
LECTURE 1-2

Reading, Presenting, and Critiquing Papers

How Thursday sessions work, with a model session on Moor et al. (2023)



Parisa Rashidi · Herbert Wertheim College of Engineering, University of Florida



Today

Period 7 · How Thursday sessions work

- 1 Reading a paper
- 2 Presenting a paper
- 3 Critiquing a paper
- 4 Discussion and reading responses

Period 8 · Model session

- 5 Moor et al., Nature 2023:
presentation, critique, discussion

Anatomy of a Thursday

Two papers, one per period. Each paper gets 50 minutes.

Period 7 · paper 1

Presentation group · 20 min

Critique group · 15 min

Discussion · 15 min

Period 8 · paper 2

Presentation group · 20 min

Critique group · 15 min

Discussion · 15 min

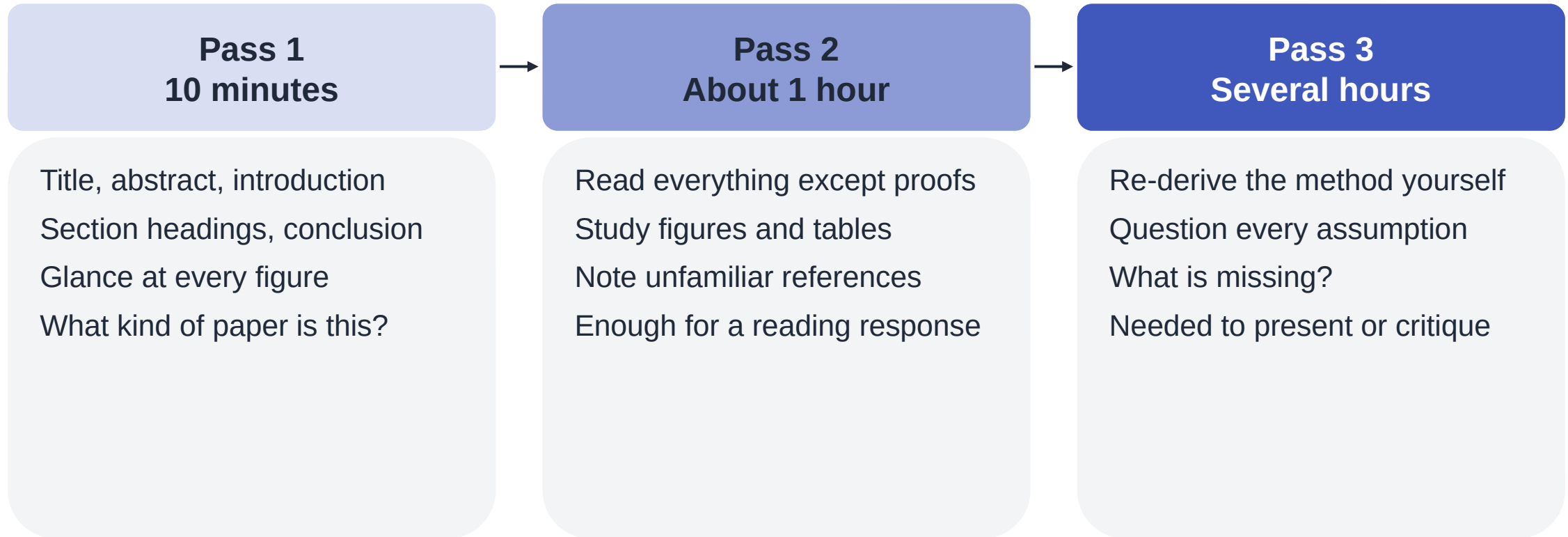
Groups are graded independently · everyone else submits a reading response



Reading a Paper

Read in Three Passes

Do not read a paper once from start to finish. Read it three times, each time deeper.



S. Keshav, "How to Read a Paper," ACM SIGCOMM Computer Communication Review, 2007

Find the Core Claim

Before judging a paper, state its claim in one sentence.

The authors show that [method] improves [outcome] on [data], measured by [metric], compared with [baseline].

Method

What is new?

Outcome and data

On which task, which patients, which sites?

Metric

Is it the right one for the clinical use?

Baseline

Is the comparison fair and strong?

Questions to Ask of a Biomedical AI Paper

Data

Where from, how many patients, which sites?
How were labels made?
Split by patient? Any leakage?

Model

What is actually new?
Strong, fairly tuned baselines?
Ablations for each design choice?

Evaluation

Metric fits the clinical use?
Confidence intervals?
External test set?
Subgroups? Calibration?

Claims

Shown, or only suggested?
Clinical utility, or only accuracy?
Limitations acknowledged?

Reporting guidelines: TRIPOD+AI (prediction models) · CONSORT-AI and SPIRIT-AI (trials) · CLAIM (imaging)



Presenting a Paper

The Presentation Rubric

Five criteria, 20 points each, scored in bands. Suggested time for a 20-minute talk:

4 min	7 min	6 min	3 min	
Overview and motivation	Problem, why it matters, main contribution			20 pts
Methods	Data, model, key design choices			20 pts
Results and evidence	Main results, metrics, baselines, interpreted			20 pts
Strengths	Concrete, relative to prior work			20 pts
Presentation quality	Structure, slides, delivery, timing (throughout)			20 pts

A Suggested Slide Outline

About 12 slides for 20 minutes. Every member presents.

1	Title and authors	7	Key design choices
2	The problem	8	Main result
3	Why it matters	9	Second result or ablation
4	The contribution	10	Strengths
5	Data	11	Take-home message
6	Model and method	12	Hand over to critique

Presenting: Do and Don't

Do

- One message per slide
- Redraw key figures in your own words
- Define notation before using it
- Interpret each result: what does it show?
- Share speaking time equally

Don't

- Paste paragraphs from the paper
- Read your slides aloud
- Show every table and figure
- Skip the data and the baselines
- Run over time



Critiquing a Paper

The Critique Rubric

Five criteria, 20 points each, scored in bands. Suggested time for a 15-minute critique:

2 min	4 min	4 min	3 min	2 min
Summary of core claims	Accurate restatement before critiquing			20 pts
Methodological critique	Data, design, validation, missing controls			20 pts
Results and interpretation	Do results support the claims?			20 pts
Limitations, biases, ethics	Beyond what the authors state			20 pts
Constructive suggestions	Concrete, realistic improvements			20 pts

Shown Versus Claimed

The most useful critique question: does the evidence support the claim as stated?

Claimed (abstract)

"Our model generalizes across hospitals and is ready for clinical use."

Shown (results)

Trained at one hospital
Tested at one other hospital:
AUROC fell from 0.91 to 0.83
No calibration, no prospective test

Gap: one external site and a retrospective test do not show readiness for clinical use.

Make Every Point Specific

Vague criticism earns few points. Point to the evidence in the paper.

Weak

Strong

"The dataset is small."



"The test set has 212 patients and 19 positives, so the AUROC confidence interval is wide, and none is reported."

"They should test more."



"All data come from one health system; an external site or a temporal split would test generalization."

"The paper is well written."



"The ablation in Table 3 isolates the effect of pretraining, which earlier work did not do."

A Critique Toolbox

Questions that work on almost any biomedical AI paper.

Data

Leakage between splits? Label quality? Who is missing from the cohort?

Design

Were baselines tuned as carefully? Was the test set used for model selection?

Evaluation

Right metric for the use? Calibration? Subgroups? External or prospective validation?

Claims

Causal language from correlational results? Utility claimed from accuracy alone?

Ethics

Consent and privacy? Bias against groups? Who benefits, who bears the risk?



Discussion and Reading Responses

The Discussion

Both groups lead; the questions come from the reading responses.

Who leads

Presentation and critique groups together

Where questions come from

Your reading responses

Good contributions

Connect to other papers, clinical practice, or your project

Counts toward

Participation (20% part of the grade)

Reading Responses

For each paper you are not presenting or critiquing.

One paragraph (about 150 words) on the claim and the evidence, plus one question · due 1:00 PM Thursday

Example (Moor et al., 2023): The authors argue that foundation models will enable generalist medical AI that handles new tasks from plain-language instructions across many data types. As a perspective, the paper offers early general-domain results and use cases rather than clinical evidence, so the claim of feasibility is plausible but the claim of clinical value is not yet supported.

Question: Who should be responsible for validating a model that can be asked to do any task?

Reading Responses: How They Are Graded

1 point each, graded automatically by Canvas when you submit on time.

1

Submitted on Canvas before 1:00 PM Thursday

0

Missing or late (the survey locks at 1:00 PM); lowest two dropped

A good response:

States the main claim in your own words

Judges the evidence, pointing to a specific result, figure, or design choice

Asks an open question, not a factual or yes/no one

How Sessions Are Graded

Item	Graded by	When	How
Presentation	Instructor	Live; final within a week	Banded rubric, 100 pts
Critique	Instructor	Live; final within a week	Banded rubric, 100 pts
Reading responses	Canvas (automatic)	On submission	1 point if on time
Audience feedback	Classmates	During discussion	Ungraded, shared anonymously

What Happens Next

Fri, Jan 15

Paper and group assignments posted on Canvas

Tue, Jan 19

Transformer Review, Part 1

Thu, Jan 21

First student sessions: BERT (period 7) and GatorTron (period 8)

Before Jan 21

Reading responses for both papers



Model Session: Moor et al., 2023

The Paper

Foundation models for generalist medical artificial intelligence

Moor, Banerjee, Shakeri Hossein Abad, Krumholz, Leskovec, Topol, Rajpurkar

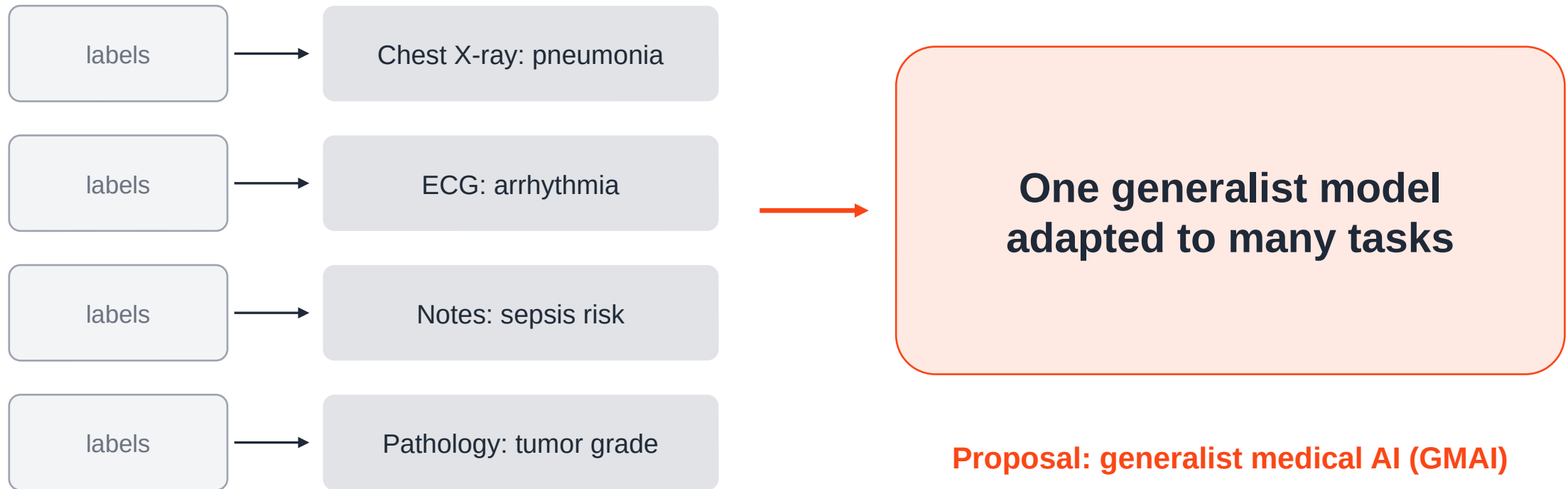
Nature 616, 259-265 (April 2023) · doi:10.1038/s41586-023-05881-4

Type: perspective
No new experiments or results tables

Rubric adapted:
methods = proposed framework
results = evidence and arguments

Motivation: One Model per Task

Most medical AI models do one task, on one data type, and need new labeled data for each new task.



Today: one model per task

PRESENTATION

The Proposal: Three Capabilities

Dynamic task specification

New tasks are described in plain language; no retraining

"Explain this finding to the patient in simple terms."

Flexible multimodal inputs and outputs

Any combination of images, records, labs, genomics, text in; text, speech, or annotations out

CT + notes + labs in; annotated image + report out

Medical domain knowledge

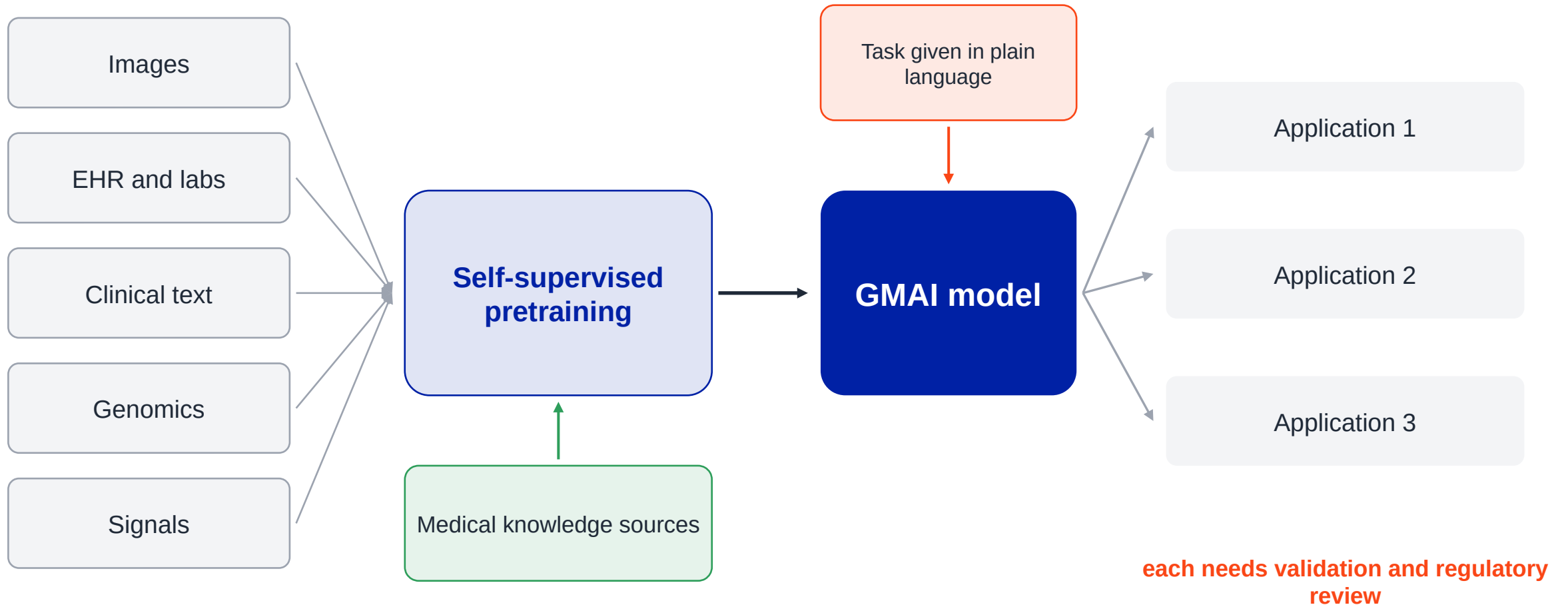
Represents medical knowledge to reason about new tasks and explain outputs

Explains why a finding suggests a diagnosis

Moor et al., 2023

PRESENTATION

How GMAI Would Be Built



PRESENTATION

Six Proposed Applications

Grounded radiology reports

Findings linked to the image regions they describe

Augmented procedures

Real-time help during surgery or endoscopy

Bedside decision support

Combines records, audio, and text to flag risk

Interactive note-taking

Drafts notes from the clinical encounter

Chatbots for patients

Answers patient questions with personal context

Text-to-protein generation

Designs proteins from text descriptions

PRESENTATION

Challenges the Authors Name

Validation	Verification	Social biases	Privacy	Scale
Failure modes are hard to anticipate	Complex outputs are hard to check	May grow with model scale	Memorization, prompt attacks	Data, training, and hardware cost

Strengths

Clear framing

A precise definition of GMAI through three capabilities

Concrete use cases

Six applications spanning radiology, surgery, notes, patients, and proteins

Honest about risks

Names validation, verification, bias, privacy, and scale

Policy relevance

Flags that current device regulation does not fit generalist models

Critique: The Core Claims

First, restate what the paper claims, accurately.

1

Feasibility

Foundation-model techniques make generalist medical AI possible

2

Value

GMAI would deliver substantial clinical value

3

Regulation

GMAI will strain how medical AI is validated and regulated

CRITIQUE

Critique: Shown Versus Claimed

Claim	Evidence offered	Verdict
Feasibility	Early general-domain foundation models	Plausible
Clinical value	Hypothetical scenarios, no outcome data	Not yet shown
Medical reasoning	Described, but no way to verify it proposed	Underspecified

A fair critique of a perspective judges its arguments, not missing experiments it never set out to run.

CRITIQUE

Critique: Limitations Beyond the Paper

Concentration

Only a few companies and health systems have the data and compute

Evaluation

No agreed way to evaluate open-ended, multimodal outputs

Liability

Who is responsible when a generalist model errs?

Automation bias

Fluent outputs invite over-trust by clinicians

Equity

Will it work in low-resource settings and underrepresented groups?

CRITIQUE

Critique: Constructive Suggestions

1

Define an evaluation protocol for each proposed application, not for the model as a whole

2

Stage deployment: low-risk documentation tasks before diagnosis and treatment

3

Benchmarks that check grounding: does each statement point to supporting evidence?

4

Prospective studies with clinical outcomes, not only accuracy

5

Open models so independent groups can audit them

CRITIQUE

Then and Now: 2023 to 2026

Which parts of the GMAI vision exist today?

Plain-language tasks



Routine: all frontier models are prompted

Multimodal inputs



Routine: natively multimodal models, MedGemma 1.5

Grounded outputs



Partial: e.g., boxes on chest X-rays

Medical reasoning



Partial: reasoning models, hard to verify

Validation and regulation



Largely open

DISCUSSION

Discussion

Is one generalist model safer or riskier than many narrow models, each validated separately?

Who should validate a model that can be asked to do any task: developers, hospitals, or regulators?

Which GMAI application would you deploy first at UF Health, and how would you evaluate it?

DISCUSSION