Beyond Data: Multimodal AI in Tomorrow's Medicine

Kode

What is Kode?

Kode is a software development company specialized in transforming customer's needs into **tailor-made solutions and vertical tools** through the development of mathematical and statistical models based on data and rules.

Somebody call this **AI**.

Olga Cozzolino
Data Scientist
PM of Dialogo

Dialogo is a solution that, through a conversational interface and AI Agents, that work on operational tasks and support your company's decisions.

Let’s talk about...

Part 1

The Context & The Clinical Need

Part 2

Under the Hood - Architecture & Training

Part 3

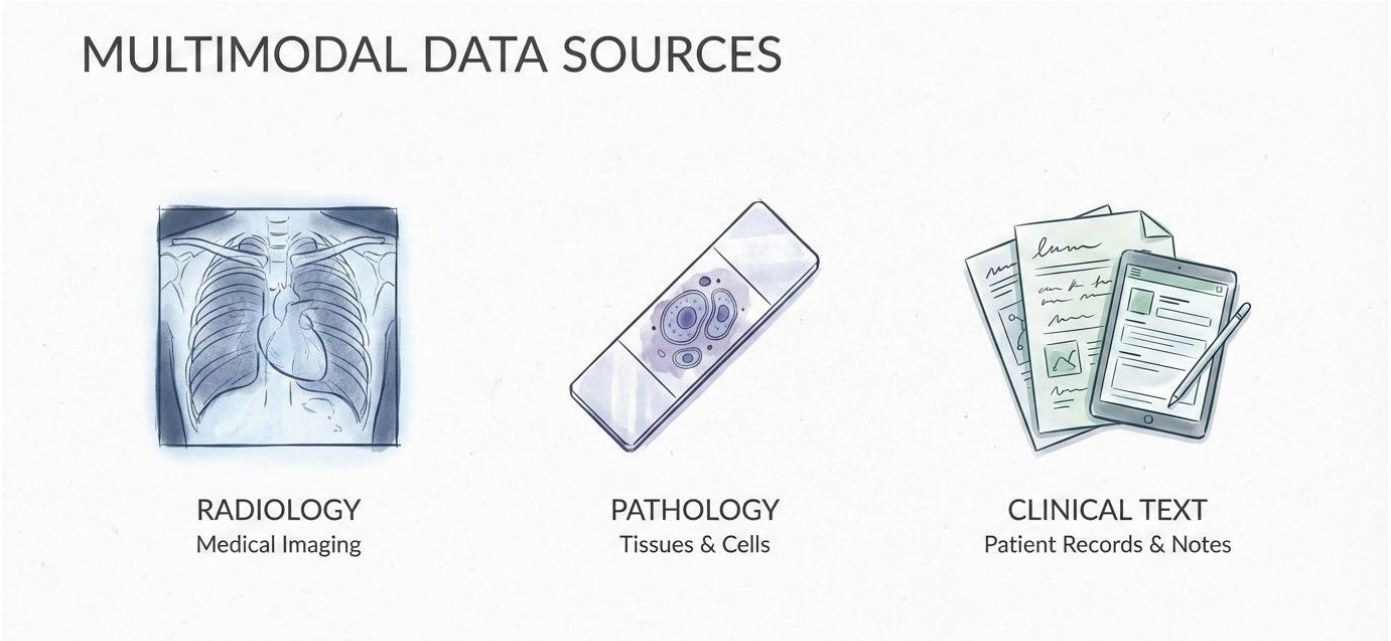
Clinical Performance & Evidence

Part 4

Safety, Ethics & The Future

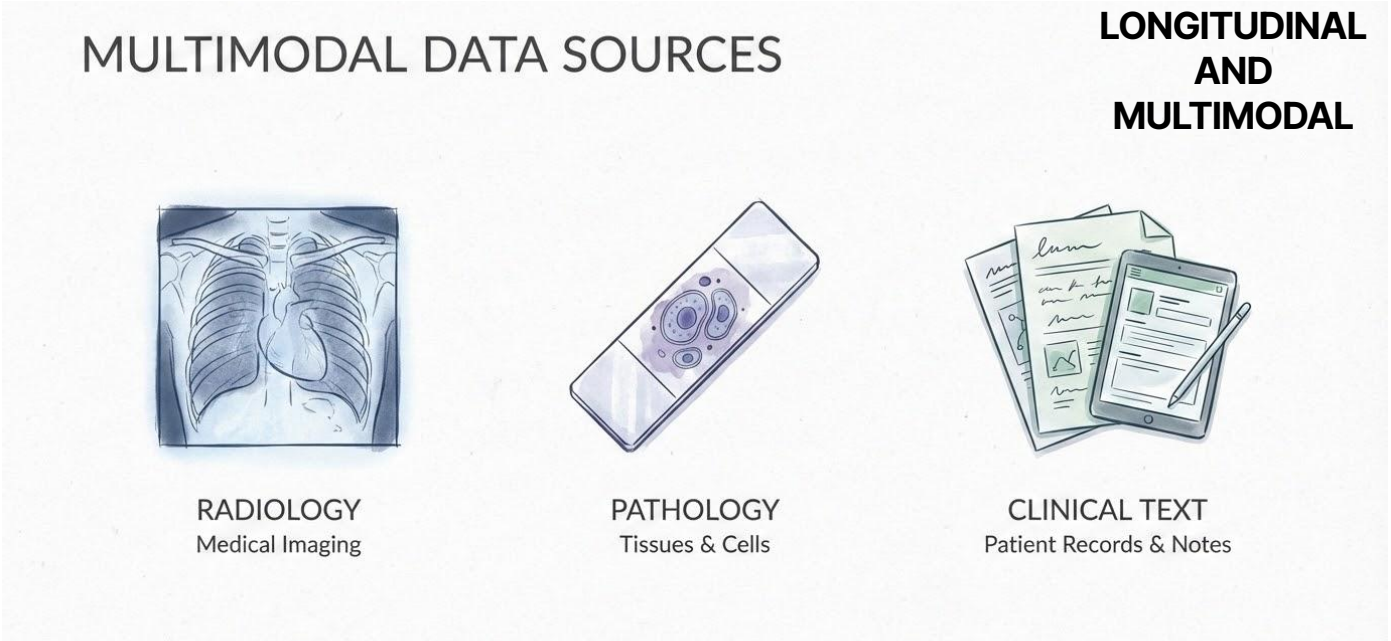
Kode

The Fragmentation Problem



The Fragmentation Problem

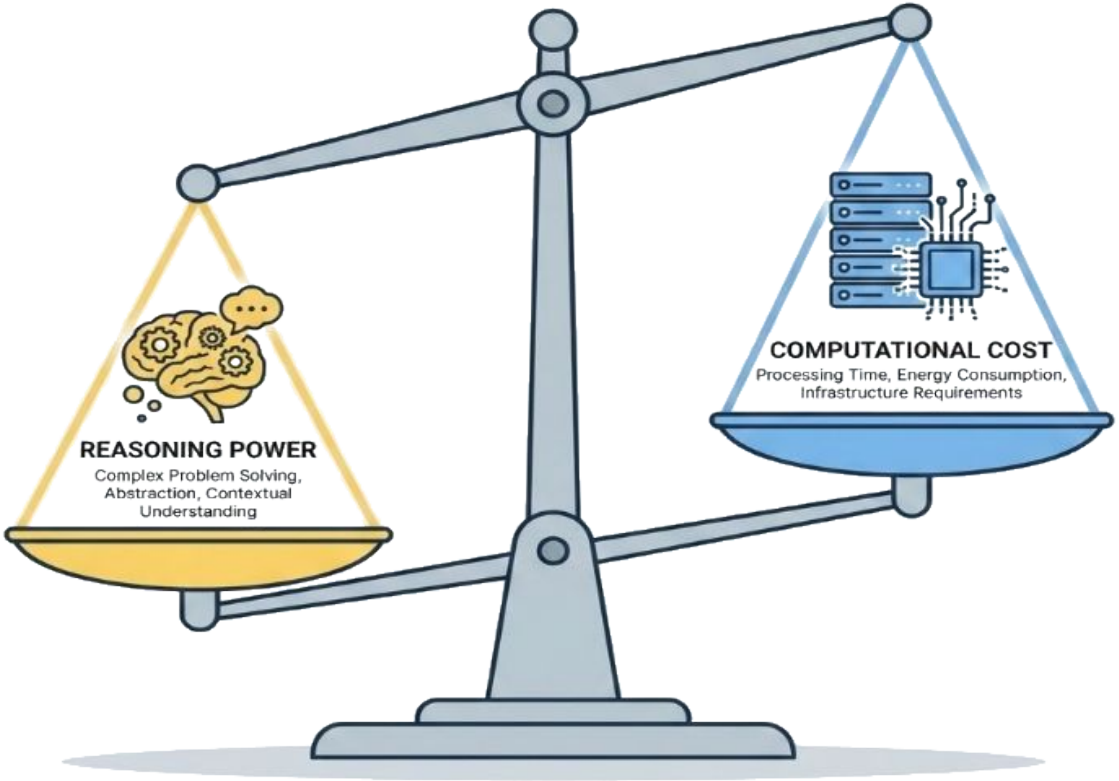
DECISION-MAKING IS LONGITUDINAL AND MULTIMODAL



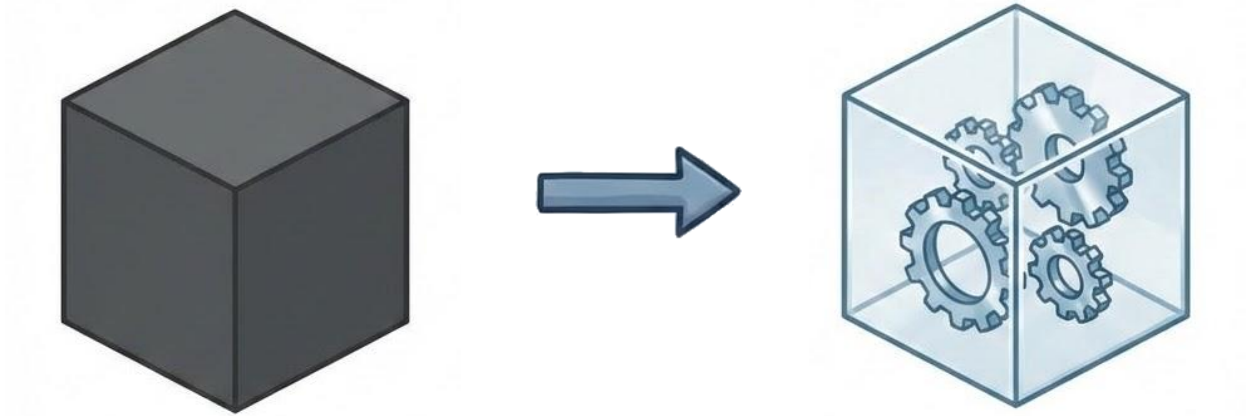
The Deployment Gap & PHI

Why 20 Billion Parameters?

GPT-1 (**2018**): ~117 Million
GPT-3 (**2020**): 175 Billion



Defining the "Black Box"



The AI as a Partner, Not a Tool

Technical Report

MEDGPT-OSS: TRAINING A GENERAL-PURPOSE VISION-LANGUAGE MODEL FOR BIOMEDICINE

Kai Zhang[1], Zhengqing Yuan[2], Cheng Peng[3], Songlin Zhao[1], Mengxian Lyu[3], Ziyi Chen[3], Yanfang Ye[2], Wei Liu[4], Ying Zhang[5], Kaleb E Smith[6], Lifang He[1], Lichao Sun[1,*], Yonghui Wu[3,*]

[1]Department of Computer Science and Engineering, Lehigh University
[2]Department of Computer Science and Engineering, University of Notre Dame
[3]Department of Health Outcomes & Biomedical Informatics, University of Florida
[4]Department of Radiation Oncology, Mayo Clinic
[5]Research Computing, University of Florida
[6]AI Technology Center, NVIDIA
lis221@lehigh.edu, yonghui.wu@ufl.edu

Is **MEDGPT-OSS** a more advanced stethoscope or a junior resident?

Let’s talk about...

Part 1

The Context & The Clinical Need

Part 2

Under the Hood - Architecture & Training

Part 3

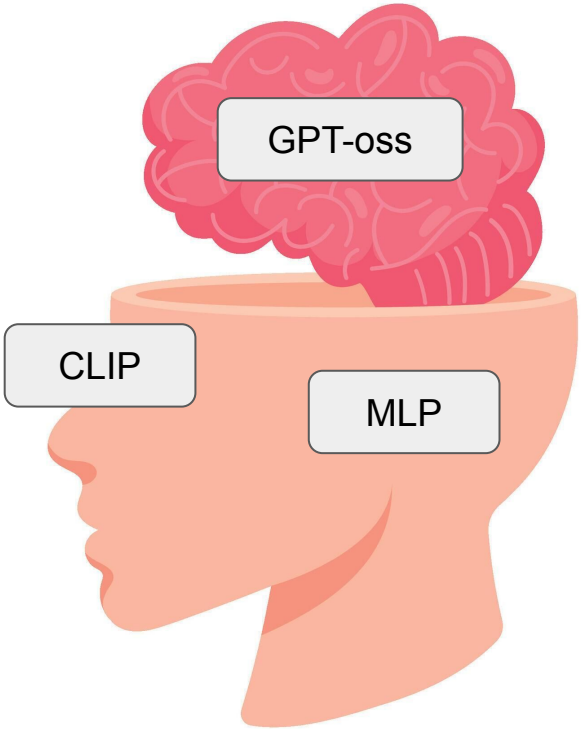
Clinical Performance & Evidence

Part 4

Safety, Ethics & The Future

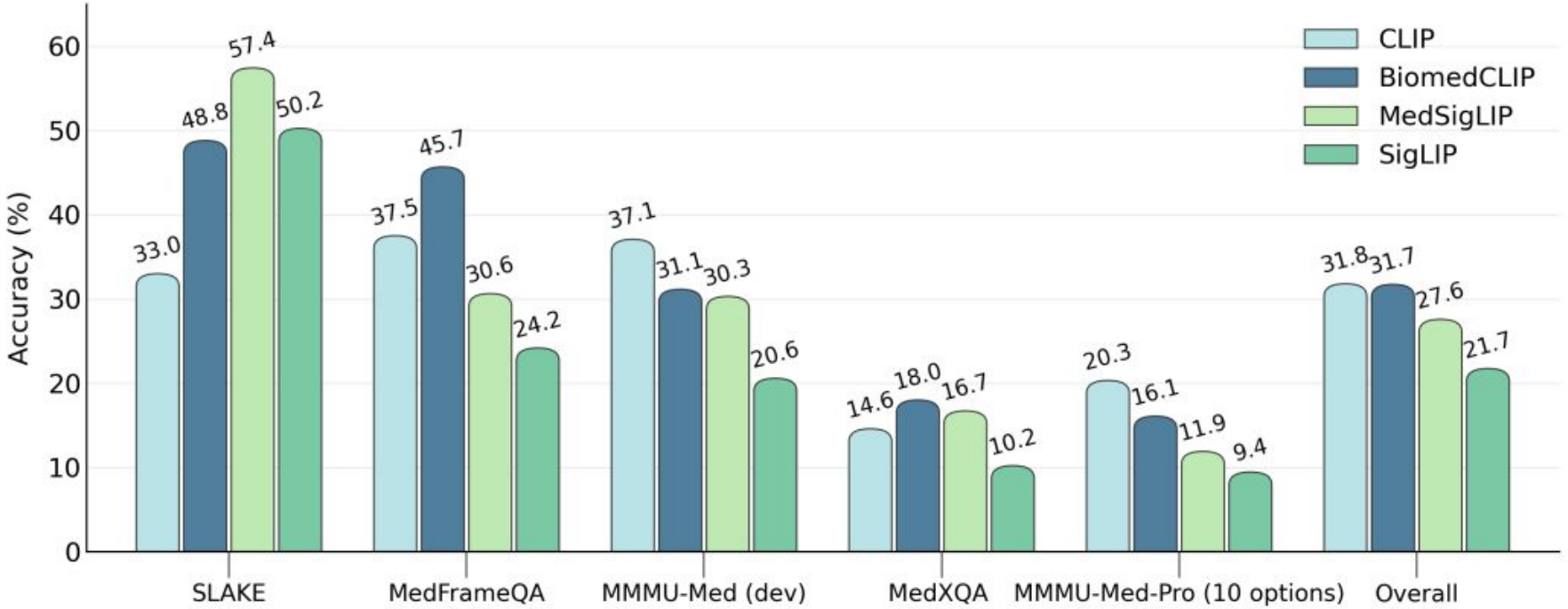
Kode

Modular Minimalism

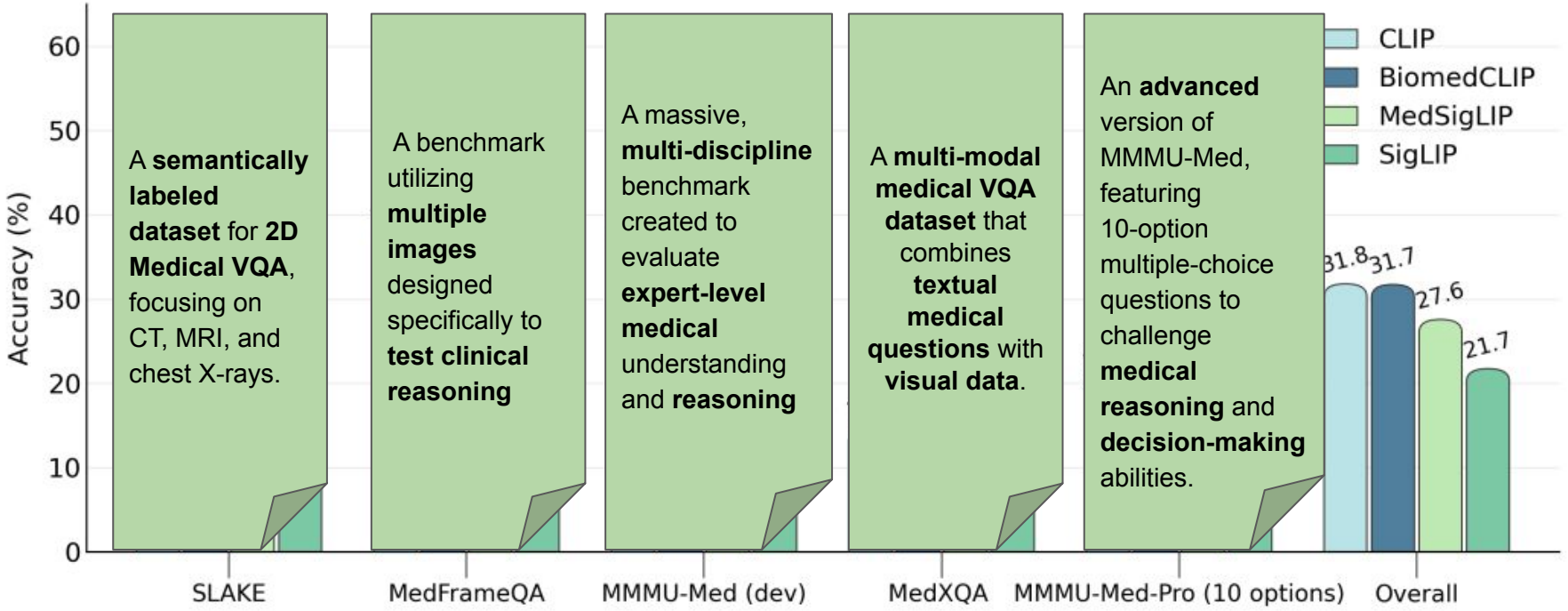


The "Eyes" of the Model: CLIP

"surprisingly, the vanilla CLIP backbone consistently outperformed the specialized medical encoders"



The "Eyes" of the Model: CLIP



The "Eyes" of the Model: CLIP

The authors claim:

1. Computational Efficiency

"This architecture was selected to strike an optimal balance between granular representational capacity and computational efficiency during inference"

2. Broader Visual Priors

The backbone maintains high domain coherence, ensuring that the model's logical transitions and instructions remain strictly within the boundaries of medical reality.

3. Superiority in Reasoning-Heavy Benchmark

MMMU-Med and MMMU-Med-Pro are specifically designed to evaluate "expert-level" reasoning

The "Eyes" of the Model: CLIP

The authors claim:

Model	Vision Encoder Size	Parameters (approx.)	Inference Cost
CLIP-ViT-L/14	ViT-Large (14x14)	~307M	High
BiomedCLIP	ViT-Base	~86M	Low (~0.387s)

1. Computational Efficiency

"This architecture was selected to strike an optimal balance between granular representational capacity and computational efficiency during inference"

2. Broader Visual Priors

The backbone maintains high domain coherence, ensuring that the model's logical transitions and instructions remain strictly within the boundaries of medical reality.

3. Superiority in Reasoning-Heavy Benchmark

MMMU-Med and MMMU-Med-Pro are specifically designed to evaluate "expert-level" reasoning

The "Eyes" of the Model: CLIP

The authors claim:

1. Computational Efficiency

"This architecture was selected to strike an optimal balance between granular representational capacity and computational efficiency during inference"

2. Broader Visual Priors

The backbone maintains high domain coherence, ensuring that the model's logical transitions and instructions remain strictly within the boundaries of medical reality.

3. Superiority in Reasoning-Heavy Benchmark

MMMU-Med and MMMU-Med-Pro are specifically designed to evaluate "expert-level" reasoning

The "Eyes" of the Model: CLIP

The authors **COULD** claim:

---

*"While **BiomedCLIP is lighter and performs well**, the choise ultimately fell on **CLIP** because it excels on more complex questions. Its **broader visual priors** also suggest it might handle **out-of-distribution (OOD) questions** better—though this, of course, still needs to be tested in the future."*

The "Brain": GPT-oss 20B

The language core is GPT-oss

1. Factual grounding

The model anchors its reasoning in verified biomedical knowledge, consistently outperforming other open-weight backbones in its ability to provide factually correct clinical responses

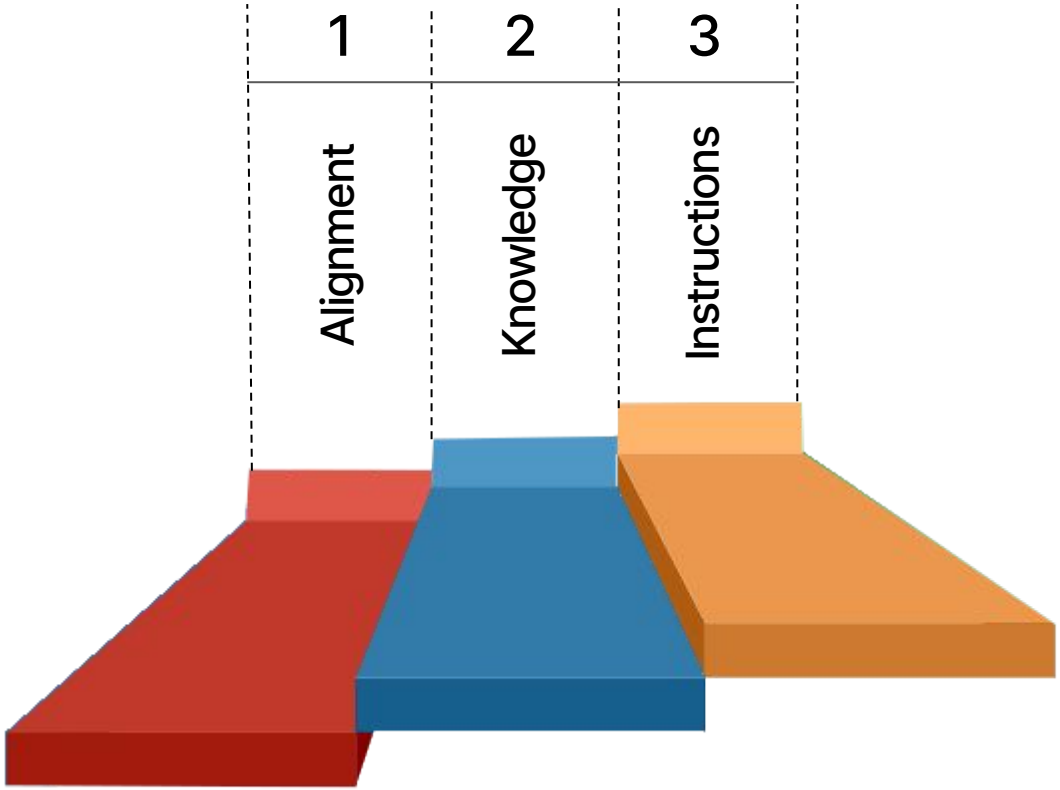
2. Domain Coherence

The backbone maintains high domain coherence, ensuring that the model’s logical transitions and instructions remain strictly within the boundaries of medical reality.

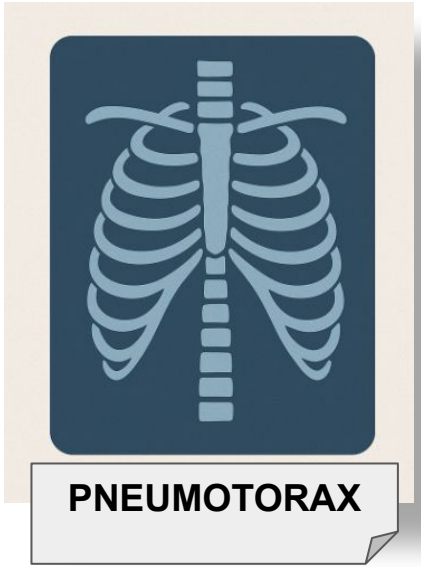
3. Moderate size

This moderate footprint ensures full compatibility with commodity GPUs, allowing healthcare institutions to deploy elite-level AI on-premises to maintain local control over sensitive patient data

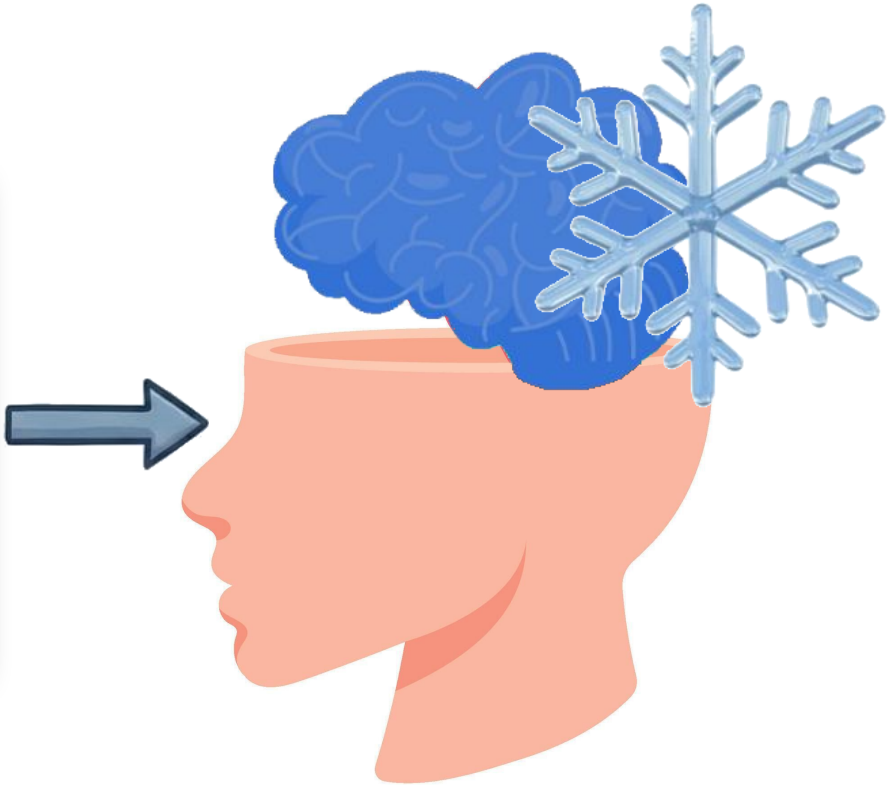
The Training Strategy:
A Three-Stage Curriculum



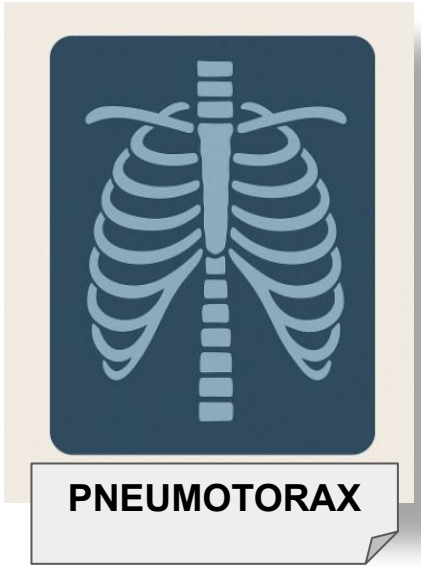
Stage 1
Short-Context Alignment



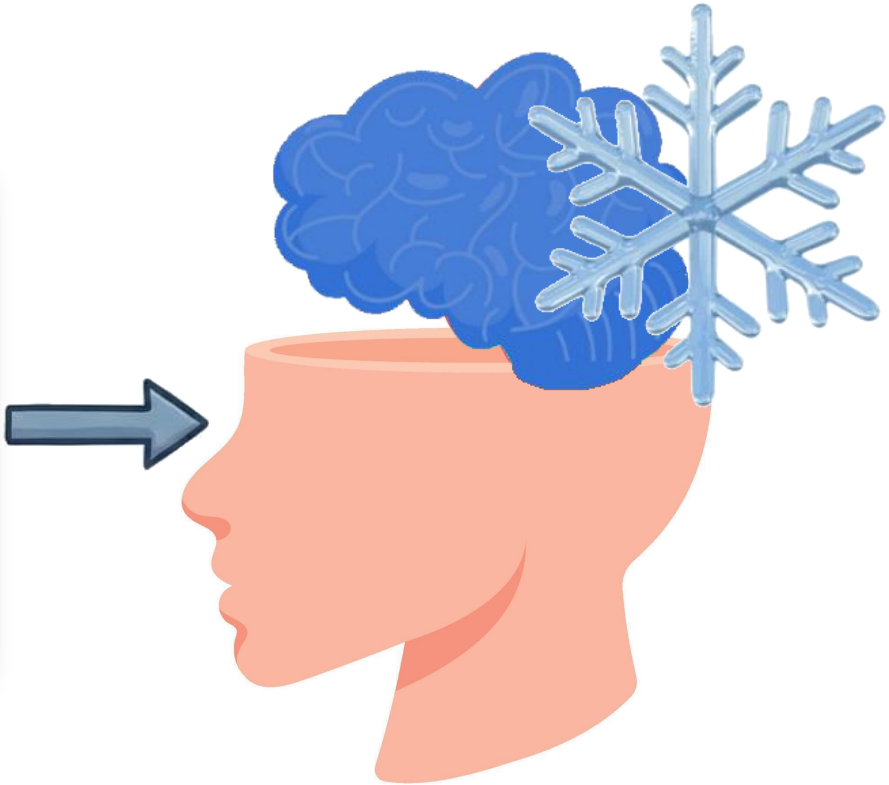
~ 1M



Stage 1
Short-Context Alignment

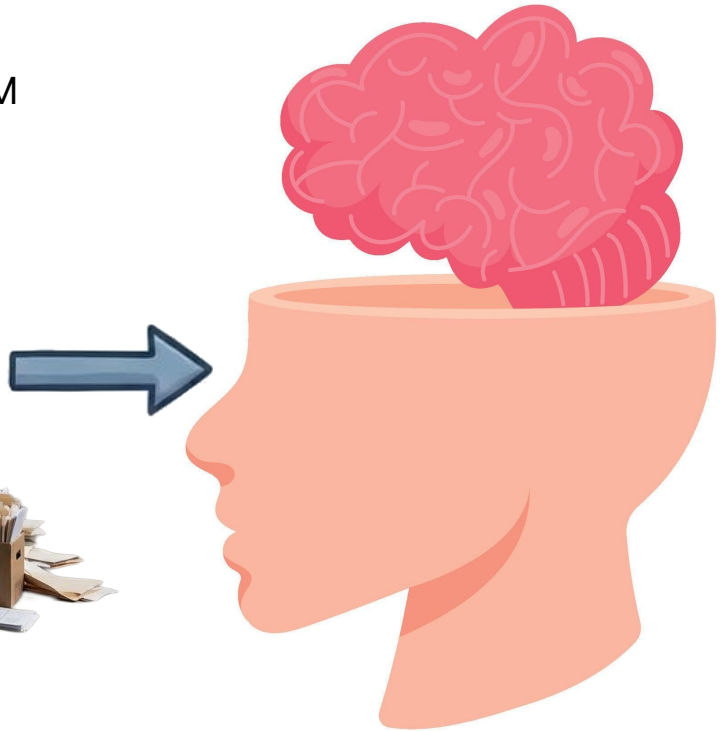


PubMed
PMC-OA



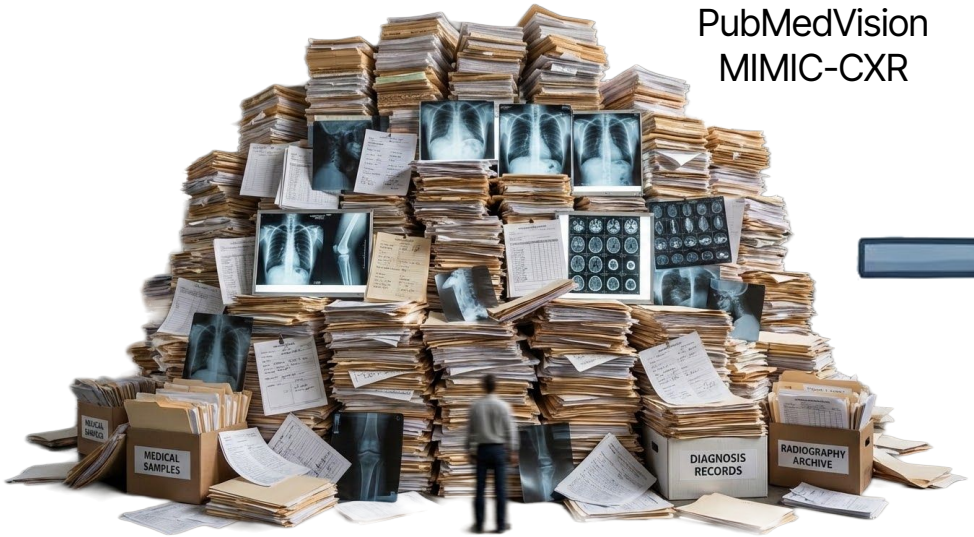
Stage 2
Mid-Training & Knowledge Integration

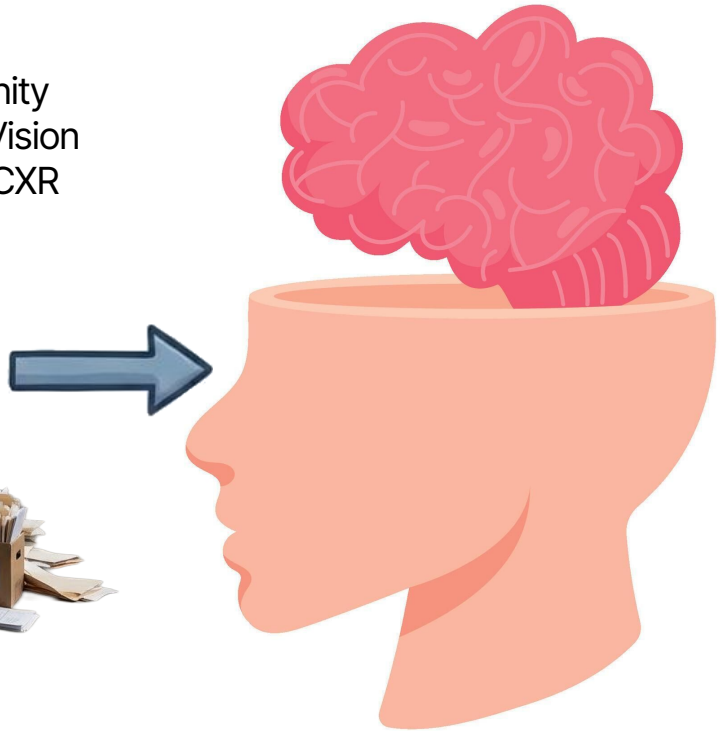
~ 19 M



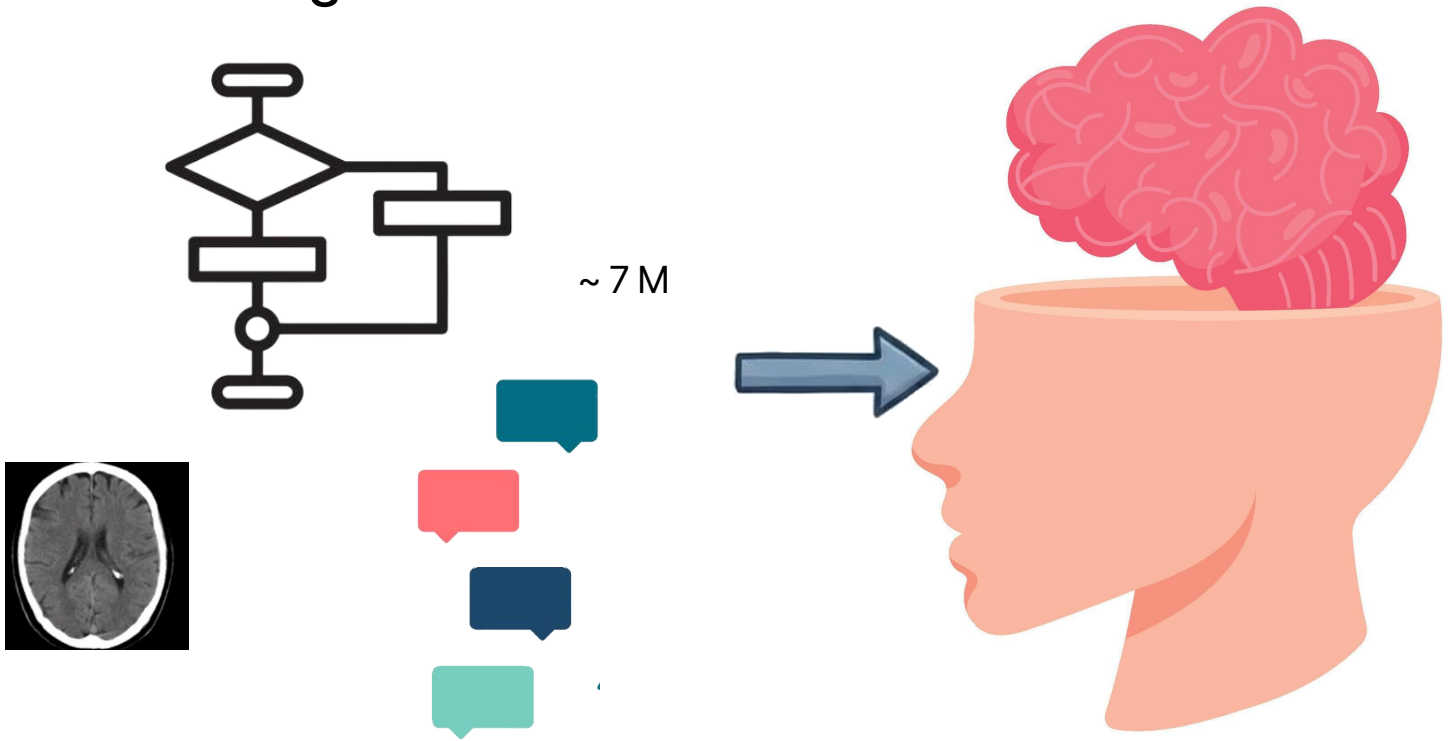
Stage 2
Mid-Training & Knowledge Integration

MedTrinity
PubMedVision
MIMIC-CXR



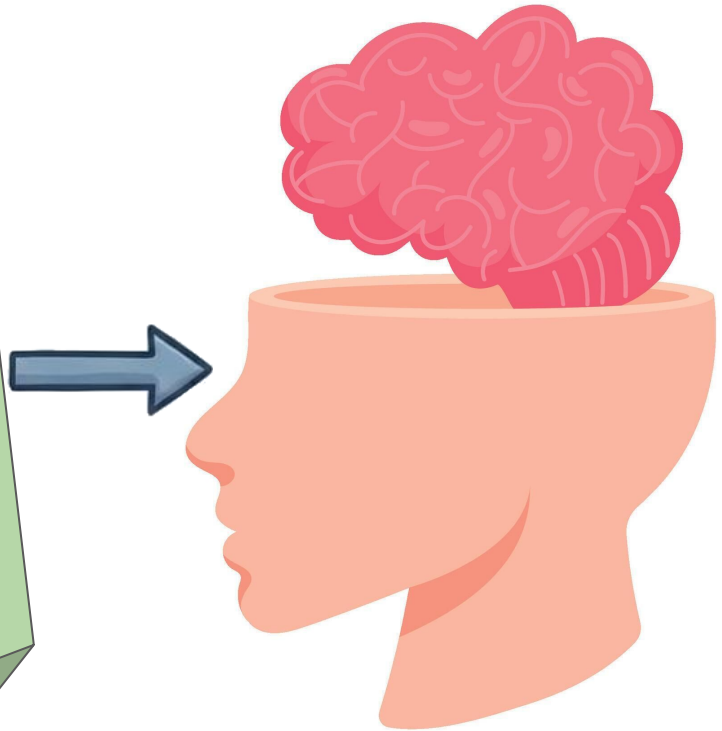


Stage 3
Instruction Tuning



Stage 3
Instruction Tuning

- **DeepSeek-R1** generates **100 reasoning traces** for a complex medical diagnosis question.
- **Reject sampling** selects the **20 most accurate and coherent traces** (e.g., those with correct differential diagnoses and logical clinical reasoning).
- **Refinement** optimizes these 20 traces (e.g., removing redundant steps or clarifying ambiguous reasoning).
- **Distillation:** A smaller model (e.g., **GPT-OSS 20B**) is then trained on these 20 high-quality traces to **replicate the same expert-level medical reasoning**



Technical Detail: Context & Hardware

Token

The basic unit of text length in NLP, typically corresponding to a word, subword, or character, depending on the tokenizer used.

1. Context length 32k

It supports a context length of over 32,000 tokens, which is vital for reading long patient histories

2. 8x B200 GPUs

It was trained on high-end B200 GPUs, taking about 330 hours for the full training

Let’s talk about...

Part 1

The Context & The Clinical Need

Part 2

Under the Hood - Architecture & Training

Part 3

Clinical Performance & Evidence

Part 4

Safety, Ethics & The Future

Kode

Benchmarking Multimodal Reasoning

Table 4: VQA results on multiple-choice benchmarks. Scores are reported in accuracy (%).

Dataset	MEDGPT-OSS	OctoMed	Hulu-Med	Lingshu†	MedGemma	QoQ-Med
MedXQA (multimodal)	**49.23**	25.60	34.35	31.43	30.90	29.64
SLAKE	71.53	65.07	69.14	**72.24**	55.98	46.53
MedFrameQA	**63.01**	42.82	62.82	61.01	47.63	55.73
MMMU-Med (dev)	**61.49**	47.65	57.71	59.43	47.43	51.84
MMMU-Med-Pro (4 opt)*	52.34	44.62	52.45	**52.67**	45.80	46.93
MMMU-Med-Pro (10 opt)	39.94	23.07	37.41	**43.45**	36.71	38.12

† Lingshu trained on the MMMU-Med dev set; *opt = options.

Punching Above Its Weight: MedXQA Results

Table 4: VQA results on multiple-choice benchmarks. Scores are reported in accuracy (%).

Dataset	MEDGPT-OSS	OctoMed	Hulu-Med	Lingshu†	MedGemma	QoQ-Med
MedXQA (multimodal)	**49.23**	25.60	34.35	31.43	30.90	29.64
SLAKE	71.53	65.07	69.14	**72.24**	55.98	46.53
MedFrameQA	**63.01**	42.82	62.82	61.01	47.63	55.73
MMMU-Med (dev)	**61.49**	47.65	57.71	59.43	47.43	51.84
MMMU-Med-Pro (4 opt)*	52.34	44.62	52.45	**52.67**	45.80	46.93
MMMU-Med-Pro (10 opt)	39.94	23.07	37.41	**43.45**	36.71	38.12

† Lingshu trained on the MMMU-Med dev set; *opt = options.

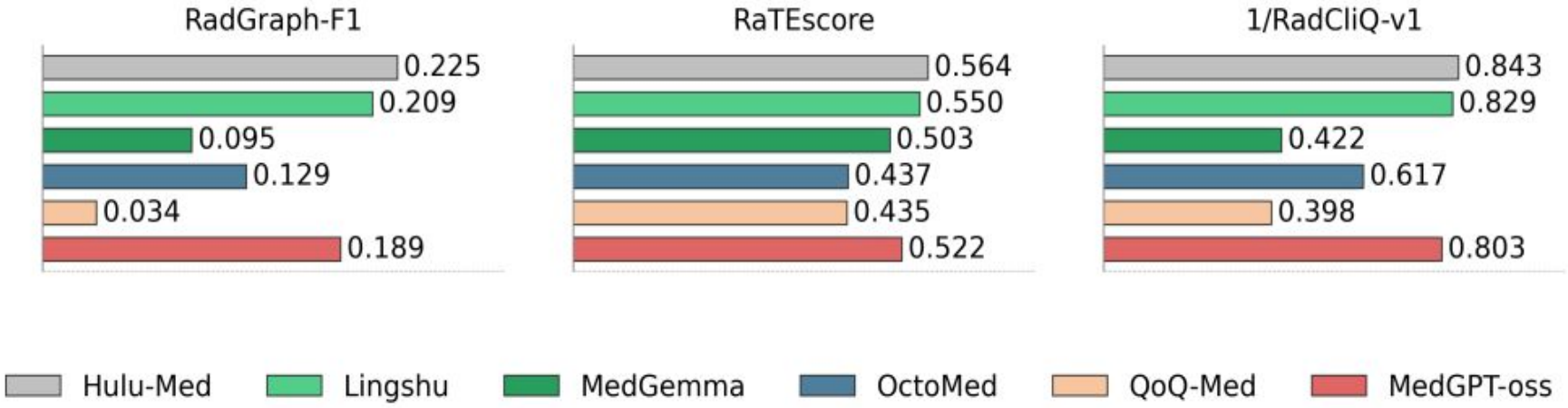
Lingshu 32B

Strong Performance in Clinical Text QA

Table 5: Text-only medical QA results. Scores are accuracy (%).

Dataset	MEDGPT-OSS	OctoMed	Hulu-Med	Lingshu	MedGemma	QoQ-Med
MedQA	70.11	44.97	**77.18**	70.71	66.75	55.25
PubMedQA	57.81	48.31	61.00	**62.44**	55.80	42.80
MedMCQA	62.53	55.32	**72.75**	65.27	65.48	51.42
MedXQA (text)	**25.38**	10.86	23.47	21.47	14.37	8.78
MMLU-Med	72.59	61.65	**87.10**	82.68	80.65	74.98
Medbullets	**68.71**	32.21	67.45	58.69	51.34	37.25

Generating Radiology Reports



The Superpower: In-Context Learning (ICL)

Table 6: Few-shot benchmarks. Scores are reported as accuracy percentages (%), with the exception of the Impression benchmark, which is evaluated using RaTEScore.

Dataset	MEDGPT-OSS	OctoMed	Hulu-Med	Lingshu	MedGemma	QoQ-Med
Patient-trial (0-shot)	48.81	40.96	51.01	**52.07**	31.03	45.20
Patient-trial (1-shot)	**55.60**	40.02	47.00	48.91	52.24	47.41
Impression (0-shot)	**47.22**	31.04	43.14	43.80	38.42	41.44
Impression (1-shot)	**47.25**	30.91	41.52	40.27		

ICL

The model's ability to instantly adapt to new clinical tasks using a single demonstration (1-shot) provided in the prompt, eliminating the need for technical retraining

Avoiding "Negative Transfer"

Table 6: Few-shot benchmarks. Scores are reported as accuracy percentages (%), with the exception of the Impression benchmark, which is evaluated using RaTEScore.

Dataset	MEDGPT-OSS	OctoMed	Hulu-Med	Lingshu	MedGemma	QoQ-Med
Patient-trial (0-shot)	48.81	40.96	51.01	**52.07**	31.03	45.20
Patient-trial (1-shot)	**55.60**	40.02	47.00	48.91	52.24	47.41
Impression (0-shot)	**47.22**	31.04	43.14	43.80	38.42	41.44
Impression (1-shot)	**47.25**	30.91	41.52	40.27	38.29	40.71

Negative Transfer

A phenomenon where model performance decreases when extra context is added.

Summary of Results

Medical Visual Question Answering

- ☑ MedXQA (multimodal): I Place
- ☐ SLAKE: II Place
- ☑ MedFrameQA: I Place
- ☑ MMMU-Med (dev): I Place
- ☐ MMMU-Med-Pro (4 opts): II Place
- ☐ MMMU-Med-Pro (10 optis): II Place

Medical Text-only Question-Answering

- ☐ MedQA: III Place
- ☐ PubMedQA: III Place
- ☐ MedMCQA: III Place
- ☑ MedXQA (text): I Place
- ☐ MMLU-Med: III Place
- ☑ Medbullets: I Place

Chest X-ray Report Generation

- ☐ MIMIC-CXR : III Place

- ☐ **In-context Learning Ability**

Patient-trial (0-shot): III Place

- ☑ Patient-trial (1-shot): III Place
- ☑ Impression (0-shot): I Place
- ☑ Impression (1-shot): I Place

Let’s talk about...

Part 1

The Context & The Clinical Need

Part 2

Under the Hood - Architecture & Training

Part 3

Clinical Performance & Evidence

Part 4

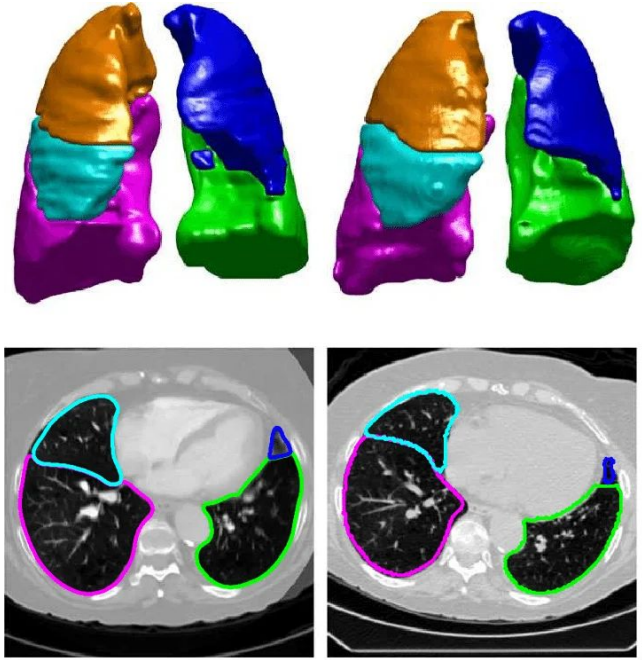
Safety, Ethics & The Future

Kode

The Persisting Challenge: Hallucinations

The 2D vs. 3D Frontier

Shi et al., 2024 DOI: 10.1002/mp.16781

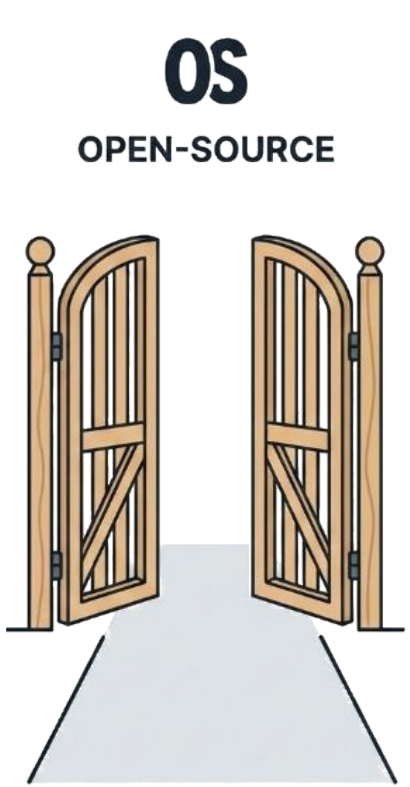


Lapointe et al., 2017 DOI:10.1002/mp.12475

Bias & Fairness in Training

Who Is Liable?

Transparency vs. Security



Transparency vs. Security

nature

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Subscribe | Sign up for alerts | RSS feed

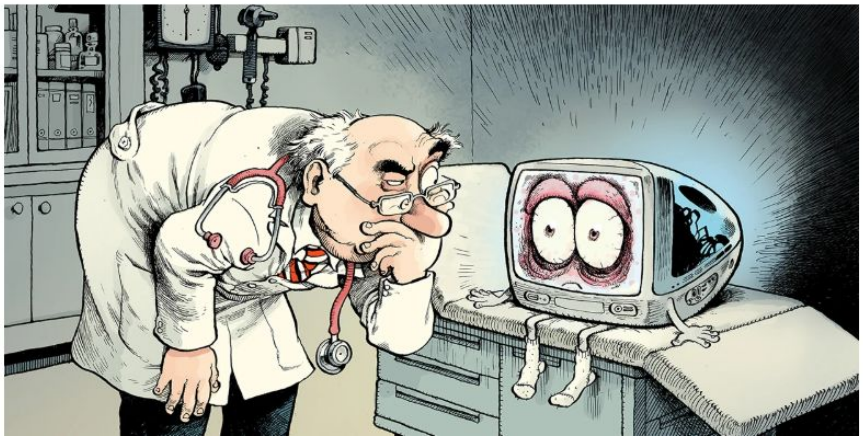
nature > news feature > article

NEWS FEATURE | 07 April 2026

Scientists invented a fake disease. AI told people it was real

Bixonimania doesn't exist except in a clutch of obviously bogus academic papers. So why did AI chatbots warn people about this fictional illness?

By Chris Stokel-Walker



Transparency vs. Security

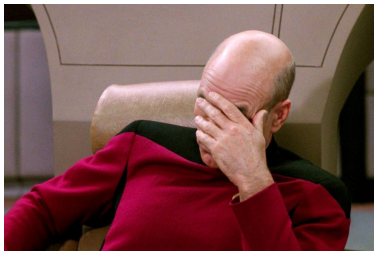
university called Asteria Horizon University in the equally fake Nova City, California. One paper's acknowledgements thank "Professor Maria Bohm at The Starfleet Academy for her kindness and generosity in contributing with her knowledge and her lab onboard the USS Enterprise". Both papers say they were funded by "the Professor Sideshow Bob Foundation for its work in advanced trickery. This works is a part of a larger funding initiative from the University of Fellowship of the Ring and the Galactic Triad".

Even if readers didn't make it all the way to the ends of the papers, they would have encountered red flags early on, such as statements that "this entire paper is made up" and "Fifty made-up individuals aged between 20 and 50 years were recruited for the exposure group".

AI-generated images of bixonimania, a fictitious illness. Source: Preprints.org https://doi.org/qzm4 (2024).

"AI Darth Vader can't hurt you."

AI Darth Vader



Transparency vs. Security

university called Asteria Horizon University in the equally fake Nova City, California. One paper's acknowledgements thank "Professor Maria Bohm at The Starfleet Academy for her kindness and generosity in contributing with her knowledge and her lab onboard the USS Enterprise". Both papers say they were funded by "the Professor Sideshow Bob Foundation for its work in advanced trickery. This works is a part of a larger funding initiative from the University of Fellowship of the Ring and the Galactic Triad".

Even if readers didn't make it all the way to the ends of the papers, they would have encountered red flags early on, such as statements that "this entire paper is made up" and "Fifty made-up individuals aged between 20 and 50 years were recruited for the exposure group".

AI-generated images of bixonimania, a fictitious illness. Source: Preprints.org https://doi.org/qzm4 (2024).

Preprints.org 10

Preprint | Brief Report

This version is not peer-reviewed.

Withdrawn: Bixonimania: Exploring the Influence of Blue Light on Periorbital Hyperpigmentation on the Palpebrae - an RCT with an r-BS design.

Lazljiv Izgubljenovic *, Betsy Thurberg, Andi Deep

Withdrawal Statement

This preprint has been withdrawn due to fabricated and non-authentic content that does not represent valid scientific research.

Feedback 4

Cureus
Part of SPRINGER NATURE

Article | Authors etc. | Metrics | Media | Comments

Original Article • Open Access Peer-Reviewed • More info

Retracted: Clinical and Dermoscopic Evaluation of Periorbital Melanosis and Its Psychological Impact and Effect on Quality of Life: A Descriptive Study

Saumya Banchhor • Sanjeev Gupta • Aneet Mahendra

9 | 10

This article has been retracted. See Banchhor S, Gupta S, Mahendra A (March 30, 2026) Retraction: Clinical and Dermoscopic Evaluation of Periorbital Melanosis and Its Psychological Impact and Effect on Quality of Life: A Descriptive Study. Cureus 18(3): r223. doi:10.7759/cureus.r223.

Transparency vs. Security

Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis

Mahmud Omar, Vera Sorin, Lothar H Wieler, Alexander W Charney, Patricia Kovatch, Carol R Horowitz, Panagiotis Korfiatis, Benjamin S Glicksberg, Robert Freeman, Girish N Nadkarni, Eyal Klang**

Summary

Background Large language models (LLMs) are increasingly used in health care but remain vulnerable to medical misinformation. We aimed to evaluate how often these models accept or reject fabricated medical content, and how framing that content as a logical fallacy changes results.

Methods In this cross-sectional benchmarking analysis, we probed 20 LLMs with more than 3·4 million prompts that all contained health misinformation drawn from three sources: public-forum and social-media dialogues, real hospital discharge notes in which we inserted a single false recommendation, and 300 physician-validated simulated vignettes. Logical fallacies—common patterns of flawed reasoning such as appeals to authority, popularity, or emotion—were used to test how rhetorical framing influences model behaviour. Each prompt was posed once in a neutral base form and ten times with a named logical fallacy. For every run we logged

Lancet Digit Health 2026; 8: 100949

Published Online February 9, 2026
https://doi.org/10.1016/j.landig.2025.100949

See Comment https://doi.org/10.1016/j.landig.2025.100978

*Contributed equally

The Windreich Department of Artificial Intelligence and

Transparency vs. Security

Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis

Mahmud Omar, Vera Sorin, Lothar H Wieler, Alexander W Charney, Patricia Kovatch, Carol R Horowitz, Panagiotis Korfiatis, Benjamin S Glicksberg, Robert Freeman, Girish N Nadkarni*, Eyal Klang*

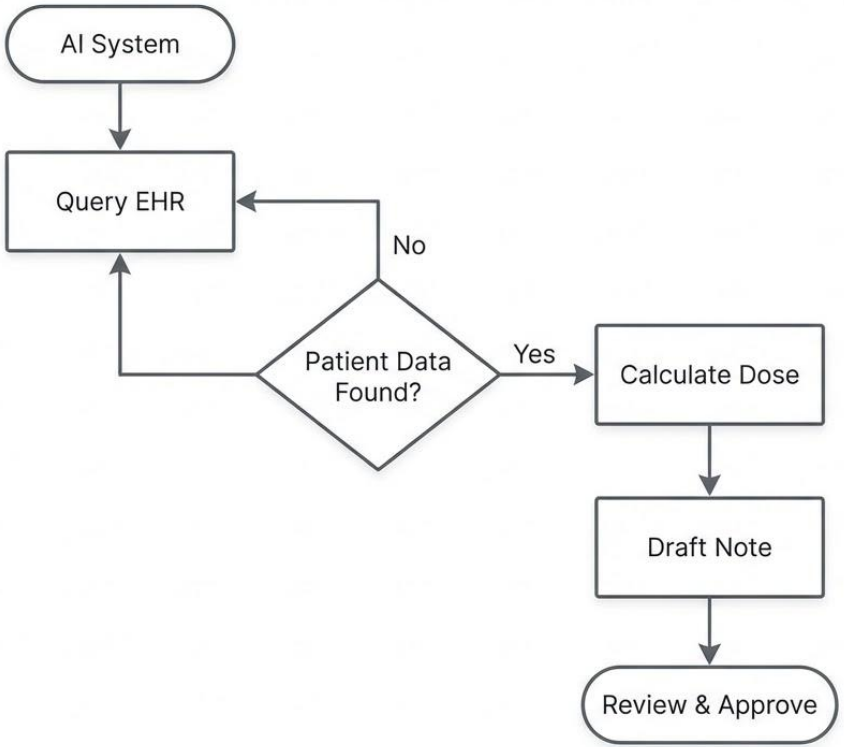
Summary

Background Large language models (LLMs) are increasingly used in health care bu misinformation. We aimed to evaluate how often these models accept or reject fabr framing that content as a logical fallacy changes results.

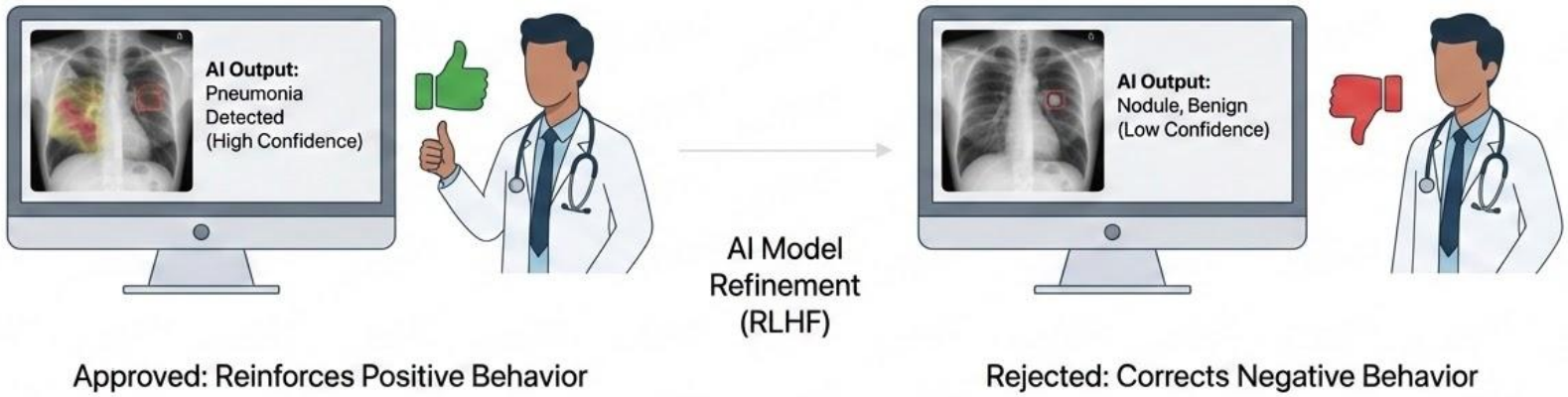
Methods In this cross-sectional benchmarking analysis, we probed 20 LLMs with mo all contained health misinformation drawn from three sources: public-forum a hospital discharge notes in which we inserted a single false recommendatio simulated vignettes. Logical fallacies—common patterns of flawed reasoning popularity, or emotion—were used to test how rhetorical framing influences mod posed once in a neutral base form and ten times with a named logical fall

in Gemma-3–4b-it to 0·0% in MediPhi. However, the apparent immunity of MediPhi reflected task refusal rather than true accuracy. Practically, the lowest interactive susceptibility was observed in gpt-oss-20b (a model comparable to OpenAI o3-mini, according to OpenAI), which maintained a 0·7% (598 of 8600) susceptibility rate across all prompts with a 74·1% (64 398 of 86 900) correct fallacy-detection rate (all models were compared with gpt-oss-20b for susceptibility). For every model, the appeal-to-popularity (bandwagon) prompt produced the greatest decrease in fallacy detection rate compared with

The Rise of "Agentic" AI



Reinforcement Learning with Clinical Rubrics



Thank you for the attention

Lungarno G. Galilei 1, Pisa 56125 Italy
P.IVA: 02040400505
info@kode-solutions.ne