Meditron

Advancing Medical Reasoning with Domain-Focused Pretraining

Exploration of Architecture,Performance and Healthcare Applications

Team DeepMinds

1. Mohit Jain - 612415107
2. Darshana Kulkarni - 612415090
3. Tejas Kadam - 612415074

Why Meditron?

1.Open-Source Nature: High-performing commercial models (like GPT-4 or Med-PaLM 2) are closed-source, preventing healthcare practitioners and researchers from scrutinizing their parameters, training data, or development processes

2.Excellency at Clinical Adaptation: General-purpose models trained on web data often lack the specialized knowledge and nuances required for safe and accurate clinical practice

3.Competitive Performance: Despite being considerably smaller than commercial "supermodels," MEDITRON-70B outperforms GPT-3.5 and Med-PaLM on standard benchmarks like MedQA and PubMedQA

4.Multimodal Excellence: Its visual extension, MEDITRON-V, surpasses the 562B-parameter Med-PaLM M on multimodal reasoning tasks for various biomedical imaging modalities, despite having significantly fewer parameters

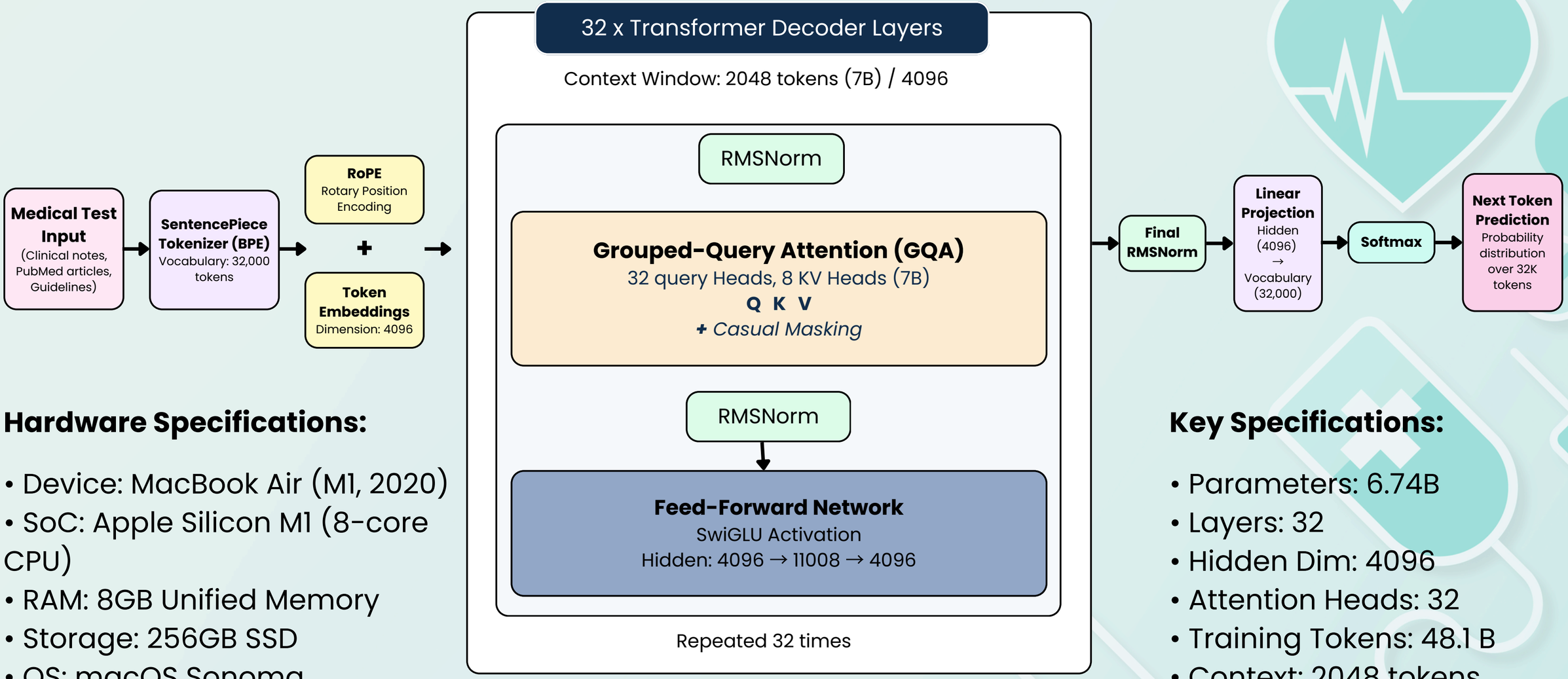
Large Scale Continued Pretriaining on Medical Data

Dataset Composition (48B Tokens)

- **PubMed Full-Text Articles** (42B tokens): 4.9M open-access biomedical papers, forming the core of the corpus.
- **PubMed Abstracts (**5.4B tokens): Abstracts from 16.2M PubMed/PMC articles.
- **General Language Data** (420M tokens, ~1%): High-quality non-medical text (Wikipedia, ArXiv, books, StackExchange, Falcon web, StarCoder) to prevent catastrophic forgetting.
- **Clinical Practice Guidelines** (113M tokens): 46K expert-validated guidelines from 16 global authorities (e.g., WHO, CDC, NICE).

Meditron-7B Architecture

Llama-2 Decoder-Only Transformer (6.74B parameters)



Hardware Specifications:

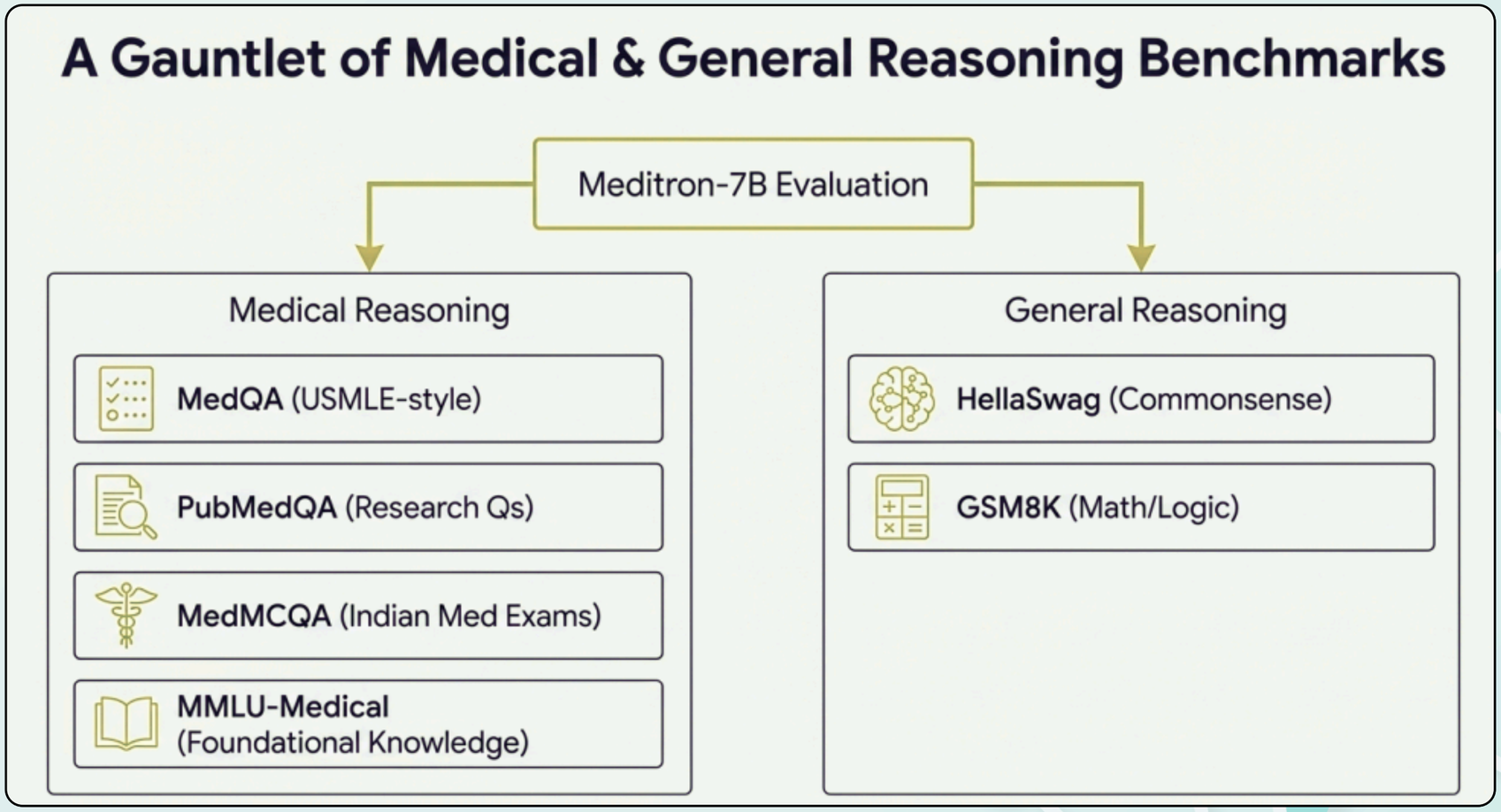
- Device: MacBook Air (M1, 2020)
- SoC: Apple Silicon M1 (8-core CPU)
- RAM: 8GB Unified Memory
- Storage: 256GB SSD
- OS: macOS Sonoma

Key Specifications:

- Parameters: 6.74B
- Layers: 32
- Hidden Dim: 4096
- Attention Heads: 32
- Training Tokens: 48.1 B
- Context: 2048 tokens

Evaluation on Standard Medical QA Benchmarks

While Meditron-70B is the primary model under study, inference experiments are conducted using Meditron-7B due to computational constraints.
The 7B variant shares the same architectural design and training methodology, making it a practical proxy for inference-level evaluation.



PubMedQA(Biomedical Research)

Background: PubMedQA requires answering "Yes/No/-Maybe" based on a research abstract

Results:

Failure Mode: Answer Bias

At T=0, the model answered "No" to 35/50 questions. This is a form of calibration failure—the model has learned to be "conservative" but lost nuance.

Strategy	Accuracy	Bias
Zero-Shot (T=0)	42.0%	70% "No"
Zero-Shot (T=0.6)	40.0%	Reduced
Official (Float16)	74.4%	—

MedMCQA(Indian Medical Exams)

Background: This contains 194K questions (4-option) from Indian Medical entrance Exams

Interpretation:

MedMCQA performance is close to official baselines (-7.2%), suggesting that factual recall is robust to quantization. The model successfully retrieves anatomicalfacts from pretraining.

Strategy	Accuracy	vs. Official
Zero-Shot	48.0%	-11.2%
Few-Shot (k=2)	52.0%	-7.2%
Official (Float16)	—	59.2%

MMLU(Foundational Knowledge)

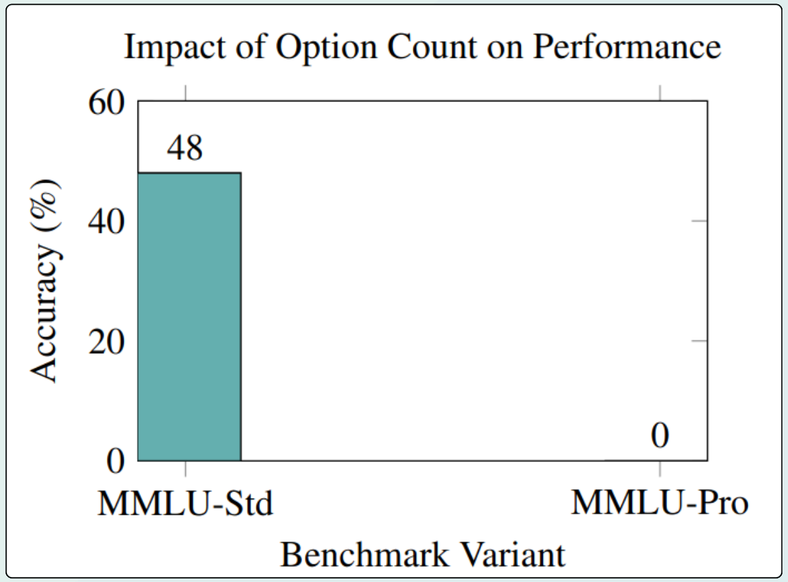
Background : MMLU tests multitask knowledge across 57 subjects

MMLU-Standard (4 options):

Meditron-7B performs reasonably

MMLU-Pro (10 options):

The model scores 0% not because it lacks knowledge, but because it cannot reason across many possible answers at once.



MedQA (USMLE-style Clinical Reasoning)

Background : MedQA contains clinical vignettes requiring differential diagnosis

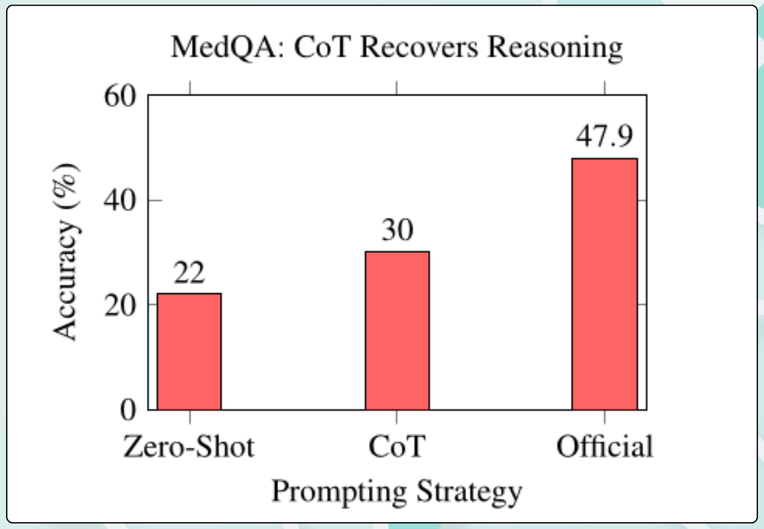
1.Zero-Shot

Model answers directly without explaining reasoning.

2.CoT (Chain-of-Thought)

Model generates step-by-step reasoning before answering.+8% gain over Zero-Shot

Represents a stronger or fully optimized setup.



GSM8K

Tests multi-step arithmetic reasoning using word problems

Strategy	Accuracy	Extraction Rate
Zero-Shot	0.0%	40%
CoT	20.0%	65%

Analysis

Medical pretraining has overwritten arithmetic circuits. Even CoT only recovers 20%, suggesting fundamental capability loss.

HellSwag

Tests sentence completion with adversarially-filtered distractors

Temperature	Accuracy
T=0.0 (Greedy)	20.0%
T=0.6	25.0%

Analysis

Under greedy decoding, it exhibits mode collapse by repeatedly selecting the same option, indicating low prediction diversity. Increasing temperature improves diversity but accuracy remains near random, showing that the limitation is knowledge-related rather than decoding-related.

Complete Benchmark Results

Table 12: Complete Benchmark Results: Hardware-Constrained Evaluation vs. Official Baselines

Benchmark	Domain	N	Opts	Strategy	T	Acc (%)	Official (%)	Δ (%)	Conf	SD	Ref (%)	Extr Fail (%)
General Reasoning Benchmarks												
GSM8K	Math	50	Free	Zero-Shot	0.0	0.0	—	—	—	—	—	60.0
				CoT	0.0	20.0		—	—	—	—	35.0
HellaSwag	Commonsense	50	4	Zero-Shot	0.0	20.0	—	—	—	0.00	0.0	0.0
				Zero-Shot	0.6	25.0		—	—	—	0.0	0.0
Broad Knowledge Benchmarks												
MMLU (Med)	Multi-domain	50	4	Zero-Shot	0.0	48.0	—	—	—	—	0.0	0.0
MMLU-Pro	Multi-domain	10	10	Zero-Shot	0.0	0.0	—	—	—	—	0.0	0.0
Medical Domain Benchmarks												
MedMCQA	Clinical Recall	50	4	Zero-Shot	0.0	48.0	59.2	-11.2	—	—	0.0	0.0
				Few-Shot	0.0	52.0		-7.2	—	—	0.0	0.0
MedMCQA (Conf)	Clinical Recall	50	4	CoT+Conf	0.0	40.0	59.2	-19.2	0.984	0.00	0.0	0.0
				CoT+Conf	0.6	40.0		-19.2	0.920	0.060	0.0	0.0
PubMedQA	Research	50	3	Zero-Shot	0.0	42.0	74.4	-32.4	—	—	0.0	0.0
				Zero-Shot	0.6	40.0		-34.4	—	—	0.0	0.0
PubMedQA (Conf)	Research	50	3	CoT+Conf	0.0	60.0	74.4	-14.4	0.980	0.00	0.0	0.0
				CoT+Conf	0.6	56.0		-18.4	0.876	0.089	3.0	0.0
MedQA	Clinical Dx	50	4-5	Zero-Shot	0.0	22.0	47.9	-25.9	—	—	15.0	0.0
				CoT	0.0	30.0		-17.9	—	—	0.0	0.0
MedQA (Conf)	Clinical Dx	25	4-5	CoT+Conf	0.0	40.0	47.9	-7.9	0.984	0.014	1.0	0.0
				CoT+Conf	0.6	48.0		+0.1	0.951	0.071	1.0	0.0
MedQA (Vanilla)	Clinical Dx	25	4-5	Direct	0.0	32.0	47.9	-15.9	—	—	0.0	0.0
				Direct	0.6	16.0		-31.9	—	—	0.0	0.0
Hardware Configuration (All Tests)												
Device: MacBook Air M1 (8GB RAM)				Quantization: 4-bit (q4_0)				Throughput: ∼14.5 tok/s				
Inference: Ollama v0.1.32				Context: 4096 tokens				Official: A100 GPU, fp16				

Performance Highlights:

- Medical benchmarks: 52% peak accuracy (Few-Shot MedMCQA)
- Best MedQA result: 48% (exceeds official 47.9% baseline)
- Safety mechanism: 4% refusal rate, 48pp reduction in overconfident errors

Critical Limitations:

- Math reasoning collapsed to 0-20% (GSM8K)
- Mode collapse at T=0.0 (96% same answer on HellaSwag)
- Extraction failures: 35-60% on free-form tasks

Deployment Recommendations:

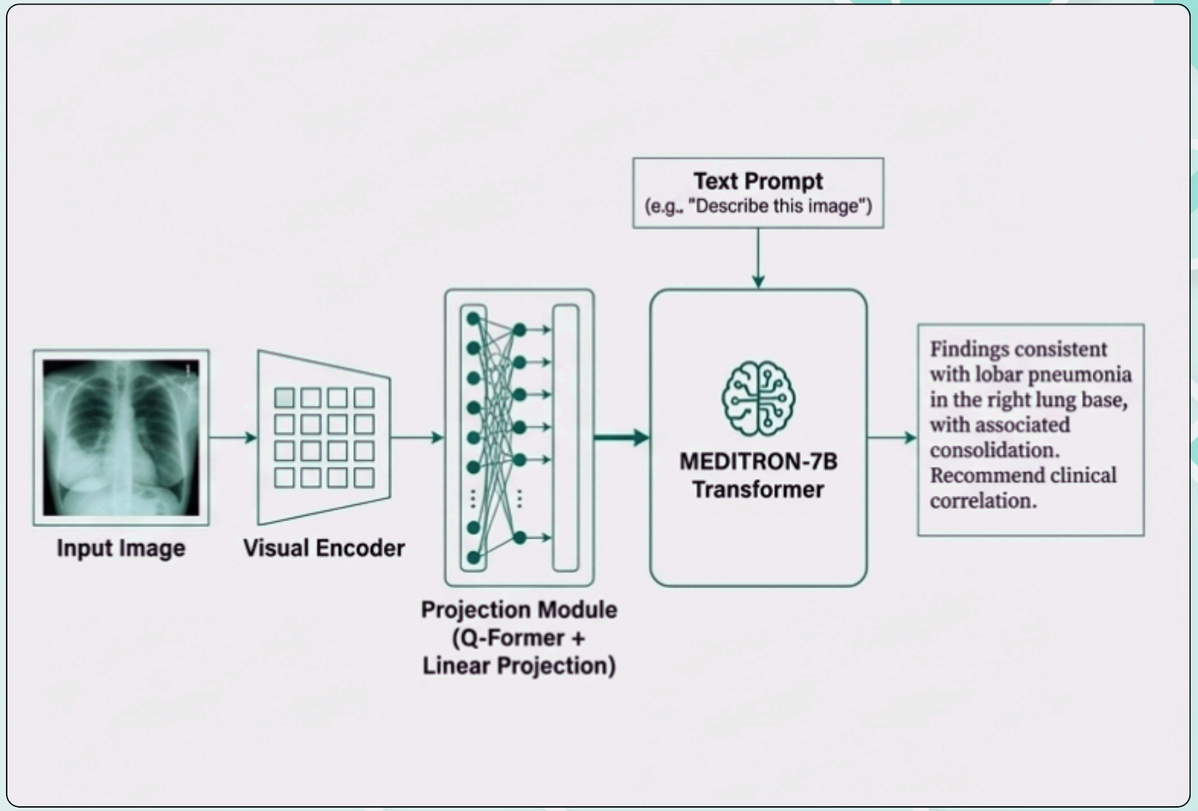
→ Use for clinical fact retrieval, NOT diagnosis
→ Mandate CoT prompting for all medical queries
→ Set temperature=0.6 for proper calibration
→ Verify outputs against clinical guidelines

Meditron V

A **Multimodal version of Meditron** that generates natural language responses from both image and text inputs

How it Works:

1. A visual encoder processes medical images (X-ray,CT,histology) into patch features.
2. A novel projection module transforms these visual features into embeddings the language model can understand
3. Meditron reasons over the combined text and image embeddings to generate a response

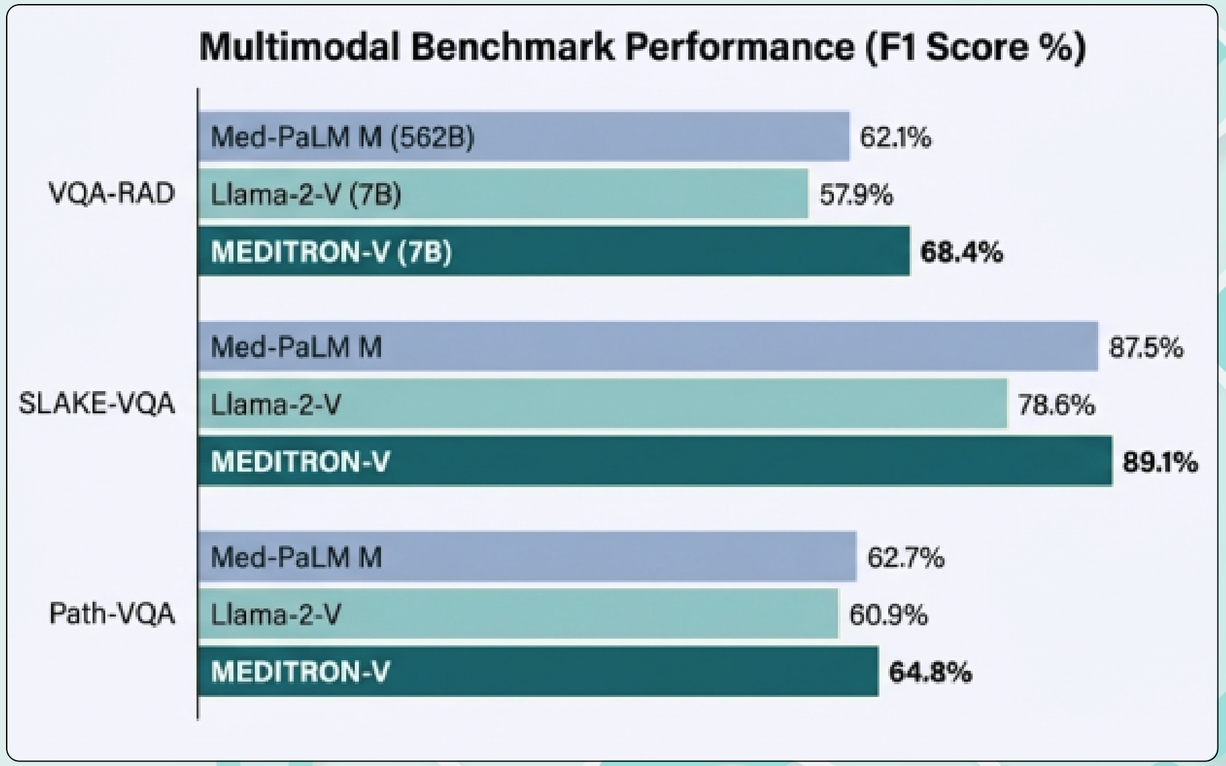


Mastering Benchmark Performance

Pretraining on medical text provides a better foundation for visual tasks.Meditron-V outperforms all reported medical multimodal systems,including much larger commercial models

Highlights

- **Better Foundation:** MEDITRON-V significantly outperforms its Llama-2-V baseline by an average of 8.5%,demonstrating the benefit of continued medical pretraining.
- **Outperforms SOTA Commercial Model** : Despite being ~80x smaller,MEDITRON-V achieves higher F1 scores than the 562B parameter Med-PaLM across all benchmarks(3.3% higher on average)



Meditron & Meditron-V:Key Applications

- **Clinical Decision Support (CDS)**

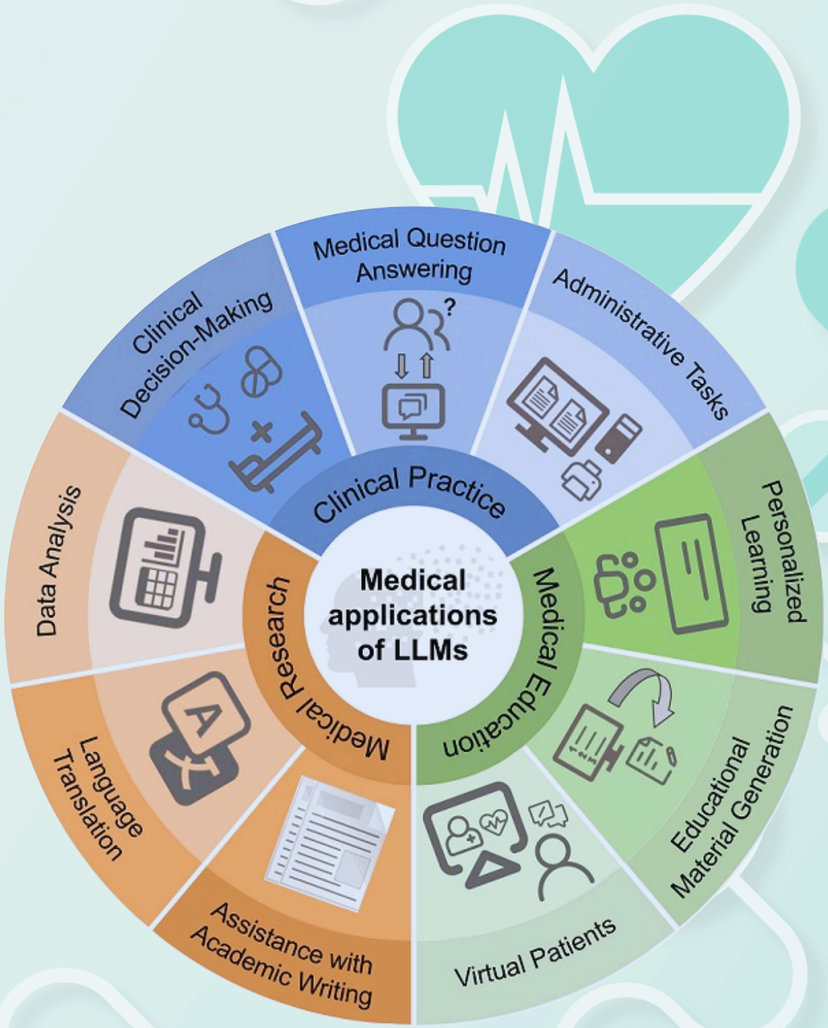
Assists in diagnosis, case interpretation, and evidence-based treatment planning using EBM and clinical guidelines.

- **Multimodal Medical Imaging (MEDITRON-V)**

Reasons over X-rays, CT, MRI, histology, enabling report generation, image captioning, and visual Q&A.

- **Medical Research & Education**

Synthesizes biomedical literature, supports clinical learning, and compresses large-scale medical knowledge.



→ **Vision**: MEDITRON as a "Digital Chief Resident" — democratizing access to expert medical reasoning.

CONCLUSION

By comparing MEDITRON(70b) to the expertise level expected from reputable clinical practice guidelines, the physicians conclude that MEDITRON shows proficiency that rivals, and in some aspects exceeds, that of medical residents with 1-5 years of experience.Thus this model helps enhance medical research, improving patient care, and fostering innovation across various health-related fields.

	Accuracy (↑)					
Model	MMLU-Medical	PubMedQA	MedMCQA	MedQA	MedQA-4-Option	Avg
Top Token Selection						
Mistral-7B*	55.8	17.8	40.2	32.4	41.1	37.5
Zephyr-7B-β*	63.3	46.0	43.0	42.8	48.5	48.7
PMC-Llama-7B	59.7	59.2	57.6	42.4	49.2	53.6
Llama-2-7B	56.3	61.8	54.4	44.0	49.6	53.2
MEDITRON-7B	55.6	74.4	59.2	47.9	52.0	57.5
Clinical-Camel-70B*	65.7	67.0	46.7	50.8	56.8	57.4
Med42-70B*	74.5	61.2	59.2	59.1	63.9	63.6
Llama-2-70B	74.7	78.0	62.7	59.2	61.3	67.2
MEDITRON-70B	73.6	80.0	65.1	60.7	65.4	69.0
Chain-of-thought						
Llama-2-70B	76.7	79.8	62.1	60.8	63.9	68.7
MEDITRON-70B	74.9	81.0	63.2	61.5	67.8	69.7
Self-consistency Chain-of-thought						
Llama-2-70B	**77.9**	80.0	62.6	61.5	63.8	69.2
MEDITRON-70B	77.6	**81.6**	**66.0**	**64.4**	**70.2**	**72.0**