaining interpretability. Future work should refine the balance between transparency and performance, extend the framework to additional modalities, and evaluate these systems in real-world clinical settings. These advancements contribute to the development of AI systems that are both medically accurate and interpretable, better supporting healthcare professionals in clinical decision-making.

Acknowledgments

I want to thank again Professor Annie Hartley for inviting us to Yale to complete our thesis. I also want to thank Alexandre Sallinen for his always insightful supervision. I thank Iris Segard and Maud Dupont-Roc for their collaboration on Synthetic data. Finally I wish to thank Etienne Boisson and Bastien Bernath for being such incredibly awesome coworkers and friends.

Data and code availability

The code to reproduce the reasoning methods is available on the Meditron protocol GitHub repository: GitHub repository for Meditron protocol. The code for the synthetic framework is available on the Synthetic dataset repository: GitHub repository for synthetic data. Both repository are under the Open Meditron group.

References

[1] Xin Yi, Ekta Walia and Paul Babyn, 'Generative adversarial network in medical imaging: A review,' *Medical image analysis*, vol. 58, p. 101 552, 2019.

[2] Mehran Pesteie, Purang Abolmaesumi and Robert N Rohling, 'Adaptive augmentation of medical data using independently conditional variational auto-encoders,' *IEEE transactions on medical imaging*, vol. 38, no. 12, pp. 2807–2820, 2019.

[3] Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun and Bradley A Malin, 'Synteg: A framework for temporal structured electronic health data simulation,' *Journal of the American Medical Informatics Association*, vol. 28, no. 3, pp. 596–604, 2021.

[4] Karan Singhal *et al.*, *Towards expert-level medical question answering with large language models*, 2023. arXiv: `2305.09617 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2305.09617`.

[5] A. J. Thirunavukarasu *et al.*, 'Trialling a large language model (chatgpt) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care,' *JMIR Medical Education*, vol. 9, e46599, 2023. DOI: `10.2196/46599`. [Online]. Available: `https://doi.org/10.2196/46599`.

[6] Renqian Luo *et al.*, 'Biogpt: Generative pre-trained transformer for biomedical text generation and mining,' *Briefings in bioinformatics*, vol. 23, no. 6, bbac409, 2022.

[7] Zeming Chen *et al.*, 'Meditron-70b: Scaling medical pretraining for large language models,' *arXiv preprint arXiv:2311.16079*, 2023.

[8] Michael Moor *et al.*, 'Foundation models for generalist medical artificial intelligence,' *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

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

[10] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo and Yusuke Iwasawa, 'Large language models are zero-shot reasoners,' *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[11] Jason Wei *et al.*, 'Chain-of-thought prompting elicits reasoning in large language models,' *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[12] Daya Guo *et al.*, 'Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,' *arXiv preprint arXiv:2501.12948*, 2025.

[13] Niklas Muennighoff *et al.*, *S1: Simple test-time scaling*, 2025. arXiv: `2501.19393 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2501.19393`.

[14] Omar Khattab *et al.*, 'Dspy: Compiling declarative language model calls into self-improving pipelines,' *arXiv preprint arXiv:2310.03714*, 2023.

[15] Alistair EW Johnson *et al.*, 'Mimic-iii, a freely accessible critical care database,' *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[16] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang and Peter Szolovits, *What disease does this patient have? a large-scale open domain question answering dataset from medical exams*, 2020. arXiv: `2009.13081 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2009.13081`.

[17] Ankit Pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu, *Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering*, 2022. arXiv: `2203.14371 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2203.14371`.

[18] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen and Xinghua Lu, '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)*, 2019, pp. 2567–2577.

[19] DeepSeek-AI *et al.*, *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*, 2025. arXiv: `2501.12948 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2501.12948`.

[20] Dan Hendrycks *et al.*, 'Measuring massive multitask language understanding,' *arXiv preprint arXiv:2009.03300*, 2020.

A Appendix

A.1 Other results

Model	MedMCQA Log	MedQA Log	PubMedQA Log	MMLU Med Log	Mean Log
Medicouenne	57.23	61.9	74.2	74.74	67.02
Medicouenne s1.1	56.73	61.43	75.2	76.72	67.52
Medicouenne medical-deepseek-only	54.98	60.96	75.6	77.96	67.38
Medicouenne combined	54.53	60.02	74.6	77.36	66.63
Medicouenne s1.1 no-reasoning	57.92	60.96	75.4	78.33	68.15
Medicouenne medical-deepseek-only no-reasoning	59.38	61.59	75.8	77.22	68.5
Medicouenne combined no-reasoning	58.31	61.67	75.6	76.58	68.04
Meditron	60.29	62.84	79.0	75.67	69.45
Meditron s1.1	57.71	60.96	78.6	75.85	68.28
Meditron medical-deepseek-only	57.28	60.64	78.4	74.29	67.65
Meditron combined	56.56	60.49	79.2	72.91	67.29
Meditron s1.1 no-reasoning	56.97	61.90	78.6	75.48	68.24
Meditron medical-deepseek-only no-reasoning	57.45	61.04	78.2	74.47	67.79
Meditron combined no-reasoning	56.61	60.33	78.2	74.47	67.40

Table A.1: Meditron and Medicouenne variants results on Logits evaluation on medical benchmarks. Log stands for Logits based evaluation. medical-deepseek-only stands for the Medical-QA-reasoning dataset.

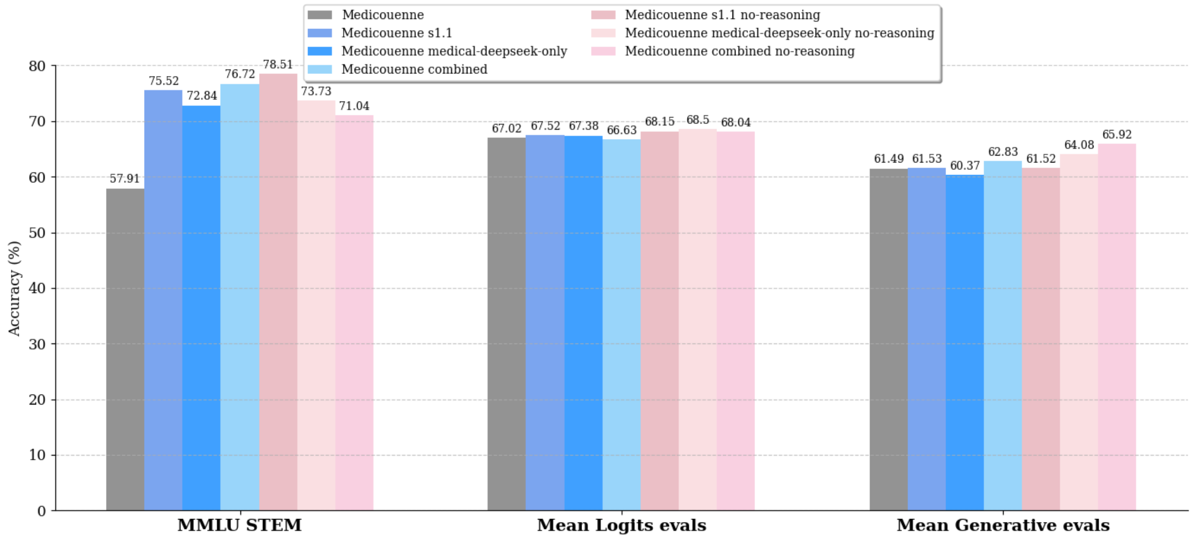


Figure A.1: Benchmark performance of Medicouenne variants. medical-deepseek-only stands for the Medical-QA-reasoning dataset.

A.2 Reasoning generation example 1

MedMCQA Prompt

Q: A mother is Hbs Ag positive and anti Hbe Ag positive. Risk of transmission of Hep. B in child is: (A) 20% (B) 50% (C) 0% (D) 90%A: Provide the final answer enclosed in boxed{the answer}.

Meditron Answer

Explanation:Ans. (C) 0%(Ref: Harrison 19/e p 1930, 18/e p 2001)Risk of transmission of Hepatitis B in child is 0% if mother is Hbs Ag positive and anti Hbe Ag positive. Answer: C

Meditron no-reasoning s1.1 Answer

The mother is HBsAg positive and HBeAg positive, indicating active viral replication and high infectivity. The risk of transmission of Hepatitis B to the child is significantly high in this scenario.
Key Points:

1. **HBsAg (Hepatitis B surface antigen)**: Indicates current infection.
2. **HBeAg (Hepatitis B e-antigen)**: Suggests high viral replication and infectivity.
3. **Mother's status**: Active replication (HBeAg positive) increases the viral load, significantly raising the transmission risk.
4. **Transmission risk**: Without intervention (e.g., vaccination, antiviral treatment), the risk of perinatal transmission can be as high as 90%.

Conclusion: The risk of transmission in this case is **90%**. $\boxed{D}$

Meditron s1.1 Thinking

Okay, so I need to figure out the risk of a child getting hepatitis B from a mother who is HbsAg positive and anti-HBeAg positive. Let me recall what each of these markers means. HbsAg is the hepatitis B surface antigen, right? That usually indicates current infection with hepatitis B. The presence of HbsAg in the blood means the person is infectious because they are actively producing the virus. Anti-HBeAg is the antibody against the hepatitis B e-antigen. Now, the e-antigen is another protein produced by the virus. The presence of anti-HBeAg usually indicates that the virus is not replicating as actively.

[...]

So, anti-HBs would be protective for the mother, but the child's risk would depend on the mother's viral load. Since anti-HBeAg indicates lower viral load, the child's risk is lower. Yes, that makes sense. So, the presence of anti-HBeAg in the mother reduces the viral load, which in turn reduces the risk of transmission. So, without any intervention, the risk is around 20%. So the answer is A) 20%. But wait, let's cross-check. If the mother is HbsAg positive, even with anti-HBeAg, the child is still at risk. The presence of anti-HBeAg doesn't eliminate the risk, but lowers it. So, from 90% maybe to 20%. So the answer is A. Yes, I think the correct answer is 20%. So option A.

Meditron s1.1 Answer

The mother's status as HBsAg positive and anti-HBeAg positive indicates current infection with hepatitis B, but with lower viral replication. Vertical transmission risk from an HBsAg-positive mother is significantly reduced with anti-HBeAg positivity. Without prophylactic measures (e.g., hepatitis B immunoglobulin (HBIG) and vaccination), the estimated risk of transmission is approximately **20%**. This is because anti-HBeAg suggests a lower viral load, reducing the likelihood of perinatal transmission.

Final Answer: $\boxed{A}$

Answer: $\boxed{A}$

A.3 Reasoning generation example 2

MedQA Prompt

Q: A 35-year-old man comes to the physician because of itchy, watery eyes for the past week. He has also been sneezing multiple times a day during this period. He had a similar episode 1 year ago around springtime. He has iron deficiency anemia and ankylosing spondylitis. Current medications include ferrous sulfate, artificial tear drops, and indomethacin. He works as an elementary school teacher. His vital signs are within normal limits. Visual acuity is 20/20 without correction. Physical examination shows bilateral conjunctival injection with watery discharge. The pupils are 3 mm, equal, and reactive to light. Examination of the anterior chamber of the eye is unremarkable. Which of the following is the most appropriate treatment? (A) Erythromycin ointment (B) Ketotifen eye drops (C) Warm compresses (D) Fluorometholone eye dropsA: Provide the final answer enclosed in boxed{the answer}.

Meditron Answer

Answer: B

Meditron s1.1 no-reasoning Answer

The patient is a 35-year-old man with itchy, watery eyes and sneezing, symptoms consistent with allergic conjunctivitis. His occupation as an elementary school teacher exposes him to allergens, which likely triggers his symptoms, especially during springtime.

Key considerations:

1. **Diagnosis**: Allergic conjunctivitis is likely due to environmental allergens, common in spring.
2. **Medication Review**:
    - **Ferrous sulfate** for iron deficiency anemia.
    - **Artificial tear drops** for dry eye relief.
    - **Indomethacin** for ankylosing spondylitis pain management.
3. **Treatment Options**:
    - **Erythromycin ointment (A)** is an antibiotic, not indicated for allergic conjunctivitis.
    - **Ketotifen eye drops (B)** are antihistamine drops that relieve itching and redness, appropriate for allergic conjunctivitis.
    - **Warm compresses (C)** can provide temporary relief for conjunctival injection but are not a primary treatment.
    - **Fluorometholone eye drops (D)** are corticosteroids, reserved for severe cases or when other treatments fail due to potential side effects.

Most Appropriate Treatment: Ketotifen eye drops (B) are the best choice as they directly address the allergic symptoms (itching and redness) without the side effects associated with corticosteroids.

Answer: $\boxed{B}$

The Meditron s1.1 Answer is on the next page.

MedQA Prompt

Q: A 35-year-old man comes to the physician because of itchy, watery eyes for the past week. He has also been sneezing multiple times a day during this period. He had a similar episode 1 year ago around springtime. He has iron deficiency anemia and ankylosing spondylitis. Current medications include ferrous sulfate, artificial tear drops, and indomethacin. He works as an elementary school teacher. His vital signs are within normal limits. Visual acuity is 20/20 without correction. Physical examination shows bilateral conjunctival injection with watery discharge. The pupils are 3 mm, equal, and reactive to light. Examination of the anterior chamber of the eye is unremarkable. Which of the following is the most appropriate treatment? (A) Erythromycin ointment (B) Ketotifen eye drops (C) Warm compresses (D) Fluorometholone eye dropsA: Provide the final answer enclosed in boxed{the answer}.

Meditron s1.1 Thinking

Okay, let's see. The question is about a 35-year-old man with itchy, watery eyes and sneezing. He had a similar episode a year ago in springtime. He has iron deficiency anemia and ankylosing spondylitis, and he's taking ferrous sulfate, artificial tear drops, and indomethacin. He works as an elementary school teacher. The physical exam shows bilateral conjunctival injection with watery discharge. The options are erythromycin ointment, ketotifen eye drops, warm compresses, or fluorometholone eye drops. First, I need to figure out what's causing his symptoms. The key here is probably allergies, right? The symptoms—itchy, watery eyes, sneezing—are classic for allergic conjunctivitis. The fact that he had a similar episode in springtime also points to seasonal allergies. Plus, he works in an elementary school, which is full of allergens like dust, pollen, and maybe even pet dander if there are classroom pets. But wait, he's also taking indomethacin.

[...]

But the symptoms here are itchy and watery, which are more consistent with allergies than with NSAID-induced conjunctivitis. Therefore, even though the patient is on indomethacin, the primary issue seems allergic, so treating with antihistamines (ketotifen) would be more appropriate. Another point: the patient works in an elementary school, which is a high-exposure environment for allergens. So even if the indomethacin contributed a bit, the main trigger is likely environmental allergens. Therefore, ketotifen eye drops are the way to go. **Final Answer** boxed{B}

Meditron s1.1 Answer

The patient presents with itchy, watery eyes and sneezing, symptoms consistent with allergic conjunctivitis, particularly given the seasonal pattern (springtime exacerbation) and occupational exposure as an elementary school teacher. Key findings include bilateral conjunctival injection and watery discharge, which align with allergic reactions rather than infectious or inflammatory causes.

Analysis:

1. **Symptomatology**: Itchy, watery eyes and sneezing are classic signs of allergic conjunctivitis.
2. **Environmental Triggers**: Working in an elementary school exposes him to common allergens (pollen, dust, pet dander).
3. **Previous Episode**: Similar symptoms last spring reinforce a seasonal allergic etiology.
4. **Medications**: Indomethacin may contribute to dry eye but is less likely to cause the current symptoms; artificial tears manage dryness but not allergies.

Treatment Options:

- **A. Erythromycin ointment**: Antibiotic; inappropriate for allergic conjunctivitis.
- **B. Ketotifen eye drops**: Antihistamine; effectively targets histamine-mediated itching and discharge.
- **C. Warm compresses**: More useful for inflammatory conditions like blepharitis; less effective for allergies.
- **D. Fluorometholone eye drops**: Corticosteroids; potentially harmful and unnecessary for allergies.

Conclusion: Ketotifen eye drops (Option B) are the most appropriate treatment as they directly address the allergic component by blocking histamine, providing rapid relief from itching and watery discharge. $\boxed{B}$

A.4 Reasoning generation example 3

MedMCQA Prompt

Q: A 29 yrs old woman with a pregnancy of 17 week has a 10 years old boy with down syndrome. She does not want another down syndrome kid; best advice to her is (A) No test is required now as her age is below 35 years (B) Ultra sound at this point of time will definitely tell her that next baby will be down syndromic or not (C) Amniotic fluid samples plus chromosomal analysis will definitely tell her that next baby will be down syndromic or not (D) blood screening at this point of time will clear the exact picture. A: Provide the final answer enclosed in boxedthe answer.

Medicouenne Answer

Understanding Down Syndrome Risk
Down syndrome is a genetic disorder caused by an extra copy of chromosome 21. The risk of having a child with Down syndrome increases with maternal age.
Factors Influencing Down Syndrome Risk

- **Maternal Age**: The risk of having a child with Down syndrome increases significantly after the age of 35.
- **Previous Child with Down Syndrome**: If a woman has already had a child with Down syndrome, she is more likely to have another child with the condition.

Current Scenario

- **Age of the Mother**: The mother is 29 years old, which is below the age of 35.
- **Previous Child with Down Syndrome**: The mother has a 10-year-old son with Down syndrome.

Best Advice for the Mother
Given the current scenario, the best advice for the mother would be to undergo prenatal testing to determine the risk of having another child with Down syndrome.
Prenatal Testing Options

- **Blood Screening**: Blood tests can detect certain markers associated with Down syndrome, but they are not definitive and may require further testing.
- **Amniocentesis**: This involves taking a sample of amniotic fluid from the uterus to analyze the chromosomes. It is a more accurate test but carries a small risk of miscarriage.
- **Chorionic Villus Sampling (CVS)**: This involves taking a sample of tissue from the placenta to analyze the chromosomes. It is also a more accurate test but carries a slightly higher risk of miscarriage than amniocentesis.

Conclusion Considering the mother's age and previous history, the best advice would be to undergo prenatal testing to determine the risk of having another child with Down syndrome.
Answer: (C) Amniotic fluid samples plus chromosomal analysis will definitely tell her that next baby will be down syndromic or not.

Medicouenne combined no-reasoning Answer

Axonal transport involves the movement of vesicles and other cargos along axons, which can occur in two directions: antegrade (from the cell body towards the axon terminal) and retrograde (from the axon terminal back towards the cell body). This bidirectional transport is essential for neuronal function. answer{C}

The Medicouenne medical-deepseek-only Answer is on the next page.

MedMCQA Prompt

Q: A 29 yrs old woman with a pregnancy of 17 week has a 10 years old boy with down syndrome. She does not want another down syndrome kid; best advice to her is (A) No test is required now as her age is below 35 years (B) Ultra sound at this point of time will definitely tell her that next baby will be down syndromic or not (C) Amniotic fluid samples plus chromosomal analysis will definitely tell her that next baby will be down syndromic or not (D) blood screening at this point of time will clear the exact picture. A: Provide the final answer enclosed in boxedthe answer.

Medicouenne medical-deepseek-only Thinking

Okay, let's tackle this question. So, we have a 29-year-old woman who's 17 weeks pregnant and already has a 10-year-old son with Down syndrome. She doesn't want another child with Down syndrome. The question is asking for the best advice among the four options given.First, I need to recall the standard prenatal screening methods for Down syndrome. Down syndrome is caused by trisomy 21, right? So chromosomal abnormalities. The usual screening tests include maternal serum screening (blood tests) and ultrasound scans. But the timing matters here because some tests are more effective earlier in the pregnancy.At 17 weeks, the options might be limited. Let me go through each option:Option A says no test is needed because she's under 35. Wait, age 29 is still considered low risk, but just because someone is younger doesn't mean they can skip testing. Many women still opt for testing, especially if there's a family history or previous child with DS. So A seems incorrect.Option B claims ultrasound at this point will definitely tell. I remember that detailed ultrasounds around 18-22 weeks can detect structural anomalies associated with DS, like congenital heart defects, but they can't confirm the diagnosis. It's more about ruling out other issues rather than confirming trisomy 21. So B might not be accurate.Option C mentions amniotic fluid samples with chromosomal analysis. That sounds like amniocentesis, which is typically done after 15-20 weeks. Amniocentesis can diagnose chromosomal abnormalities, including Down syndrome. However, the question is whether it's definitelytelling. Maybe, but there's always a small chance of false negatives or errors. So C could be a candidate.Option D refers to blood screening. At 17 weeks, the first trimester combined screening (triple screen or quad screen) is available, but its effectiveness depends on combining maternal serum markers with nuchal translucency from ultrasound. However, the blood screening alone might not be definitive at this stage. Also, cell-free DNA testing (non-invasive prenatal testing, NIPT) is becoming more common and can detect trisomy 21 with high accuracy from as early as 10 weeks. Since she's 17 weeks, NIPT would be possible. But the options don't mention NIPT specifically. Blood screening here might refer to the standard serum markers, which have lower sensitivity compared to later tests. So D might not be the best choice.Putting it all together: The best advice would be to perform a test that can provide a definitive diagnosis. Amniocentesis (option C) is a diagnostic test, not just screening. Blood tests like NIPT (which isn't listed here) would be more accurate at this stage, but since those options aren't available, the closest is C. However, some sources might say that amniocentesis is the gold standard and can definitively confirm or rule out Down syndrome. So even though there's a tiny risk of error, it's still the most reliable method mentioned here.Wait, but the options don't include NIPT. If NIPT is available and recommended, but since it's not listed, maybe the intended answer is C. Alternatively, perhaps the answer is C because amniocentesis is the definitive test. But I need to check if at 17 weeks, amniocentesis is the appropriate test. Yes, amniocentesis is usually done between 15-20 weeks, so 17 weeks is within that window. So performing amniocentesis (option C) would give a definite result, whereas blood tests and ultrasounds are less certain.Therefore, the best answer is C. The other options are either incorrect (A, B) or not the most accurate (D). Even though some newer tests like NIPT exist, since they're not listed, the most appropriate answer from the given options is C.

Medicouenne medical-deepseek-only Answer

The best advice for the patient is to undergo amniotic fluid analysis via amniocentesis, which involves chromosomal analysis (karyotyping or molecular tests) to definitively diagnose or rule out Down syndrome (trisomy 21) in the fetus. While non-invasive prenatal testing (NIPT) using cell-free DNA is highly accurate and available, it is not listed among the options. Among the provided choices, **(C)** most accurately reflects the definitive diagnostic approach at 17 weeks gestation, as ultrasound and blood screening lack the certainty offered by invasive genetic testing. answer{C}

A.5 Synthetic generation example 1

Multimodal synthetic patient record example

Patient history

Allergies

No known drug allergies

Medications

- Atorvastatin 40 mg daily
- Metformin 1000 mg daily
- Lisinopril 20 mg daily
- Aspirin 81 mg daily

Past Medical History

- Hypertension
- Type 2 Diabetes Mellitus
- Hyperlipidemia
- Gout (managed with allopurinol)
- History of myocardial infarction (5 years ago)

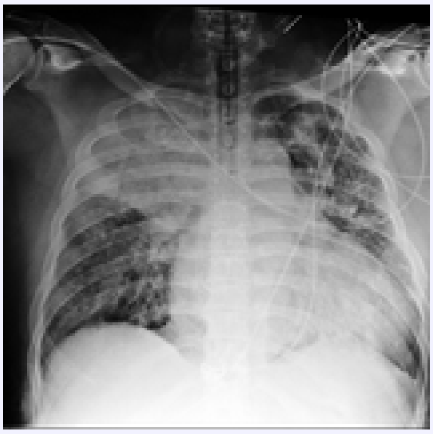
Family History

- Father with hypertension and heart disease
- Mother with type 2 diabetes

Social History

- Occupation: Senior Project Manager at **Company 2567**
- Drugs: Denies any illicit drug use
- Tobacco: Former smoker, quit 10 years ago
- Alcohol: Drinks socially, approximately 2–3 glasses of wine per week

Modality



Problematic

Chief Complaint

cough and shortness of breath

HPI

The patient is a 52-year-old male with a history of hypertension, type 2 diabetes mellitus, hyperlipidemia, gout, and a prior myocardial infarction, presenting with cough and shortness of breath that have progressively worsened over the past three days. He reports a productive cough with yellowish sputum and states that he feels increasingly fatigued. The patient also mentions experiencing fever and chills, which began the same day his symptoms worsened. He denies any chest pain but notes that he feels tightness in his chest when he coughs. He has not had any recent travel or known sick contacts but did mention that he had a cold about two weeks ago that seemed to resolve.

On arrival to the ED, the patient's vital signs were as follows: T 100.8°F, BP 130/85 mmHg, HR 92 bpm, RR 20 breaths per minute, and O2 saturation 92% on room air. A chest X-ray was performed, revealing bilateral infiltrates consistent with aspiration pneumonia. The patient was started on broad-spectrum IV antibiotics and given supplemental oxygen to maintain his oxygen saturation. He was advised to remain in a semi-upright position and was monitored closely for any signs of respiratory distress.

Assessment and Plan

52-year-old male with a history of hypertension, type 2 diabetes, hyperlipidemia, and a prior myocardial infarction presenting with cough, fever, and shortness of breath, diagnosed with aspiration pneumonia. Initiate IV antibiotics and supportive care, monitor respiratory status closely.