



OctoMed: Data Recipes for State-of-the-Art Multimodal Medical Reasoning

Supplementary Material

A. Case Studies

We qualitatively analyze the performance of OctoMed on various multimodal medical queries. In particular, we note the following capabilities:

- **Task Versatility:** OctoMed is capable of various clinical reasoning tasks by simply changing the prompt. As shown in Figure S1-1 and Figure S2-2, OctoMed can provide advice on medication assessment or provide image reports.
- **Modality Versatility:** OctoMed is able to answer queries about images from many different imaging modalities, such as Pathology (Figure S2-1), Chest X-Ray (Figure S2-2), Dermatology (Figure S2-3), Fundus (Figure S2-4), and Text-Only (Figure S2-5).
- **Task-Aware Thinking:** OctoMed displays a propensity to adapt its reasoning trace length depending on the downstream task. For more challenging tasks such as MedXpertQA, OctoMed has longer average reasoning traces (Figure S3-1). However, simpler queries such as those from PMC-VQA result in shorter chain of thought length (Figure S3-2).

B. Question Sources for Evaluation Datasets

To comprehensively evaluate multimodal medical reasoning, we considered three benchmark categories. **Text-Only** tasks evaluate the model’s ability to apply its knowledge to solve multi-step clinical reasoning tasks without visual cues. **Multimodal Reasoning** benchmarks test the model’s capability to combine its text-only reasoning capabilities with visual evidence. **Multimodal Classification** tasks measure how well models can use perceptual grounding to extract clinical relevant features and arrive at diagnoses. Together, these categories provide comprehensive assessment of reasoning, modality integration, and perception, each critical for building a trustworthy multimodal reasoning model. We provide more details about the evaluation datasets we used for each task category below.

Text-Only

- **MedQA** [7] is a large-scale medical multiple-choice QA dataset drawn from U.S. medical licensing exams. We use the english subset which contains 12,723 English questions, with 10,178 for training, 1,272 for validation, and 1,273 for testing. Each question is text-only and has four answer options (one correct), making it a 4-way multiple-choice classification task. There is a corresponding 5-option version of the dataset, which we used to curate

our training data. We combine the training and validation splits of the 5-option version to use as SFT data, and evaluate on the 4 option version test split.

- **MMLU-PRO** (Health Subset) [16] is a challenging multiple-choice benchmark which builds off of the original MMLU benchmark [6] by expanding the number of multiple choice options to 10 per question. We evaluate on the health subset which contains 818 questions.
- **MedMCQA** [13] is a 4-option multiple-choice benchmark of real-world medical exams containing 182,822, 4,183, and 6,150 questions in the train, val, and test splits respectively.
- **MedXpertQA** (text) [19] is a comprehensive reasoning benchmark consisting of 17 medical specialties and 11 body systems. In the text-only portion of the dataset contain 2455 questions, each question has 10 multiple choice options. We only use this dataset for evaluation to test the generalization ability of OctoMed.
- **HeadQA** [15] contains questions from medical exams designed to evaluate readiness to access specialized portions of the Spanish medical system. There are 2657, 1366, and 2742 questions in the train, val, and test splits respectively. We merge the train and val splits for distillation and supervised finetuning, and evaluate on the unseen test split.

Multimodal Reasoning

- **PMC-VQA** [18] is a large-scale medical VQA benchmark consisting of over 227,000 image-question pairs and 149,000 unique images. The authors also provide a manually verified test split consisting of 2000 image-question pairs. We evaluate on this clean test split and carefully remove any questions from our training data with image overlap.
- **MedXpertQA** (multimodal) [19] contains 2000 questions related to advanced multimodal reasoning about various medical specialties and body systems. Each question is multiple choice and has 5 options.
- **MMMU-PRO** [17] is a widely used college-exam level multimodal reasoning benchmark. Each question is multiple choice and may refer to one or more images required to correctly arrive at the answer. We evaluate on the medical subset of the pro version of the dataset, which augments the answer choices to include up to 10 options. There are a total of 286 questions in the medical split.
- **NEJM Image Challenge** [12] is a weekly challenge hosted by the New England Journal of Medicine in which participants are tasked to perform differential diagnoses

to assess patient condition based on their case report summary and accompanying images. We follow prior work which scraped the questions from past weeks resulting in a total of 947 5-option multiple choice questions about multimodal patient differential diagnosis.

Multimodal Classification

- **Brain Tumor Classification** [3] is a MRI image dataset consisting of T1-Weighted images of brain tumors. The objective is to classify the tumor into one of 4 categories: Glioma, Meningioma, Pituitary, or no tumor. We follow the code in the CLIMB [4] codebase to obtain a train and test split of 2,870 and 394 samples respectively.
- **Coronahack** [11] is a Chest X-ray dataset in which models must predict if the patient has Bacterial Pneumonia, Viral Pneumonia, or is normal. We follow the code in the CLIMB codebase to obtain a train and test split of 5,284 and 624 records respectively.
- **Aptos** [2] is a fundus imaging dataset collected by the Asia Pacific Tele Ophthalmology Society Symposium. The task is to classify the diabetic retinopathy rating of a patient’s fundus image into one of 5 categories: No DR, Mild, Moderate, Severe, or Proliferative DR. Following the code in the CLIMB codebase, we obtain the same train test split of 2,929 and 733 image-question pairs respectively.
- **MESSIDOR-2** [5] Similar to Aptos, the MESSIDOR-2 dataset contains fundus images labelled with their corresponding diabetic retinopathy grades. We follow the CLIMB codebase to obtain a train-test split of 1,394 and 350 records.
- **BCSS** [1] is a crowd-sourced dataset of pathology slides of breast tissue. The task is to classify the tissue into one of 4 categories: Tumor, Stroma, Inflammatory, or Necrosis. We follow the CLIMB codebase to obtain our train-test split.

C. Prompts

Distillation Prompts Figure S3-3 and Figure S3-4 show the distillation prompt used to prompt GPT-4o and DeepSeek-R1 for multimodal and text-only tasks respectively.

Evaluation Prompts To ensure fair comparison, we did our best to evaluate all models using their suggested prompts. For transparency, we show the prompts used for each model in this section in Figures S3-5 to S3-8.

D. Additional Medical Benchmarks Results

Table 1 shows additional comparison between OctoMed and other models on benchmarks such as VQA-RAD and SLAKE. Many recent medical reasoning methods (e.g.,

Dataset	Ours	Chiron-o1	MedReason	ReasonMed
VQA-RAD	84.9	76.8	-	-
SLAKE	88.8	83.2	-	-
PathVQA	65.3	74.0	-	-
MedQA	90.8	48.2	68.4	66.9
MedMCQA	72.7	49.5	57.5	65.1
Aptos	75.4	37.0	-	-
Brain Tumor	80.9	38.8	-	-

Table 1. Performance of various medical chain of thought approaches on established medical benchmarks.

MedCoT[10]) rely on multi-stage pipelines with external teacher-generated rationales and are sometimes text-only, limiting cross-task generalization and making them not directly comparable to our single-model, multimodal setting (hence the empty table entries). We report results on VQA-RAD [8] and SLAKE [9] here, but note that these benchmarks are largely saturated and evaluated under inconsistent splits and metrics across prior work, motivating our focus on more recent and challenging benchmarks.

E. Discussion of Hallucination and Factuality

While rejection sampling alone does not guarantee faithful or clinically supported rationales, we believe our curated teacher reasoning traces provide a useful foundation for future work. In one sampled reasoning trace, DeepSeek-R1 produced the statement:

“... in medical texts, it’s described that long-standing obstruction leads to loss of nephrons and atrophy.”

While plausible, this claim would benefit from an explicit citation or supporting evidence. This example highlights the importance of developing stronger verification mechanisms for generated reasoning traces. A potential solution could be to add a citation or reference check in addition to answer correctness when filtering reasoning traces.

Training Setting	Text	MM Reasoning	CLS
Text-Only <small>(100k subset)</small>	2631	2160	2444
All Tasks <small>(300k subset)</small>	1609	2111	275
All 8M data	1156	1740	325

Table 2. Avg Trace lengths of various student models.

F. Task-Aware Thinking Further Analysis

We believe task-aware reasoning emerges from both task diversity and teacher diversity. As shown in Table 2, training

on a single task type with one teacher yields relatively uniform trace lengths, whereas training across tasks with multiple teachers produces task-dependent reasoning lengths, even with a smaller subset.

References

- [1] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467, 2019. 2
- [2] Asia Pacific Tele-Ophthalmology Society. Aptos 2019 blindness detection. <https://www.kaggle.com/c/aptos2019-blindness-detection/data>, 2019. [Dataset]. 2
- [3] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri). <https://www.kaggle.com/dsv/1183165>, 2020. [Dataset]. 2
- [4] Wei Dai, Peilin Chen, Malinda Lu, Daniel Li, Haowen Wei, Hejie Cui, and Paul Pu Liang. Climb: Data foundations for large scale multimodal clinical foundation models. *arXiv preprint arXiv:2503.07667*, 2025. 2
- [5] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234, 2014. 2
- [6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 1
- [7] 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. *Applied Sciences*, 11(14):6421, 2021. 1
- [8] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018. 2
- [9] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 2
- [10] Jiayang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024. 2
- [11] Nida Nasir, Afreen Kansal, Feras Barneih, Omar Al-Shaltona, Talal Bonny, Mohammad Al-Shabi, and Ahmed Al Shammaa. Multi-modal image classification of covid-19 cases using computed tomography and x-rays scans. *Intelligent Systems with Applications*, 17:200160, 2023. 2
- [12] New England Journal of Medicine. Nejm image challenge. <https://imagechallenge.nejm.org/>, 2025. Accessed: 2025-09-15. 1
- [13] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022. 1
- [14] Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. *arXiv preprint arXiv:2506.04353*, 2025. 6
- [15] David Vilares and Carlos Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy, 2019. Association for Computational Linguistics. 1
- [16] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhramil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024. 1
- [17] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 1
- [18] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 1
- [19] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025. 1

Medication Assessment Example

A 55-year old woman with active smoking history, type II diabetes, hypertension, history of coronary artery disease was recently hospitalized for heart failure with an ejection fraction of 35%. She is presenting for follow-up post discharge. Her discharge medications included lisinopril 10 mg daily, carvedilol 2.5 mg twice daily, and eplerenone 25 mg daily, in addition to her usual medications for diabetes. In clinic, she reports feeling symptomatic with dyspnea after going up 1 flight of stairs though is asymptomatic at rest. Vital signs at rest: BP 130/90, HR 75, RR 10, Temperature 97.2F, SpO2 98% on room air. Labs are remarkable for Na 132 mEq/L, K 4.9 mEq/L, BUN 22 mg/dL, Cr 1.04 mg/dL.

The patient reports having a friend on Entresto. Please review whether Entresto is indicated in this patient and why or why not.

OctoMed: <think> Okay, let's tackle this case step by step. So the patient is a 55-year-old woman with multiple comorbidities: active smoking, type II diabetes, hypertension, history of coronary artery disease, and now she's been hospitalized for heart failure with reduced ejection fraction (HFrEF) at 35%. She's on several meds including lisinopril, carvedilol, eplerenone, and others for diabetes.

First, I need to recall the guidelines for managing HFrEF. The mainstay therapies include ACE inhibitors (or ARBs if intolerant), beta-blockers, and mineralocorticoid receptor antagonists (MRAs) like eplerenone. Additionally, SGLT2 inhibitors have become important lately. But wait, the question is about Entresto, which is sacubitril/valsartan. That's an ARNI—angiotensin receptor-neprilysin inhibitor.

So, when is Entresto indicated? From what I remember, it's used in HFrEF patients as an alternative to ACE inhibitors or ARBs. The PARADIGM-HF trial showed that sacubitril/valsartan was better than enalapril in reducing cardiovascular death and hospitalization for HF. The indication typically requires NYHA class II-IV symptoms despite being on optimal therapy including an ACE-I or ARB, MRA, and beta-blocker. Also, they need to be stable hemodynamically, no hypotension, renal issues, hyperkalemia, etc.

Looking at this patient: She has HFrEF EF 35%, so definitely reduced. She's symptomatic with dyspnea on exertion (1st flight of stairs). Post-discharge, but maybe still in NYHA class II. Her current meds include lisinopril (ACE-I), carvedilol (beta-blocker), and eplerenone (MRA). So she's already on all three pillars. However, the next step might be substituting the ACE-I with Entresto if appropriate.

But there are contraindications. Let's check her labs: Sodium is low at 132, potassium 4.9 (which is okay, upper limit but acceptable). Creatinine is 1.04, BUN 22. So renal function seems okay, though creatinine isn't super high. She doesn't have significant hyperkalemia here. Her blood pressure is controlled at 130/90. No mention of angioedema or cough from lisinopril, so probably can switch.

Wait, but the guidelines say that Entresto should replace the ACE-I or ARB, not add to them. Since she's on lisinopril, switching to Entresto would make sense. But does she meet criteria? She has persistent symptoms despite being on those meds. The fact that she's still symptomatic (dyspnea on exertion) suggests that her current regimen isn't fully controlling her symptoms. Therefore, moving to Entresto could provide additional benefit.

Another consideration is that she has diabetes. Entresto can be used in diabetics, although monitoring for hyperkalemia and renal function is needed. Her potassium is normal, and her creatinine is slightly elevated but not critically. Her BUN is 22 mg/dL and her creatinine is 1.04 mg/dL. To estimate her GFR using the CKD-EPI formula for a 55-year-old female, we use: $eGFR = 141 \times (Scr)^{-1.209} \times (Age)^{-0.203} \times 0.742$. Assuming $Scr = 1.04$ mg/dL and age = 55 years, we approximate: $141 \times (1.04)^{-1.209} \approx 141 \times 0.82 \approx 114$, and then $114 \times 0.742 \approx 83$. Thus, her estimated GFR is approximately 83 mL/min/1.73 m². This corresponds to stage 1 chronic kidney disease but is considered mild, and therefore Entresto would be safe here unless other issues arise.

Also, she's on eplerenone. Combining MRAs with Entresto increases risk of hyperkalemia. Her potassium is 4.9, which is within normal range (normal up to 5.0 or 5.1 depending on lab). So maybe manageable with close monitoring. Guidelines recommend avoiding combination unless necessary, but in some cases, benefits outweigh risks. Given her EF is 35%, and symptoms persisting, adding Entresto while continuing eplerenone may be considered, but perhaps with caution.

Wait, but standard practice is to substitute ACE-I/ARB with Entresto. If she's already on eplerenone and carvedilol, then replacing lisinopril with Entresto would be the way to go. Then continue eplerenone. Alternatively, sometimes MRA can be held temporarily when starting Entresto due to hyperkalemia risk, but recent studies suggest that continuation may be possible with monitoring. The 2022 AHA/ACC/HFSA guidelines state that combining Entresto with MRA is reasonable in certain patients, especially if they're stable, but requires close monitoring of potassium and renal function.

In this case, since her potassium is normal and renal function is preserved, adding Entresto while keeping eplerenone might be acceptable. However, some clinicians prefer to hold the MRA for a week after initiating Entresto and monitor. But given that she's symptomatic, the benefit of Entresto likely outweighs the risk.

Alternatively, if her potassium were higher, maybe hold eplerenone. But here, K is 4.9, which is just under the cutoff (usually 5.0 or 5.5). So perhaps proceed with Entresto substitution, keep eplerenone, monitor K closely.

Other factors: she's an active smoker. Smoking cessation counseling is essential. Also, optimizing diabetes control, encouraging exercise, low-sodium diet, etc. But focusing on medication: Entresto is indicated here because she's on ACE-I, beta-blocker, MRA, and remains symptomatic. Replacing ACE-I with Entresto would align with guideline-directed medical therapy (GDMT) for HFrEF.</think>

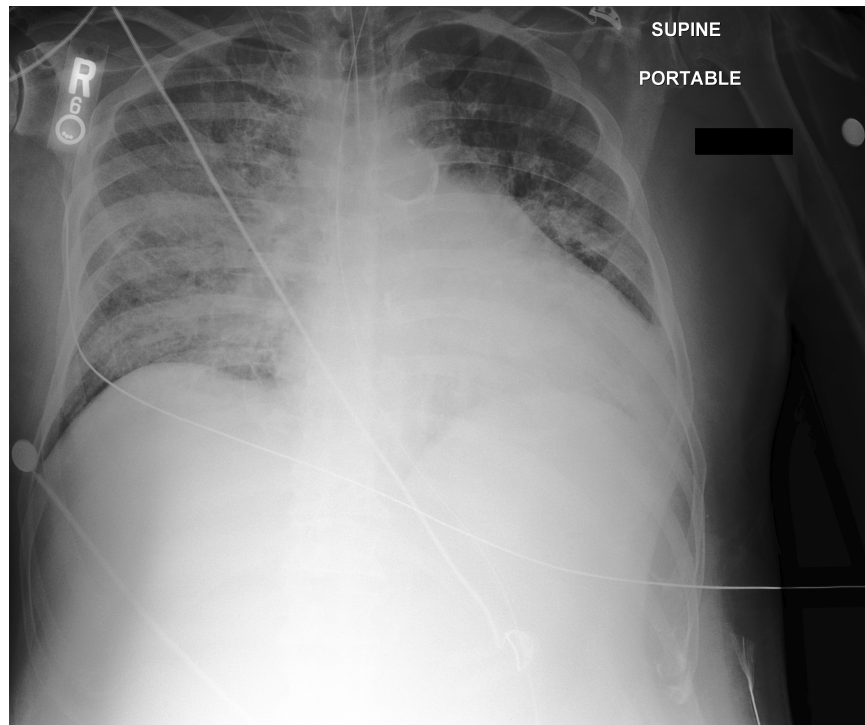
Entresto (sacubitril/valsartan) is indicated for this patient based on the following rationale:

Key Indications for Entresto:

- HFrEF with Persistent Symptoms:** The patient has heart failure with reduced ejection fraction (HFrEF; EF 35%) and remains symptomatic (dyspnea on exertion despite optimal therapy). This aligns with NYHA Class II symptoms, where Entresto improves outcomes compared to ACE inhibitors (e.g., enalapril).
- Guideline-Directed Medical Therapy (GDMT):** Current ACC/AHA/HFSA guidelines recommend replacing ACE inhibitors or ARBs with Entresto in HFrEF patients who remain symptomatic despite ≥ 3 months of evidence-based therapy (ACE-I + beta-blocker + MRA). This patient is on lisinopril (ACE-I), carvedilol (beta-blocker), and eplerenone (MRA), fulfilling these criteria.
- Superior Efficacy:** The PARADIGM-HF trial demonstrated Entresto reduces cardiovascular mortality and HF hospitalizations more effectively than enalapril. Substituting lisinopril with Entresto addresses residual symptoms and optimizes prognosis.

Figure S1-1. In addition to multiple-choice VQA, OctoMed can be prompted to perform other tasks such as medication assessment, shown above.

Chest X-Ray Report Generation Example



Describe the findings in this chest x-ray.

OctoMed: <think>When analyzing this chest X-ray, I begin by systematically examining the image to identify any abnormalities or significant findings.

1. **Image Orientation and Quality:**

The X-ray is labeled as "SUPINE" and "PORTABLE," indicating that it was taken with the patient lying down (supine position) using a portable machine. This positioning can affect the appearance of certain structures, such as the heart size and lung markings, so these factors must be considered during interpretation.

2. **Lines and Tubes:**

Multiple lines and tubes are visible: A central venous catheter (CVC) is seen entering from the right side and traveling toward the mediastinum. An endotracheal tube is present, likely used for mechanical ventilation. Additional monitoring electrodes and wires are also visible but do not interfere significantly with the assessment.

3. **Cardiac Silhouette:**

The cardiac silhouette appears enlarged. While supine films may exaggerate heart size due to magnification effects, the enlargement here seems disproportionate even accounting for positional changes. This suggests cardiomegaly or possible pericardial effusion.

4. **Pulmonary Findings:**

There is increased opacity throughout both lungs, particularly prominent in the lower zones bilaterally. These opacities appear patchy and diffuse, consistent with alveolar filling processes. The pattern of opacification resembles pulmonary edema, which could result from fluid accumulation in the interstitial and alveolar spaces. No clear focal consolidation or cavitation is noted, making pneumonia less likely as the primary cause of these findings. The vascular markings are more pronounced, suggesting vascular congestion, which aligns with pulmonary edema.

5. **Pleural Spaces:**

Both costophrenic angles are obscured, raising suspicion for pleural effusions. In the supine position, pleural effusions layer posteriorly and can obscure the diaphragms rather than creating the classic meniscus sign seen on upright films.

6. **Diaphragm and Abdomen:**

The diaphragmatic contours are poorly defined, especially on the left side, further supporting the presence of pleural effusions. Gas patterns within the abdomen suggest bowel loops filled with air, which are incidental findings unrelated to the thoracic pathology.

7. **Differential Diagnosis:**

The combination of bilateral diffuse opacities, obscured costophrenic angles, and an enlarged cardiac silhouette strongly suggests pulmonary edema secondary to congestive heart failure. Other considerations include acute respiratory distress syndrome (ARDS), though ARDS typically lacks significant cardiomegaly unless complicated by cardiac dysfunction.

8. **Summary:**

Based on the observed features—cardiomegaly, bilateral pulmonary opacities consistent with alveolar flooding, and evidence of pleural effusions—the most likely diagnosis is pulmonary edema. This conclusion fits well with the clinical context of a critically ill patient requiring intensive support (e.g., intubation and CVC).

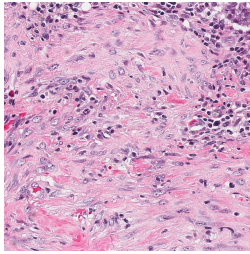
By carefully evaluating each component of the image step-by-step, I arrive at the final interpretation.

</think>

The chest x-ray shows signs of extensive bilateral pulmonary edema, characterized by widespread opacification of both lung fields, indicative of fluid accumulation within the alveolar spaces. Additionally, there is evidence of pleural effusions, as suggested by blunted costophrenic angles and obscuration of the diaphragmatic borders. Cardiomegaly is apparent, indicated by an enlarged cardiac silhouette. The presence of medical devices like the central venous catheter and endotracheal tube indicates the patient is receiving intensive care.

Figure S1-2. Example of prompting OctoMed to perform a Chest X-Ray report generation task.

Histopathology Example



Above is a histopathological image patch from breast cancer tissue. What type of tissue is shown in this image patch?

- A: Tumor
- B: Stroma (ground truth)
- C: Inflammatory
- D: Necrosis

Please reason step-by-step, and put your final answer within `\\boxed{}`.

OctoMed: `<think>` To determine the type of tissue shown in the histopathological image, I need to carefully analyze the visual characteristics. The image appears to be stained with hematoxylin and eosin (H&E), which is commonly used for histological examination.

First, I notice a dense pink background with spindle-shaped cells that are characteristic of stromal tissue. The pink areas represent extracellular matrix components, such as collagen, which are abundant in the stroma. The spindle-shaped cells are fibroblasts or myofibroblasts, which are also typical of stromal tissue.

Next, I observe scattered darker purple-stained nuclei. These nuclei are elongated and irregular, which is consistent with stromal cells rather than the round or pleomorphic nuclei typically seen in tumor cells. Tumor tissue would generally exhibit a higher degree of nuclear atypia, mitotic figures, and cellular crowding, which are not prominent here.

Inflammatory tissue would show a significant infiltration of immune cells such as lymphocytes, plasma cells, or neutrophils, which would appear as small, round, dark-stained nuclei. While there are some immune cells in the image, they are sparse and not the dominant feature.

Necrotic tissue would show areas of cell death with ghost-like cell outlines, loss of nuclear staining, and eosinophilic debris. There is no evidence of necrosis in this image, as the cells appear intact with visible nuclei and organized structure.

Based on these observations, the tissue in the image is most consistent with stroma, as it shows abundant extracellular matrix, spindle-shaped fibroblasts, and a lack of features characteristic of tumor, inflammatory, or necrotic tissue. `</think>`

`\\boxed{B: Stroma}`

Figure S2-1. Example of a multimodal classification based on histopathological tissue.

Chest X-Ray Example



What finding is observed over the left lateral chest on this chest X-ray?

- A: Subcutaneous emphysema (ground truth)
- B: Pleural effusion
- C: Rib fracture
- D: Mediastinal mass

Please reason step-by-step, and put your final answer within `\\boxed{}`.

OctoMed: `<think>` To analyze the findings in this chest X-ray, I need to carefully evaluate the image and consider each option systematically.

First, let me examine the area of interest: the left lateral chest. The question specifically asks about an abnormality in this region, so my focus will be there.

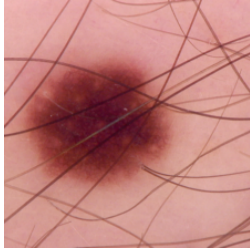
1. **Option A: Subcutaneous emphysema** Subcutaneous emphysema appears as streaky or radiolucent (dark) areas within the soft tissues due to air trapped under the skin. On the X-ray, I should look for unusual dark lines or pockets outside the bony thoracic cage, particularly in the soft tissue planes. In this image, there does appear to be some lucency extending along the left lateral chest wall, which could suggest subcutaneous emphysema. However, I need to rule out other possibilities before concluding.
2. **Option B: Pleural effusion** Pleural effusion typically manifests as blunting of the costophrenic angle or a meniscus-shaped opacity at the lung base. It would cause increased density (whiteness) rather than lucency. Here, the left lateral chest shows no significant fluid accumulation or opacification suggestive of pleural effusion, so this is less likely.
3. **Option C: Rib fracture** Rib fractures may show discontinuity or irregularity in the cortical outline of the ribs. While it can sometimes be subtle, I don't see any clear evidence of rib fractures in the left lateral chest on this X-ray. Additionally, rib fractures alone wouldn't explain the lucency seen here unless associated with secondary complications like pneumothorax or subcutaneous emphysema.
4. **Option D: Mediastinal mass** A mediastinal mass would present as a well-defined opacity in the central chest, near the heart or great vessels. This finding is not located laterally in the chest, making this option unlikely based on the location described.

Given these considerations, the most plausible explanation for the lucent streaks in the left lateral chest wall is subcutaneous emphysema. This conclusion aligns with the characteristic appearance of air tracking through the soft tissues. `</think>`

`\\boxed{A: Subcutaneous emphysema}`

Figure S2-2. Example response for a Chest X-Ray multiple-choice question from RexVQA [14].

Dermatology Example



Above is a dermoscopic image of a pigmented skin lesion. Classify the image into one of the following diagnostic categories. Answer with just one of the options below.

- A: Actinic keratoses and intraepithelial carcinoma
- B: Basal cell carcinoma
- C: Benign keratosis-like lesions
- D: Dermatofibroma
- E: Melanoma
- F: Melanocytic nevi (ground truth)
- G: Vascular lesions

Please reason step-by-step, and put your final answer within `\ \boxed{}`.

OctoMed: `<think>` To classify this dermoscopic image, I need to carefully analyze its features and compare them against typical characteristics associated with each diagnostic category. The lesion appears as a uniformly dark brown structure with an overall symmetrical appearance. There are no visible ulcerations, vascular structures, or irregular borders that might suggest malignancy like basal cell carcinoma (B) or melanoma (E).

Actinic keratoses and intraepithelial carcinoma (A) typically present as scaly, erythematous patches rather than a pigmented lesion like this one. Vascular lesions (G) usually exhibit red, purple, or blue hues due to blood vessel involvement, which is not evident here. Dermatofibromas (D) often show a central white scar-like area surrounded by a pigment network, but this lesion lacks such features.

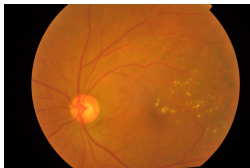
Benign keratosis-like lesions (C), including seborrheic keratosis, can have varied appearances, including warty or stuck-on textures, but they generally lack the uniform pigmentation seen here. This leaves us with melanocytic nevi (F), which are benign moles characterized by symmetric pigmentation, smooth borders, and uniform coloration—features consistent with this image.

Given the symmetry, even distribution of pigmentation, and absence of concerning malignant features, this lesion most closely resembles a melanocytic nevus. `</think>`

`\ \boxed{F: Melanocytic nevi}`

Figure S2-3. Example response for a multimodal classification question based on a dermoscopic image.

Fundus Example



Grade the diabetic retinopathy of the patient in the retinal image on the Davis Scale of Diabetic Retinopathy.

- A: No DR
- B: Mild DR
- C: Moderate DR (ground truth)
- D: Severe DR
- E: Proliferative DR

Please reason step-by-step, and put your final answer within `\ \boxed{}`.

OctoMed: `<think>` To assess the diabetic retinopathy grade of the patient using the Davis Scale, I must carefully analyze the retinal image provided. The Davis Scale categorizes diabetic retinopathy into five stages: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR.

Looking at the retinal image, I notice several key features. The optic disc appears normal, and the retinal vasculature is visible. However, there are yellowish-white spots scattered across the retina, particularly in the area away from the optic disc. These spots are likely hard exudates, which are lipid residues from leaky blood vessels. The presence of hard exudates suggests some level of diabetic retinopathy.

There are no signs of neovascularization, which would indicate Proliferative DR. Additionally, I do not observe extensive hemorrhages, venous beading, or intraretinal microvascular abnormalities (IRMA), which are hallmarks of Severe DR. The hard exudates and possible mild retinal changes suggest the condition is less severe than Severe DR.

To differentiate between Mild and Moderate DR, I need to consider the extent of the abnormalities. Mild DR is characterized by microaneurysms only, while Moderate DR includes additional features such as hard exudates, more prominent hemorrhages, or cotton-wool spots. The presence of hard exudates in this image suggests the condition is beyond Mild DR and falls into the Moderate DR category.

After careful consideration of the retinal features and the Davis Scale criteria, I conclude that the patient has Moderate DR. `</think>`

`\ \boxed{C: Moderate DR}`

Figure S2-4. Example response for a multimodal classification question based on a fundus image.

MedQA Example

A man is brought into the emergency department by the police department. The officer state that the man has been arrested multiple times for public alcohol intoxication, but recently became homeless. On exam, the man is behaving erratically. His vitals are all within normal limits. He appears confused and has a slurred speech. On gait exam, the patient is ataxic and cannot stand without support for more than a few seconds. Labs return with the following values: Na 140, K 4, Cl 106, BUN 8, Cr 2. His ABG has pH 7.3, PaCO₂ 13mm, PaO₂ 130mm, HCO₃ 7. His urinalysis is shown in Figure 1. Blood salicylate levels return as normal. While you await other diagnostic tests, which of the following should be administered next to treat this patient?

Answer with just one of the options below.

- A. Ethanol
- B. Naltrexone
- C. Naloxone
- D. Fomepizole (ground truth)

OctoMed: <think> Okay, let's try to figure out this case. So, the patient is a man with a history of public intoxication arrests and is now homeless. He's brought in by police, acting erratically, confused, slurred speech, and ataxic. Vitals are normal. Labs show normal sodium, potassium, chloride, BUN, creatinine. ABG is pH 7.3 (which is acidic), PaCO₂ 13 mmHg (low, so respiratory alkalosis?), HCO₃ 7 (very low, metabolic acidosis). Wait, pH is 7.3, which is acidotic. So combining the low HCO₃ and low PaCO₂. Hmm, that's a mixed acid-base disorder? Because in metabolic acidosis, the body compensates by hyperventilating to lower PaCO₂. Let me check the anion gap. $Na - (Cl + HCO_3) = 140 - (106 + 7) = 140 - 113 = 27$. That's a high anion gap metabolic acidosis. Anion gap of 27. So the patient has a high anion gap metabolic acidosis. The ABG pH is 7.3, which is acidotic. The PaCO₂ is 13, which is low. So the expected compensation for metabolic acidosis would be using Winter's formula: $expected\ PaCO_2 = (1.5 \times HCO_3) + 8 \pm 2$. Let's compute that. HCO₃ is 7, so $1.5 \times 7 = 10.5 + 8 = 18.5$. So expected PaCO₂ should be around 18.5 ± 2 . The patient's PaCO₂ is 13, which is lower than expected. That suggests there's an additional respiratory alkalosis component. So maybe a mixed high anion gap metabolic acidosis and respiratory alkalosis? But why the respiratory alkalosis? Could be due to hyperventilation from the acidosis, but the PaCO₂ is lower than expected. Hmm. Alternatively, maybe there's another process causing the respiratory alkalosis, like salicylate poisoning? Wait, but salicylate levels are normal. The urinalysis is shown in Figure 1, but since we can't see it, maybe it's something else. The patient is homeless, so maybe alcohol-related issues? But the blood ethanol isn't mentioned here. The options given are ethanol, naltrexone, naloxone, fomepizole.

The high anion gap metabolic acidosis could be due to several causes: methanol, ethylene glycol, diabetic ketoacidosis, lactic acidosis, etc. Since the patient is homeless and has a history of intoxication, maybe he ingested something toxic. The options include fomepizole, which is an antidote for methanol and ethylene glycol poisoning. Ethanol is also used as an antidote for those because it competes for the enzyme alcohol dehydrogenase, preventing the conversion of the toxic alcohol into its toxic metabolites. Naltrexone is for alcohol dependence, and naloxone is for opioid overdose.

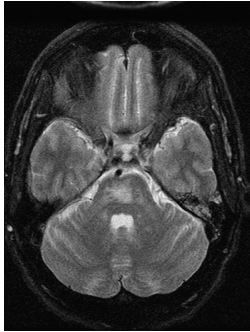
Given the high anion gap metabolic acidosis, and possible toxic alcohol ingestion, fomepizole or ethanol would be the antidotes. Since blood salicylate levels are normal, that's ruled out. So between ethanol and fomepizole. Fomepizole is preferred over ethanol because it's easier to administer and has fewer side effects. So the answer might be D. Fomepizole. But wait, the ABG shows a respiratory alkalosis as well. Could that be a clue? In ethylene glycol poisoning, you can have an initial respiratory alkalosis due to stimulation of the respiratory center, followed by metabolic acidosis. Similarly, methanol can cause similar effects. So if the patient has both high anion gap acidosis and respiratory alkalosis, maybe it's ethylene glycol or methanol. So the treatment would be fomepizole. The other options don't fit. Naltrexone is for alcohol dependence maintenance, not acute intoxication. Naloxone is for opioids. Ethanol could be used as an antidote, but fomepizole is more appropriate. So the answer is D.</think>

The patient presents with a high anion gap metabolic acidosis (anion gap = 27) and a mixed acid-base disorder (pH 7.3, PaCO₂ 13 mmHg, HCO₃ 7). The low HCO₃ and elevated PaCO₂ suggest metabolic acidosis with respiratory compensation, but the PaCO₂ is lower than expected, indicating a superimposed respiratory alkalosis. This pattern is consistent with toxic alcohol ingestion (e.g., ethylene glycol or methanol), which can cause an initial respiratory alkalosis due to direct stimulation of the respiratory center, followed by metabolic acidosis from toxic metabolites. The patient's history of homelessness and prior alcohol-related arrests raises suspicion for accidental or intentional ingestion of toxic alcohols.

Given the normal salicylate levels and lack of opioid involvement, naloxone (C) is not indicated. Naltrexone (B) is used for alcohol dependence management, not acute toxicity. Ethanol (A) could theoretically compete with toxic alcohols for alcohol dehydrogenase but is less favorable due to side effects and monitoring challenges. **Fomepizole (D)***, a competitive inhibitor of alcohol dehydrogenase, is the preferred antidote for toxic alcohol poisoning, as it prevents the formation of toxic metabolites while being safer and easier to administer than ethanol.
Answer: \boxed{D}

Figure S2-5. Example of a detailed reasoning trace on a challenging text-only MedQA question.

MedXpertQA Example



This patient was in the recovery phase of diabetic ketoacidosis (DKA). What is the diagnosis?

Answer Choices:

- (A) Basilar artery thrombosis
- (B) Osmotic demyelination syndrome (ground truth)
- (C) Cobalamin deficiency
- (D) Vermian atrophy
- (E) Cerebral edema

Please reason step-by-step, and put your final answer within `\ \ boxed{}`.

OctoMed: `<think>` Okay, let's try to figure out this question. The patient is in the recovery phase of diabetic ketoacidosis (DKA), and we need to determine the diagnosis from the given options. First, I should recall what DKA is and its possible complications, especially during recovery.

DKA is a serious complication of diabetes where the body starts breaking down fat at a rapid rate, leading to acidic ketones in the blood. It's usually treated with insulin and fluids, but sometimes there are complications during recovery.

Now, looking at the MRI image provided, it shows a T2-weighted axial view of the brain. I need to analyze the findings here. The image seems to show symmetrical hyperintensities in the pons, which might be relevant. Hyperintensities on T2 could indicate edema, demyelination, or other pathological processes.

Let me consider each option:

A. Basilar artery thrombosis: This would present with acute stroke symptoms. On MRI, you might see restricted diffusion in the affected area. However, the image shows bilateral symmetrical involvement, which is less typical for a thrombosis. Also, basilar artery issues often present with more acute deficits, not necessarily in the recovery phase of DKA.

B. Osmotic demyelination syndrome (ODS): Also known as central pontine myelinolysis. This is a known complication of rapid correction of hyponatremia or other metabolic disturbances. DKA treatment involves aggressive fluid administration, which can lead to rapid shifts in osmolality. ODS typically affects the pons, causing symmetrical lesions visible on MRI as hyperintensities on T2/FLAIR. The clinical scenario fits because the patient is recovering from DKA, which involves fluid and electrolyte shifts.

C. Cobalamin deficiency: This leads to subacute combined degeneration of the spinal cord, not typically pontine lesions. It's associated with megaloblastic anemia and neurological symptoms like ataxia, but the MRI findings here don't support this diagnosis.

D. Vermian atrophy: This refers to shrinkage of the cerebellar vermis, often seen in chronic conditions like alcoholism or certain neurodegenerative diseases. It's not related to DKA recovery and wouldn't present with acute changes in the pons.

E. Cerebral edema: While cerebral edema can occur in DKA, especially during treatment, it usually presents more diffusely and acutely. The MRI here shows localized lesions in the pons rather than diffuse swelling.

Putting this together, the key points are the recovery phase of DKA, the MRI showing symmetrical pontine lesions, and the association with osmotic shifts. Osmotic demyelination syndrome (B) is the most likely diagnosis. Basilar thrombosis doesn't fit the imaging pattern or the clinical context. Cobalamin deficiency and vermian atrophy don't match the presentation. Cerebral edema would be more widespread and acute. `</think>`

The patient is in the recovery phase of diabetic ketoacidosis (DKA), a condition known to cause significant metabolic and osmotic disturbances. During treatment, rapid correction of electrolyte imbalances—particularly sodium levels—can lead to **osmotic demyelination syndrome (ODS)**, previously termed central pontine myelinolysis. This condition is characterized by damage to the myelin sheaths in the central nervous system, most commonly affecting the pons, due to rapid shifts in osmolality.

Key Clinical and Radiological Features Supporting ODS: - **Clinical Context:** DKA management often involves aggressive rehydration and insulin therapy, which can lead to abrupt changes in serum osmolality. This creates a high-risk environment for ODS. - **Imaging Findings:** The MRI demonstrates symmetrical hyperintensities in the pons on T2-weighted imaging, which is classic for ODS. These lesions are typically bilateral and localized to the upper pons, consistent with the pathophysiology of demyelination following osmotic stress. - **Timing:** ODS typically manifests 2–4 days after the correction of hyponatremia or other metabolic derangements, aligning with the patient being in the **recovery phase** of DKA.

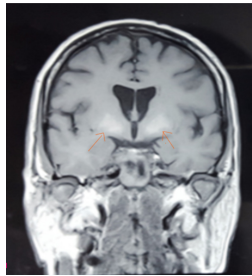
Evaluation of Other Options: - **A. Basilar artery thrombosis:** This would present with acute neurological deficits and restricted diffusion on MRI, not the symmetrical T2 hyperintensities seen here. - **C. Cobalamin deficiency:** This causes spinal cord pathology (e.g., subacute combined degeneration), not pontine lesions. - **D. Vermian atrophy:** This is a chronic, degenerative condition affecting the cerebellum, not an acute process related to DKA. - **E. Cerebral edema:** While possible in DKA, it would present with diffuse brain swelling and loss of gray-white differentiation, not localized pontine changes.

Conclusion: Given the clinical context of DKA recovery and the characteristic MRI findings, **osmotic demyelination syndrome** is the most likely diagnosis.

`\ \ boxed{B}`

Figure S3-1. Example of a detailed reasoning trace on a challenging MedXpertQA question. OctoMed displays longer reasoning traces for challenging questions.

PMC-VQA Example



What is the name of the medical imaging technique used in this case?
A: X-ray
B: Magnetic resonance imaging (ground truth)
C: Computed tomography
D: Ultrasound

Please reason step-by-step, and put your final answer within `\\boxed{}`.

OctoMed: `<think>` Okay, let's analyze the question and the provided image. The image shows a detailed view of the brain in a coronal section, with clear visualization of soft tissues, ventricles, and other intracranial structures. This level of detail, especially for soft tissues, is characteristic of Magnetic Resonance Imaging (MRI). X-rays (option A) are primarily used for imaging bones and do not provide this level of detail for soft tissues. Computed Tomography (CT, option C) can show soft tissues but is less detailed compared to MRI and often involves the use of ionizing radiation. Ultrasound (option D) is not suitable for imaging the brain in this manner due to the skull obstructing sound waves. Therefore, the medical imaging technique used in this case is Magnetic Resonance Imaging (MRI). `</think>`

`\\boxed{B: Magnetic resonance imaging}`

Figure S3-2. Example reasoning trace on a question from PMC-VQA. OctoMed displays shorter reasoning length for easier questions.

GPT-4o Distillation Prompt

You should provide your detailed thoughts within `<think>` `</think>` tags, always making sure to reflect and think about your response, then answer with just one of the options below within `<answer>` `</answer>` tags. Your response should carefully consider the options and output a very long chain of thought. (For example, if the question is 'Is the earth flat?')

A: Yes

B: No', you should answer with `<think>`Okay, let's tackle this question about whether the Earth is flat or not. The idea that the Earth is flat may feel intuitive because our everyday experience seems flat, but overwhelming evidence shows it's a sphere. First, astronomical observations reveal that stars rotate differently in the northern and southern hemispheres: Polaris is visible up north but not down south, which only makes sense on a curved surface. Wait, could that be due to perspective? No, let me double check; this change in visible stars directly correlates with latitude, which wouldn't happen on a flat plane. Ships also disappear bottom-first over the horizon; wait, maybe that's just perspective? But even with a telescope, the hull stays hidden, confirming it's curvature, not optics. Then there's air travel: planes follow great-circle routes, which look curved on flat maps but are the shortest path on a globe. Let me double check. Yes, for example, New York to Tokyo arcs over Alaska only because the Earth is round. During lunar eclipses, Earth always casts a round shadow on the Moon. Wait, could a flat disc do that? Only from one angle; a sphere is the only shape that does this consistently. And what about space photos? Are they fake? Let me double check. No, not just NASA, but independent agencies and private companies would all have to be complicit, and their satellite systems rely on spherical Earth physics to work, including GPS. Time zones also show curvature; when it's day in Tokyo, it's night in New York. Wait, could the Sun just be a spotlight above a flat Earth? That fails too; we'd see the Sun all the time just dimmer, not dipping below the horizon. Also, engineers designing long bridges or tunnels adjust for curvature, and GPS satellites wouldn't function without spherical Earth modeling. Let me double check|yes, geodetic surveying and orbital mechanics prove it. So from ancient Greek reasoning to modern engineering and spaceflight, every independent line of evidence confirms the Earth is not flat, but round.</think> `<answer>`B: No</answer>). Here is the question: {{ question }}

Figure S3-3. Prompt used with GPT-4o to curate reasoning traces for multimodal tasks.

DeepSeek-R1 Distillation Prompt

```
{{question}}

Put your final answer letter within
<answer></answer> tags.
```

Figure S3-4. Prompt used with DeepSeek-R1 to curate reasoning traces for text-only tasks.

OctoMed Evaluation Prompt

```
{{question}}

Please reason step-by-step, and put your final
answer within \\boxed{}.
```

Figure S3-5. Prompt used to evaluate OctoMed on multiple choice tasks.

Huatuo-GPT-Vision Evaluation Prompt

```
{{question}}

Answer with the option's letter from the given
choices directly.
```

Figure S3-6. Prompt used to evaluate Huatuo-GPT-Vision on multiple choice tasks. Since we found the model struggles to perform long chain of thought, we copy the template from their work directly.

LingShu Evaluation Prompt

```
Question: {{question}}
Answer with the option's letter from the given
choices and put the letter in one "\\boxed{}"
```

Figure S3-7. Prompt used to evaluate LingShu-7B on multiple choice tasks. We use the default reasoning prompt in their MedEvalToolkit without modifications, reproduced here for convenience.

MedVLThinker Evaluation Prompt

```
You will solve a problem/request. You should
provide your thoughts within <think> </think>
tags before providing the answer.
Write your final answer within <answer> </answer>
tags.
{{question}}
```

Figure S3-8. Prompt used to evaluate MedVLThinker on multiple choice tasks. We use the prompt in the appendix of their work, reproduced here for convenience.