

Llama-3-Meditron: An Open-Source Suite of Medical LLMs Based on Llama-3.1

Anonymous submission

Abstract

We introduce Llama-3-Meditron, a high-performing open-source suite of medical LLMs built on Llama-3.1 (8B and 70B). Our models are pre-trained on a carefully curated medical corpus, including textbooks, filtered PubMed Central articles, and Clinical Practice Guidelines, following an enhanced Meditron methodology. To enable robust reasoning and generalization, we synthesize a novel dataset for instruction fine-tuning, combining multi-turn Q&A, adversarial questions, medical exams, and differential diagnostics. Additionally, we propose MediTree, an inference pipeline leveraging the Tree-of-Thoughts sampling strategy, to boost the performance of the smaller 8B model. On popular benchmarks (MedMCQA, MedQA, PubMedQA), Llama-3-Meditron-8B surpasses all Llama-3.1 models by over 3%, while our 70B model outperforms competitors medical LLMs across nearly all tasks, outperforming Meditron 1 and 2, GPT-4 (fine-tuned), Flan-PaLM, and MedPaLM-2. Notably, our model only lags GPT-4-Base on these tasks and slightly underperforms MedPaLM-2 on MedMCQA. These findings demonstrate that open-source medical LLMs can set new benchmarks in physician-level question-answering, advancing the accessibility and utility of AI in healthcare.

Introduction

Access to medical knowledge and expertise is crucial for delivering high-quality healthcare, especially in low-resource settings where shortages of medical professionals are common. Recent advancements in large language models (LLMs) have demonstrated that AI can perform proficiently on medical question answering tasks (Liévin et al. 2024; Rajpurkar et al. 2022; Singhal et al. 2023a). This has spurred significant interest in the development of medical LLMs with the eventual goal of achieve physician-level capabilities.

Closed-source medical models such as the MedPaLM family (Anil et al. 2023) and even non-specialized LLMs including GPT-4, have achieved impressive performance on popular benchmarks such as MedMCQA (Pal, Umapathi, and Sankarasubbu 2022), MedQA (Jin et al. 2020a), and PubMedQA (Jin et al. 2019). However, there has been an enthusiastic response from the research community to develop open-source medical LLMs such as the BioMistral (Labrak et al. 2024), PMC-Llama (Wu et al. 2024), and Meditron (Chen et al. 2023) families. PMC-Llama, adapted from

Llama 2 (Touvron et al. 2023), was specialized to the medical domain through continued pre-training on PubMed Central (PMC) articles (Roberts 2001). Meditron-70B improved on this approach by performing continued pre-training on a richer source of medical data, including filtered PMC data, medical textbooks, and Clinical Practice Guidelines (CPG).

In this work, we introduce a new family of models, Llama-3-Meditron, based on Llama-3 (Dubey et al. 2024). The marked improved language capabilities of the Llama-3 herd provide us with a base superior to the foundation of other open-source models built on top of Llama-2. We adopt the continued pre-training methodology of Meditron-70B, but develop an improved instruction fine-tuning phase by utilizing novel Q&A datasets reformatted from DDXPlus (Tchango et al. 2022) and MedlinePlus (Miller, Lacroix, and Backus 2000), standard training splits of MedQA, MedMCQA, and PubMedQA augmented with explanations, and adversarial question-answering. To further improve the quality of the 8B model, we develop MediTree, a Tree-of-Thoughts inspired pipeline co-designed with clinicians to leverage the problem-solving ability of large language models (LLMs) for differential diagnosis.

In particular, the 8B model equipped with MediTree is capable of outperforming substantially larger models such as Meditron-70B. The 70B model is even more promising, outperforming Med-PaLM 2, and surpassing the base Llama 3 70B model by over 2% on average, while trailing GPT-4-Base by approximately 3%.

Methodology

In this section, we detail our training strategy to develop Llama-3-Meditron.

Pre-training

We constructed our pre-training dataset from a variety of authoritative medical information sources, aiming to cover both general medical knowledge and specialized clinical guidelines.

- **PubMed Central Articles:** We included peer-reviewed articles from PubMed Central (PMC) (pmc), focusing on high levels of evidence such as meta-analyses, systematic reviews, randomized controlled trials, practice guidelines, and Phase III/IV clinical trials. Articles tagged

with "Animal" or "Veterinary" were excluded to maintain clinical relevance. The selection and filtration process was rigorously validated by medical doctors from the *Centre Hospitalier Universitaire Vaudois (CHUV)*.

- **Medical Textbooks:** To provide a solid foundation of medical knowledge, we incorporated validated medical textbooks covering various specialties, including genetics, oncology, infectious diseases, and pain management. Recognizing the challenges of extracting text from PDFs, we utilized advanced tools like Surya (Paruchuri and Surya Contributors 2024), enabling us to extract approximately 34 million tokens of high-quality text.
- **Clinical Practice Guidelines:** We incorporated clinical practice guidelines (CPGs) from globally recognized entities, following the approach of (Chen et al. 2023). CPGs represent the pinnacle of evidence-based medical data, synthesizing expert analyses to offer crucial guidance for clinical decision-making. Our guidelines corpus includes 46,000 articles spanning multiple medical domains and catering to diverse geographic scopes, including both high- and low-resource settings.

Instruction Tuning Data

To enhance the models' ability to follow instructions and perform complex medical tasks, we fine-tuned them on a custom instruction-tuning dataset. This dataset was designed to make the models more useful for real-world interactions and to improve their knowledge extraction capabilities.

- **Patient Progression Dialogue Dataset:** This multi-turn chat dataset tracks patients' conditions throughout their hospital stay or across a series of appointments. Constructed from the PMC-Patient dataset's discharge summaries (<https://pmc-patients.github.io/>), it simulates interactions where the assistant suggests medical tests or treatments based on initial symptoms, and the user provides results or feedback. This setup mirrors the iterative diagnostic process in clinical practice.
- **Symptoms to Diagnosis QA:** We reformatted the DDx-Plus dataset (Tchango et al. 2022) to create a question-answering dataset where, given a list of symptoms in natural language, the model outputs a differential diagnosis containing potential diseases. This enhances the model's diagnostic reasoning capabilities.
- **Questions for Diagnosis Generation:** Another reformatting of the DDxPlus dataset, this component focuses on generating relevant diagnostic questions given a differential diagnosis. This task improves the model's ability to suggest pertinent questions, aiding in the diagnostic process.
- **Health and Lab Tests Topics QA:** We scraped the MedlinePlus website (med 2024) to construct a multi-turn question-answering dataset covering various health topics and medical tests. This dataset enriches the models' knowledge base and improves their ability to handle patient inquiries.
- **Exam MCQA:** We combined the training sets from MedQA (Jin et al. 2020a), MedMCQA (Pal, Umapathi, and Sankarasubbu 2022), and PubMedQA (Jin

et al. 2019) to create a comprehensive multiple-choice question-answering dataset with standardized formatting. To enhance instruction tuning stability, we processed these datasets through Llama 3.1 70B using web-based retrieval-augmented generation (RAG), generating explanations along with answers. This approach increased the amount of tokens in 80% of the samples, providing richer context and improving the models' performance on complex tasks.

- **Adversarial QA:** We created a synthetic dataset that critiques answers to exam questions, pointing out potential shortcomings to discredit them. This task trains the model to self-reflect and recognize incorrect or suboptimal responses, enhancing its reliability.

Following recommendations from (Longpre et al. 2023), we diversified system prompts and included few-shot examples in approximately 50% of the samples, improving the models' capabilities in both zero-shot and few-shot settings. We also incorporated a portion of the AlpacaReplay dataset to broaden the range of learned tasks and mitigate over-fitting. A summary of the instruction tuning dataset is provided in Table 1.

Training Infrastructure

We conducted training of both the 8B and 70B models on a high-performance computing cluster. Each node was equipped with 8 NVIDIA A100 SXM GPUs with 80GB of memory, connected via NVLink and NVSwitch within nodes. For inter-node communications, we have a 2-port ThinkSystem Mellanox ConnectX-6 Dx 100GbE QSFP56 Ethernet Adapter per node, utilizing RoCE to speed up communications.

After considering the number of tokens, the size of the models, and the number of nodes available for training, we ultimately decided against using 3D parallel training frameworks such as Megatron (Shoeybi et al. 2019), which would have required a significant amount of hours to implement all the features we wanted to experiment with. Instead, we opted to use multiple Hugging Face libraries: *transformers* (Wolf et al. 2020) for the models and checkpoints, *datasets* (Lhoest et al. 2021) for preprocessing and feeding data to the models during training, and *accelerate* (Gugger et al. 2022) to shard the models among multiple GPUs. For the latter, we leveraged the DeepSpeed integration included in *accelerate*, specifically DeepSpeed ZeRO-3 (Rajbhandari et al. 2020).

Experiments

In this section, we assess the medical question-answering abilities of Llama-3-Meditron in comparison to other well-known models.

Selected Benchmarks

We selected three well-known medical question-answering benchmarks. MedQA (Jin et al. 2020b) and MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) evaluate the accuracy and reasoning abilities of models in diagnosing medical conditions based on clinical information and established

Dataset	Type	Samples	Percentage (%)
Patient Progression Dialogue	Multi-turn	86,000	14
Symptoms to Diagnosis QA	Single-turn	10,000	1
Questions for Diagnosis Generation	Single-turn	24,000	4
Health and Lab Tests Topics QA	Multi-turn	3,000	0.6
Exam MCQA	MCQA with CoT	397,000	62
Adversarial QA	Single-turn	32,000	2
AlpacaReplay	Single-turn	52,000	8
Total		607,000	100

Table 1: Summary of the instruction tuning dataset.

medical knowledge. These datasets use simple multiple-choice answer evaluation strategy. PubMedQA (Jin et al. 2019) evaluates the model on a more theoretical medical knowledge. This is also a multiple-choice dataset. To systematically run these benchmarks, we used a (Gao et al. 2023).

Llama-3-Meditron evaluation

We compared our model Llama-3-Meditron to several other medical LLMs. We compared it to Llama-2[7B] and Llama-2[70B], Meditron-7B, Meditron-70B (Chen et al. 2023), MedPalm 2 (Singhal et al. 2023b) and GPT-4 (Base and fine-tuned). We observe that our 8B model achieves high performance in the 7B/8B category. On average, Llama-3[8B]-Meditron beats all models and achieves similar results to Llama-2-70B. More detailed results can be found in Table 2.

Meditree Inference Pipeline

In this section, we discuss MediTree, a novel inference pipeline designed with the supervision of trained medical doctors from start to finish. Built for helping clinical decision-making, the pipeline leverages the problem-solving ability of large language models (LLMs) for differential diagnosis. We take inspiration from the diagnostic approach used by medical doctors, the LLM sampling in the *Tree of Thoughts* architecture presented by Yao & al. (Yao et al. 2023) and the *Med-Gemini* architecture (Saab et al. 2024). The differential diagnosis (DDx) approach provides a systematic method to identify a disease and determine appropriate treatment, especially when numerous alternative diagnoses are possible. See Section for an example.

The input to the MediTree pipeline is a patient case description. The pipeline then iteratively calls four components:

- **Chat:** Interact directly with the user to add more contextual information to the patient description.
- **Generation:** Proposes a thought to elaborate on or suggests a probable diagnosis.
- **Evaluation:** Assesses the pertinence of each proposed diagnosis.
- **Selection:** Chooses the best diagnosis to explore using the evaluation results or determine the end of the pipeline if the confidence is high enough.

Chat. The chat component adds additional context through an interactive process that mimics a medical evaluation. The model aims to evaluate the temporalization, quality, and quantification of symptoms by asking questions to the patient, similar to the way medical doctors do. The model is prompted to ask questions to the user in an interactive way, to further describe the main characteristics of the symptoms. At the end of the interaction user model, the model is prompted to update the patient description based on the new information collected. This patient description/patient note serves as both the input to the pipeline and is also updated to reflect the new state of the patient, for example, including the results of any medical test. It is composed of four parts :

- **Introduction:** A brief introduction to the patient and their illness, injury, or condition.
- **Symptoms:** Observed or detectable signs, and experienced symptoms of an illness, injury, or condition.
- **Treatments/Tests:** Information of previous or current medical therapy and medical tests conducted on the patient.
- **Medical history:** Details involving the patient, and eventually people close to them, to gather reliable/objective information for managing the medical diagnosis and proposing efficient medical treatments.

Doctor guidance for the prompts. To mimic how doctors investigate a patient’s case, we craft questions to address specific points:

- Temporalization, location of the symptoms and their particular characteristics.
- The patient’s previous treatments and behaviors affecting the symptoms.
- Understanding the patient’s pain, including its nature and intensity.
 - Quantifying the pain level on a scale from 1 to 10.
 - Describing the kind of pain felt.
- Other contexts that might influence the patient’s condition, such as:
 - Geographical context.
 - Location and recent travels.
- The patient’s personal and family medical history and current medication.

Model	Accuracy (\uparrow)			
	MedmcQA	MedQA	PubmedQA	Average
Llama2-70B-Base Inst	43.08	49.73	76.80	56.54
Llama2-70B-Base	47.93	57.42	74.40	59.92
Llama 3 8B Instruct	56.99	60.25	74.20	63.81
Llama 3 8B	57.52	60.00	74.80	64.11
Meditron 70B	53.30	59.80	79.80	64.30
Llama-3-Meditron 8B (ours)	57.83	63.00	76.80	65.88
Flan-Palm	57.60	67.60	79.00	68.07
Meditron 2 70B	65.10	65.40	80.00	70.17
Meditron 2 70B - CoT	63.20	67.80	81.00	70.67
GPT-4	69.50	78.80	75.20	74.50
Meditron 2 70B - CoT/SC	66.70	75.80	81.60	74.70
Llama 3 70B	70.00	78.40	77.00	75.13
Llama 3 70B Instruct	70.01	76.36	79.81	75.39
MedPalm 2	71.30	79.70	79.20	76.73
Llama-3-70B (ours)	70.10	80.75	81.00	77.28
GPT4-Base	73.66	86.10	80.40	80.05

Table 2: **Performance of Selected Models on Medical QA Benchmarks.** This table shows the accuracy of the selected models on three medical QA benchmarks: PubMedQA, MedMCQA, and MedQA-4-Option.

Base	Instruct (Gen)	ContPre	ContPre+Instr	ContPre + prompt	
Llama-2-70B-Base	Llama-2-70B-Instruct	Meditron 70B	Meditron 2 70B	Meditron 2 70B CoT*	
GPT4-Base	GPT-4			Medprompt	
Palm	Flan-Palm			MedPalm 2	
Llama 3 8B	Llama 3 8B Instruct	Llama-3-Meditron-3 8B	Meditron 3 8B Inst		
Llama 3 70B	Llama 3 70B Instruct	Llama-3-Meditron 70B	Meditron 3 8B Inst		
Average Gain	0	-0.93	1.65	5.87	8.43

Table 3: Comparison of different models and their respective configurations.

- Lifestyle factors, such as smoking and drinking, that can significantly impact the patient’s overall health.

Generation The generation component uses the mode *sample*, similar to the method presented in Yao & al. (Yao et al. 2023), to generate multiple diagnosis. This step involves producing multiple answers from the model with a high temperature ($temperature = 1.5$) to encourage diversity in the responses. To optimize inference time, sampling is performed using batch generation with a sampling size of 8, assuring a sufficiently large sample size. Each generation represents multiple possible diagnosis, and each diagnosis is identified by parsing the model answers.

Evaluation The evaluation component assigns a score to each possible diagnosis suggested in the answers. The score is calculated as the ratio of the number of times a particular diagnosis has been suggested to the total number of suggestions. This scoring method aims to approximate the probability of each opinion, using a sampling strategy to evaluate the model’s knowledge rather than relying on the raw logits.

Selection The selection component at the end of the pipeline is inspired by Med-Gemini (Saab et al. 2024). In this part, the entropy of each generation candidate is calculated using Shannon’s formula $H = -\sum_{i \in S} p_i \log_2(p_i)$. If the entropy value is higher than a predetermined threshold, indicating that the choice is not confident enough, resampling occurs. A new set of diagnoses is generated using a modified prompt, and this process is repeated until the en-

trophy falls below the threshold. The inference pipeline then outputs the diagnosis with the highest probability.

Conclusion

We release Llama-3-Meditron, a suite of open source medical LLM foundation models. In the continuity of the Open Meditron methodology, we crafted a high-quality dataset, using continued pretraining, instruction tuning, query tools, and alignment. We developed a novel inference pipeline, Meditree, that provides potential diagnoses and explores the most likely options, mimicking a doctor’s diagnostic approach. Our 8B model, tailored for low resource settings, is state of the art in his category, and performs comparable results to bigger models. Our 70B model achieved the best performances on public benchmarks, within 2% of the closed model GPT-4-Base.

References

- ???? PMC Open Access Subset [Internet]. Bethesda (MD): National Library of Medicine. 2024. MedlinePlus. Accessed: 2024-07-07.
- Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; Chu, E.; Clark, J. H.; Shafey, L. E.; Huang, Y.; Meier-Hellstern, K.; Mishra, G.; Moreira, E.; Omernick, M.; Robinson, K.; Ruder, S.; Tay, Y.; Xiao, K.; Xu, Y.; Zhang, Y.; Abrego, G. H.; Ahn, J.; Austin, J.; Barham, P.; Botha, J.; Bradbury, J.; Brahma, S.; Brooks, K.; Catasta, M.; Cheng, Y.; Cherry, C.; Choquette-Choo, C. A.; Chowdhery, A.; Crepy, C.; Dave, S.; Dehghani, M.; Dev, S.; Devlin, J.; Díaz, M.; Du, N.; Dyer, E.; Feinberg, V.; Feng, F.; Fienber, V.; Freitag, M.; Garcia, X.; Gehrmann, S.; Gonzalez, L.; Gur-Ari, G.; Hand, S.; Hashemi, H.; Hou, L.; Howland, J.; Hu, A.; Hui, J.; Hurwitz, J.; Isard, M.; Ittycheriah, A.; Jagielski, M.; Jia, W.; Kenealy, K.; Krikun, M.; Kudugunta, S.; Lan, C.; Lee, K.; Lee, B.; Li, E.; Li, M.; Li, W.; Li, Y.; Li, J.; Lim, H.; Lin, H.; Liu, Z.; Liu, F.; Maggioni, M.; Mahendru, A.; Maynez, J.; Misra, V.; Moussalem, M.; Nado, Z.; Nham, J.; Ni, E.; Nystrom, A.; Parrish, A.; Pellat, M.; Polacek, M.; Polozov, A.; Pope, R.; Qiao, S.; Reif, E.; Richter, B.; Riley, P.; Ros, A. C.; Roy, A.; Saeta, B.; Samuel, R.; Shelby, R.; Slone, A.; Smilkov, D.; So, D. R.; Sohn, D.; Tokumine, S.; Valter, D.; Vasudevan, V.; Vodrahalli, K.; Wang, X.; Wang, P.; Wang, Z.; Wang, T.; Wieting, J.; Wu, Y.; Xu, K.; Xu, Y.; Xue, L.; Yin, P.; Yu, J.; Zhang, Q.; Zheng, S.; Zheng, C.; Zhou, W.; Zhou, D.; Petrov, S.; and Wu, Y. 2023. PaLM 2 Technical Report. arXiv:2305.10403.
- Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A.; Sakhaeirad, A.; Swamy, V.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Montariol, S.; Hartley, M.-A.; Jaggi, M.; and Bosselut, A. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv:2311.16079.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac’h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2023. A framework for few-shot language model evaluation.
- Gugger, S.; Debut, L.; Wolf, T.; Schmid, P.; Mueller, Z.; Mangrulkar, S.; Sun, M.; and Bossan, B. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>.
- Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020a. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams.
- Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020b. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv:2009.13081.
- Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 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)*, 2567–2577.
- Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv:2402.10373.
- Lhoest, Q.; Villanova del Moral, A.; Jernite, Y.; Thakur, A.; von Platen, P.; Patil, S.; Chaumond, J.; Drame, M.; Plu, J.; Tunstall, L.; Davison, J.; Šaško, M.; Chhablani, G.; Malik, B.; Brandeis, S.; Le Scao, T.; Sanh, V.; Xu, C.; Patry, N.; McMillan-Major, A.; Schmid, P.; Gugger, S.; Delangue, C.; Matussière, T.; Debut, L.; Bekman, S.; Cistac, P.; Goehringer, T.; Mustar, V.; Lagunas, F.; Rush, A.; and Wolf, T. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 175–184. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Liévin, V.; Hother, C. E.; Motzfeldt, A. G.; and Winther, O. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).
- Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; and Roberts, A. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. arXiv:2301.13688.
- Miller, N.; Lacroix, E.-M.; and Backus, J. E. 2000. MEDLINEplus: building and maintaining the National Library of Medicine’s consumer health Web service. *Bulletin of the Medical Library Association*, 88(1): 11.
- Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. arXiv:2203.14371.
- Paruchuri, V.; and Surya Contributors. 2024. Surya (OCR library).
- Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. ZeRO: Memory optimizations Toward Training Trillion Parameter Models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16.
- Rajpurkar, P.; Chen, E.; Banerjee, O.; and Topol, E. J. 2022. AI in health and medicine. *Nature medicine*, 28(1): 31–38.
- Roberts, R. J. 2001. PubMed Central: The GenBank of the published literature.
- Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; Chaves, J. Z.; Hu, S.-Y.; Schaekermann, M.; Kamath, A.; Cheng, Y.; Barrett, D. G. T.; Cheung, C.; Mustafa, B.; Palepu, A.; McDuff, D.; Hou, L.; Golany, T.; Liu, L.; baptiste Alayrac, J.; Houlsby, N.; Tomasev, N.; Freyberg, J.;

Lau, C.; Kemp, J.; Lai, J.; Azizi, S.; Kanada, K.; Man, S.; Kulkarni, K.; Sun, R.; Shakeri, S.; He, L.; Caine, B.; Webson, A.; Latysheva, N.; Johnson, M.; Mansfield, P.; Lu, J.; Rivlin, E.; Anderson, J.; Green, B.; Wong, R.; Krause, J.; Shlens, J.; Dominowska, E.; Eslami, S. M. A.; Chou, K.; Cui, C.; Vinyals, O.; Kavukcuoglu, K.; Manyika, J.; Dean, J.; Hassabis, D.; Matias, Y.; Webster, D.; Barral, J.; Corrado, G.; Sementurs, C.; Mahdavi, S. S.; Gottweis, J.; Karthikesalingam, A.; and Natarajan, V. 2024. Capabilities of Gemini Models in Medicine. arXiv:2404.18416.

Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; and Catanzaro, B. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaeckermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Sementurs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023b. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617.

Tchango, A. F.; Goel, R.; Wen, Z.; Martel, J.; and Ghosn, J. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. arXiv:2205.09148.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Xie, W.; and Wang, Y. 2024. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, ocae045.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Appendix

MediTree Example

MediTree Example

MedQA question:

A 30-year-old G1P1 woman gives birth to a healthy infant at 39 weeks gestation. The mother attended all her prenatal visits and took all her appropriate prenatal vitamins. Her past medical history is notable for diabetes mellitus, for which she takes metformin. Apgar scores were 8 and 9 at 1 and 5 minutes, respectively. The infant's temperature is 99.2°F (37.3°C), blood pressure is 65/40 mmHg, pulse is 110/min, and respirations are 22/min. On exam, the child has ambiguous genitalia. A karyotype analysis demonstrates the presence of a Y chromosome. Further workup reveals internal testes with normal levels of serum luteinizing hormone and testosterone. Which of the following additional findings would most likely be seen in this patient? (A) Decreased androgen receptor responsiveness (B) Impaired sense of smell (C) Long extremities (D) Testosterone-to-dihydrotestosterone ratio > 1 (E) Testosterone-to-dihydrotestosterone ratio > 20 Give a step-by-step explanation then always answer the question with one of the options A, B, C, D or E (Answer: correct letter).

One of the answers:

Let's solve this step-by-step, referring to authoritative sources as needed. The infant has ambiguous genitalia and a Y chromosome, indicating XY sex chromosome anatomy. However, the presence of internal testes suggests an androgen biosynthetic defect rather than a receptor defect (option A) or XY complete sex reversal. Patients with XXXY sex reversal often have Kallmann syndrome (hypolfactory deficiency, hyposmia), but XY complete sex reversal typically does not manifest with olfactory deficiencies (option B). The diagnosis is likely congenital adrenal hyperplasia (CAH) with complete virilization due to a 21-hydroxylase deficiency. CAH is often characterized by tall stature (long extremities) due to premature epiphyseal ossification. Elevated testosterone-to-dihydrotestosterone (T:DHT) ratio is consistent with CAH, as a deficiency in the conversion of testosterone to DHT leads to increased androgen levels (option E). The T:DHT ratio is typically high (> 20) in classic forms of CAH, such as 21-hydroxylase deficiency. Answer: E

Results:

'D': 2, 'E': 5, 'A': 1

Entropy: 1.2

Probability of D: 0.25

Probability of E: 0.625

Probability of A: 0.125

Answer E is the correct answer. ✓

MedMCQA

Prompt:

A 45 yrs old man presents with progressive cervical lymph nodes enlargement, since 3 months; most diagnostic investigation is

- A. Xray soft tissue
- B. FNAC
- C. Lymph node biopsy
- D. None of the above

Rejected

I would choose option A. Xray soft tissue. Explanation: The patient presenting with progressive cervical lymph node enlargement suggests the possibility of malignancy, which would require further evaluation beyond just an X-ray of the soft tissue. Lymph node biopsy is the most appropriate diagnostic investigation in this case as it can provide a definitive diagnosis by examining the tissue sample for the presence of cancer cells. Additionally, an X-ray may not provide enough information to determine the underlying cause of the lymph node enlargement.

Answer: A. Xray soft tissue

Chosen

Explanation: Ans. is 'c' i.e., Lymph node biopsy Harrison writes that in cases of lymphadenopathy, if the patient's history and physical findings are suggestive of malignancy, then a prompt lymph node biopsy (excisional biopsy) should be done. FNAC is not of much use, as it does not provide enough tissue to reach a diagnosis. ("FNAC should be reserved for thyroid nodules and for confirmation of relapse in patients whose primary diagnosis is known." - Harrison)

Answer:C. Lymph node biopsy