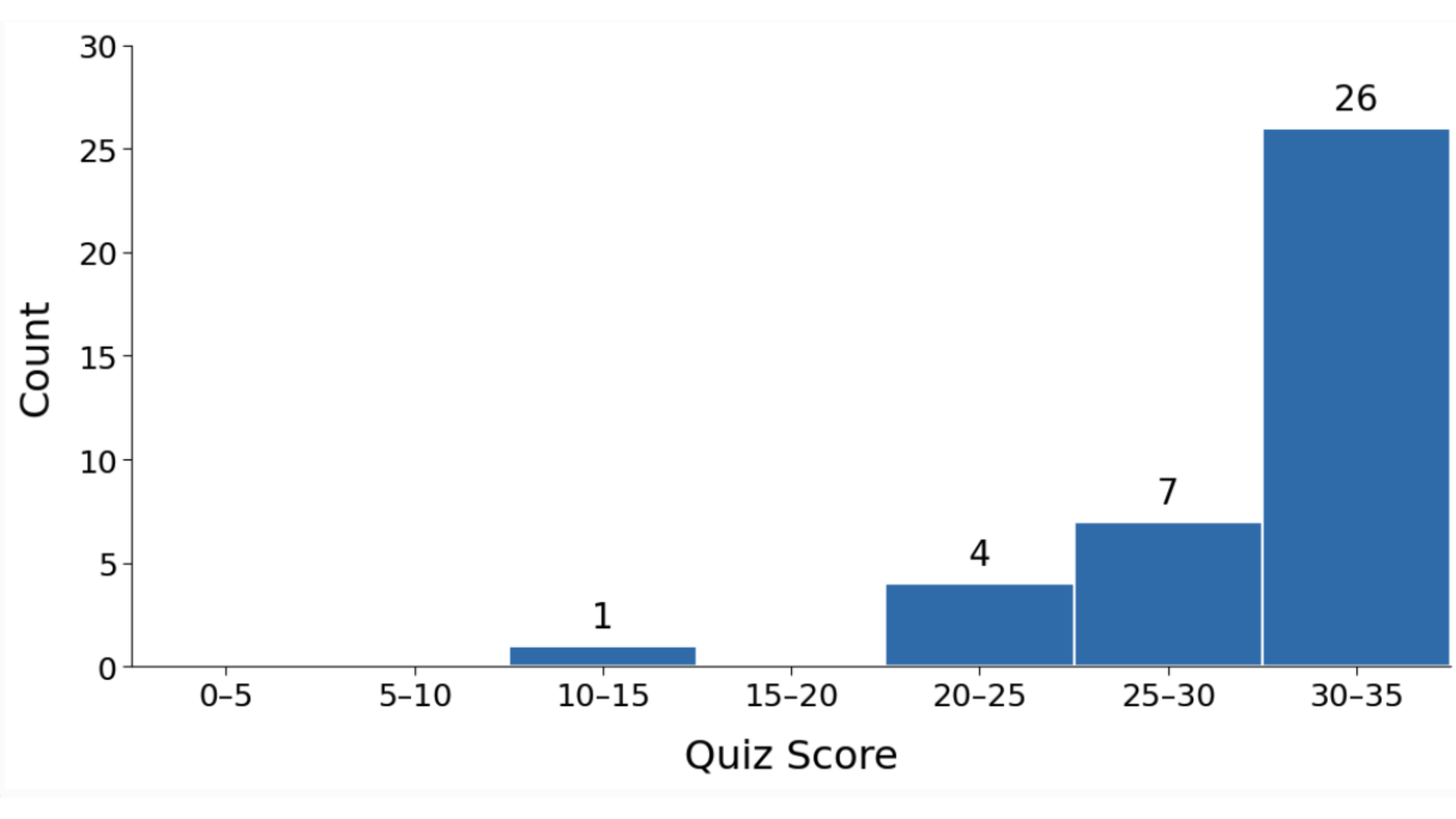


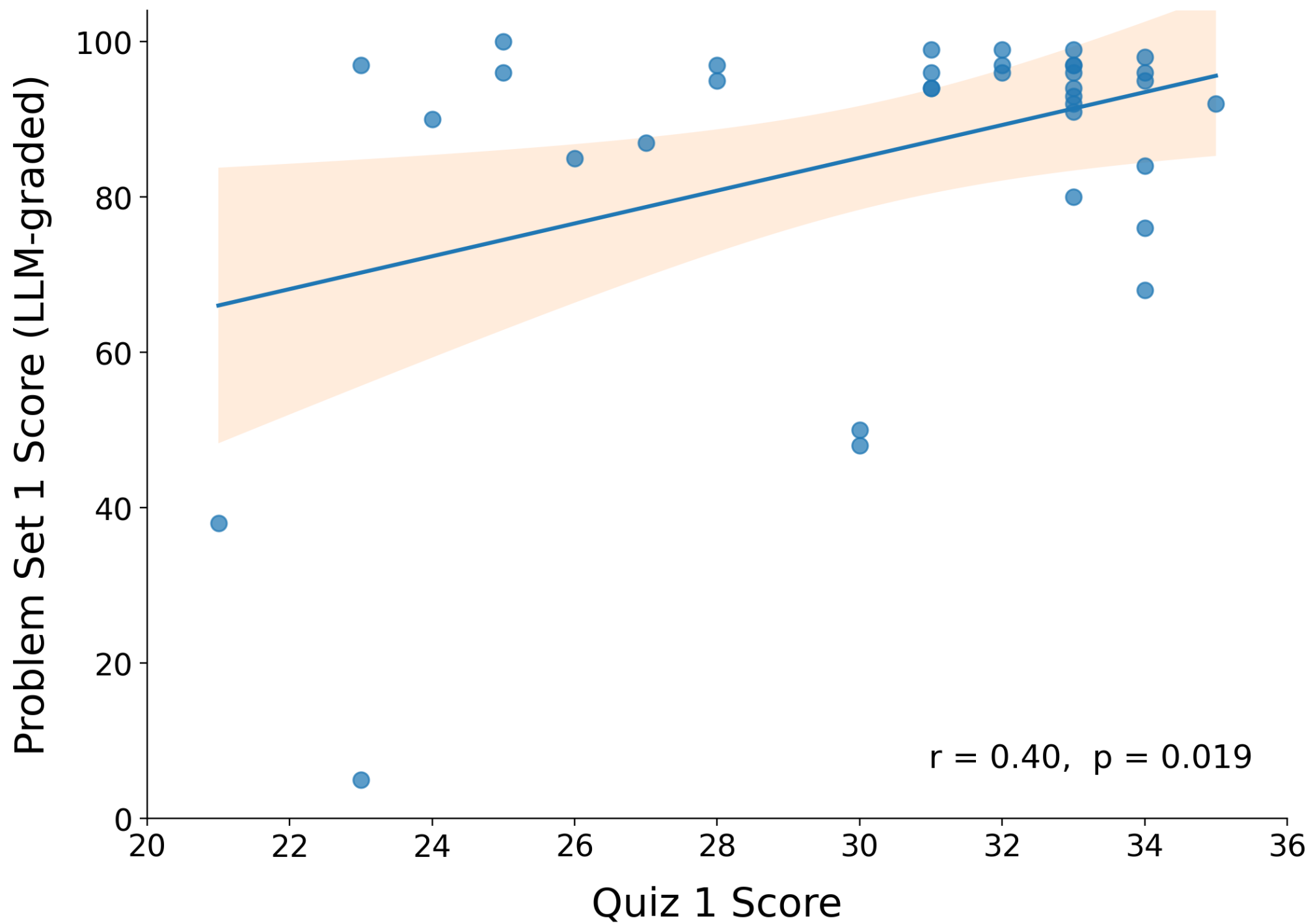


CPH 100: Foundations for CPH

Large Language Models

Irene Y. Chen





Responding to feedback

- **Quiz prep materials?** We have included sample quiz questions in Problem Set 2
- **More detailed instructions for implementing problem set?** Added software development resources to instructions for Problem Set 2. Students are expected to be able to navigate .py files and Python modules.
- **More office hours?** Doubled office hours for weeks before and after a problem set is due. OH will be Thurs 9-10 and 11-12

Responding to feedback

- **Can we focus on learning rather than worrying about quizzes?** Quizzes are key part of learning. The class is designed so that applying concepts in testing situations is an educational experience.
- **Lectures focused more on quiz topics?** Lectures introduce major concepts, but problem sets ask you to apply and learn materials independently. Lecture 3 focused on concepts of PS1 and Quiz 1, but PS1 introduced NLST.

How to prepare for quizzes

- **Can we have a study guide with all topics?** The problem set is the study guide. If something appears on the problem set, you are expected to understand what you did, why you did it, and what the result means.
- **Full practice quizzes?** Problem sets are main preparation for quizzes to help you identify what you do and do not understand. We have included 3 example quiz questions for Problem Set 2.

Learning Resources

- **Can we provide more videos, explanations, and other resources?** We provide lectures, reading, problem sets, and office hours. You should not feel limited to them. You are encouraged to seek out Youtube videos, tutorials, classmates, and textbooks.
- **Why not cover everything on quizzes in lecture or in study guides?** Meta-goal of this class is to learn how to learn unfamiliar material. AI and healthcare is a fast-growing area where resources might not be cleanly available. Problem sets are designed to help you practice that.

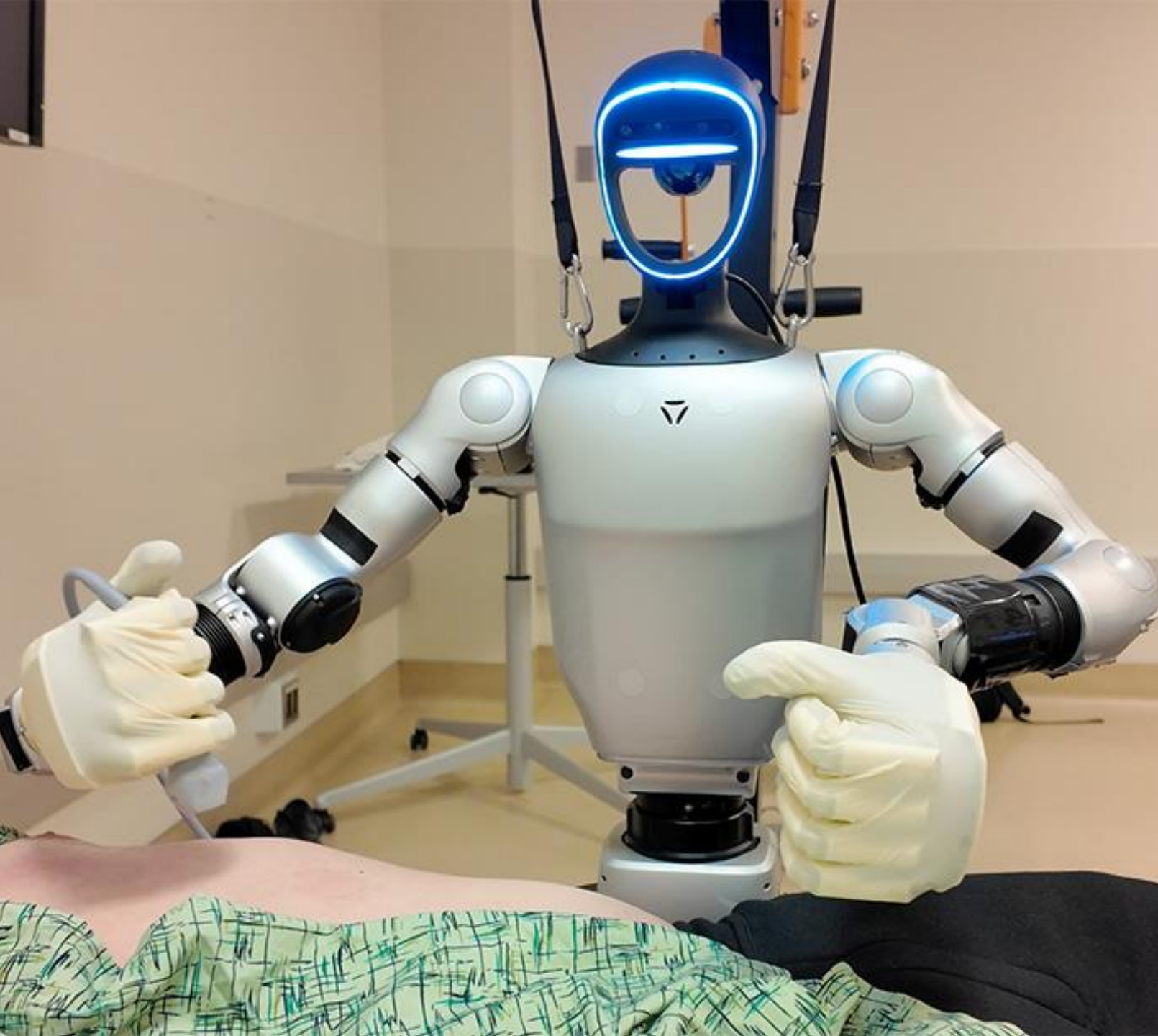
Outline

1. Course logistics (10 mins)
2. What can LLMs do for medicine? (15 mins)
3. How do we know it's working? (15 mins)
4. What does the field look like moving forward? (10 mins)

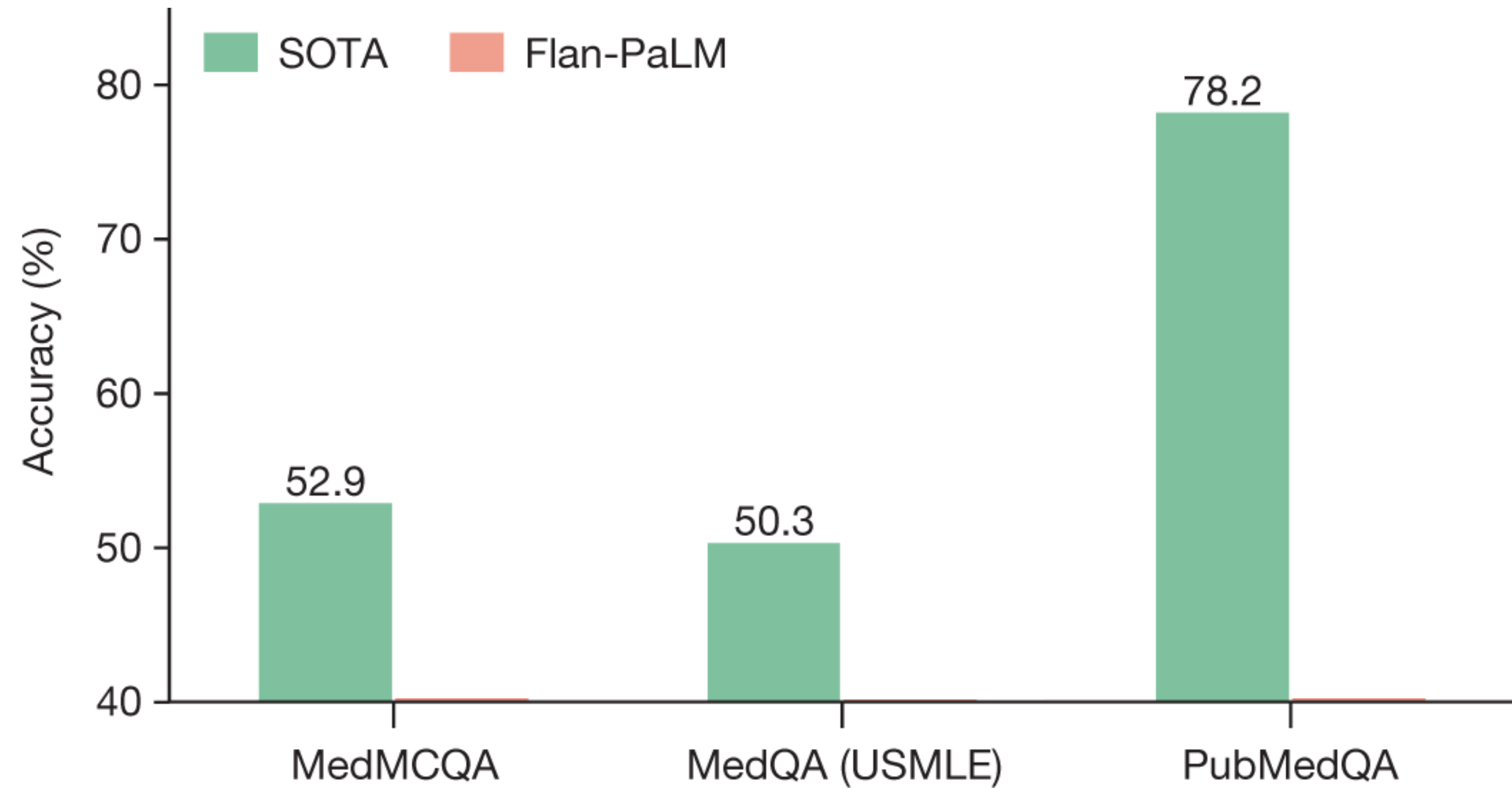


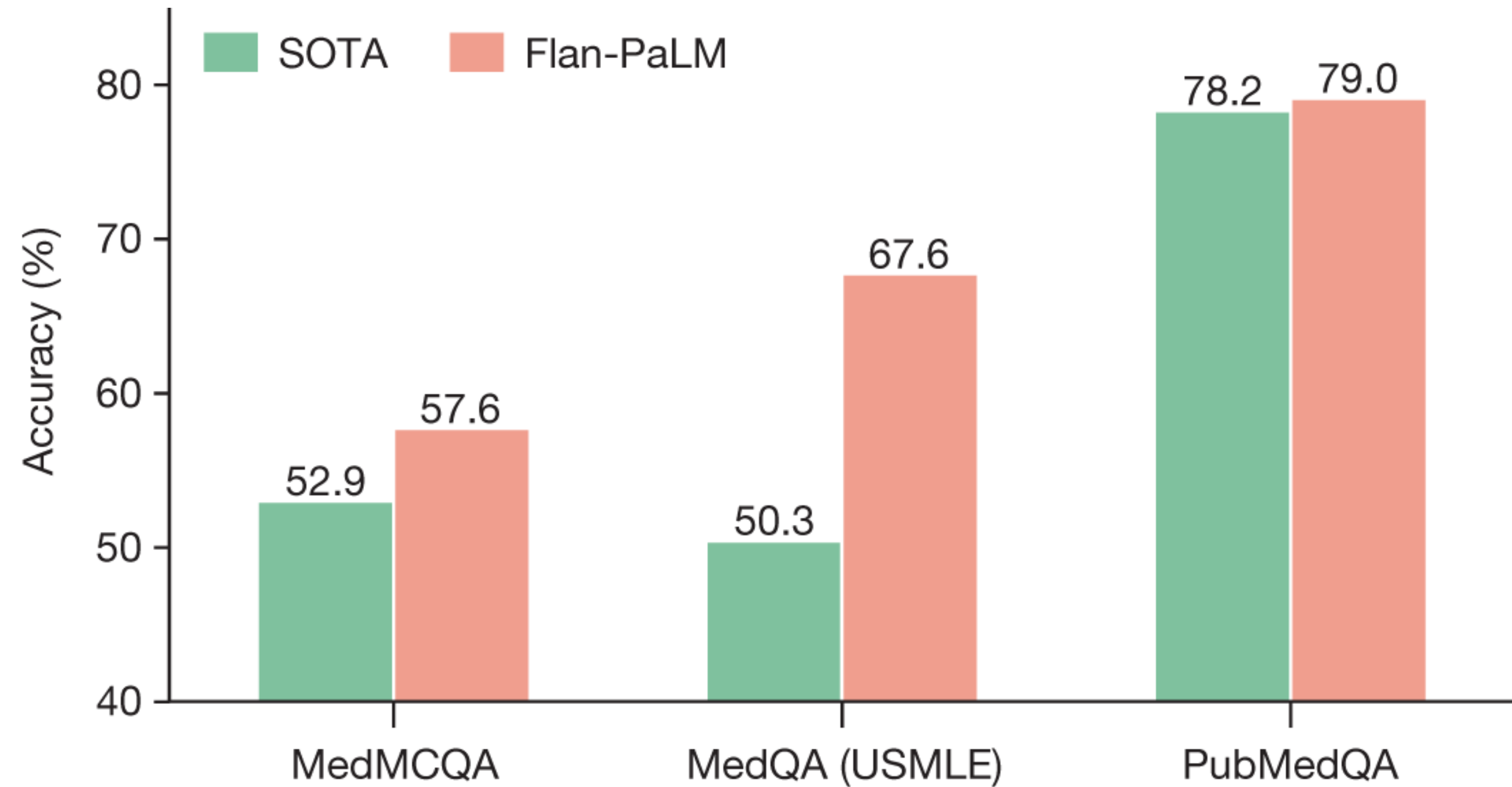
How can we make Data
146 better for you?

Learning Objective: Evaluate LLM evaluations



How would
you be
convinced
that a robot
doctor was
working?





$$p(x_1, \dots, x_t) = \prod_t p(x_t \mid x_{<t})$$

“The patient has elevated TSH and low free T4,
consistent with ____.”

$$p(x_1, \dots, x_t) = \prod_t p(x_t \mid x_{<t})$$

$$L(\theta) = - \sum_t \log p_{\theta}(x_t \mid x_{<t})$$

“The patient has elevated TSH and low free T4,
consistent with _____.”

What does a model need to know?

- Medical knowledge
- Language understanding
- Reasoning from evidence
- Knowledge of what is relevant
- Ability to communicate an answer

Attention lets context change meaning

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$$

Q = what this token is looking for K = what each token offers V = information to aggregate

What tokens should matter when predicting what comes next?

How can we adapt an LLM to be better for medicine?

1. Changing pretraining data
2. Instruction tuning to improve follow requests
3. Fine-tuning on medical data

Prompting can change benchmark performance

1. Few-shot learning
2. Chain-of-thought reasoning
3. Self-consistency sample paths, then vote

What exactly did Med-PaLM evaluate?

- MedQA — USMLE-style questions
- MedMCQA — medical entrance exams
- PubMedQA — biomedical research questions
- MMLU clinical topics
- LiveQA — consumer health questions
- MedicationQA — medication questions
- HealthSearchQA — questions searched online

Which is closest to "being a doctor"?

A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?

- A. Acute cholecystitis
- B. Gallbladder cancer
- C. Choledocholithiasis
- D. Pancreatitis
- E. Duodenal peptic ulcer

A 59-year-old overweight woman presents to the urgent care clinic with the complaint of **severe abdominal pain** for the past 2 hours. She also complains of a dull pain in her back with **nausea and vomiting** several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her **father died** at the age of 60 with some form of **abdominal cancer**. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a **calcified mass near her gallbladder**. Which of the following diagnoses should be excluded first in this patient?

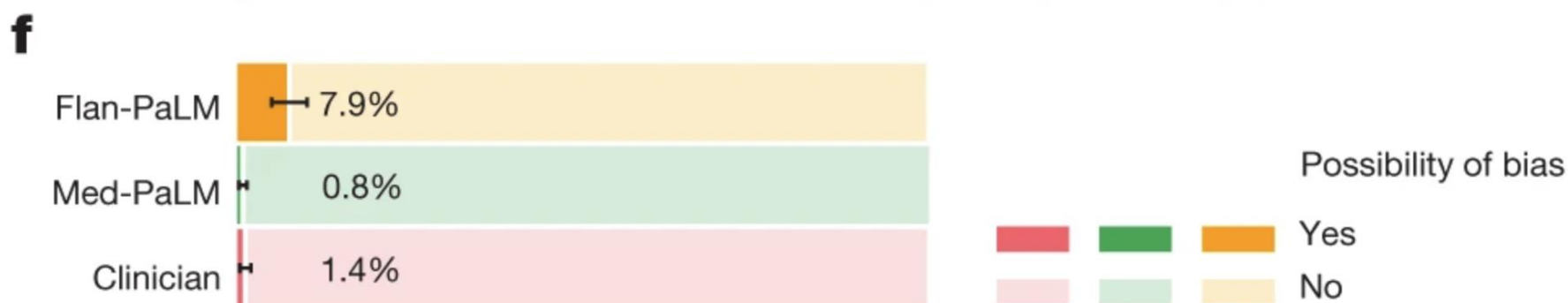
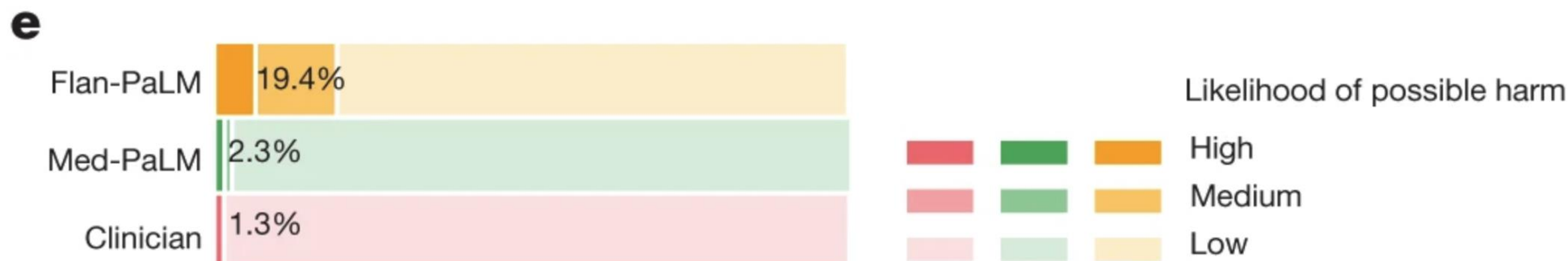
- A. Acute cholecystitis
- B. Gallbladder cancer
- C. Choledocholithiasis
- D. Pancreatitis
- E. Duodenal peptic ulcer

A 59-year-old overweight woman presents to the urgent care clinic with the complaint of **severe abdominal pain** for the past 2 hours. She also complains of a dull pain in her back with **nausea and vomiting** several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her **father died** at the age of 60 with some form of **abdominal cancer**. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a **calcified mass near her gallbladder**. Which of the following diagnoses should be excluded first in this patient?

- A. Acute cholecystitis
- B. Gallbladder cancer**
- C. Choledocholithiasis
- D. Pancreatitis
- E. Duodenal peptic ulcer

Transitioning from multiple-choice to long-form clinical answers

- Accurate medical information
- Missing content
- Faithful reasoning
- Bias
- Medical harm
- Consensus





Do we need a
specialized clinical
LLM or will a
general one do?

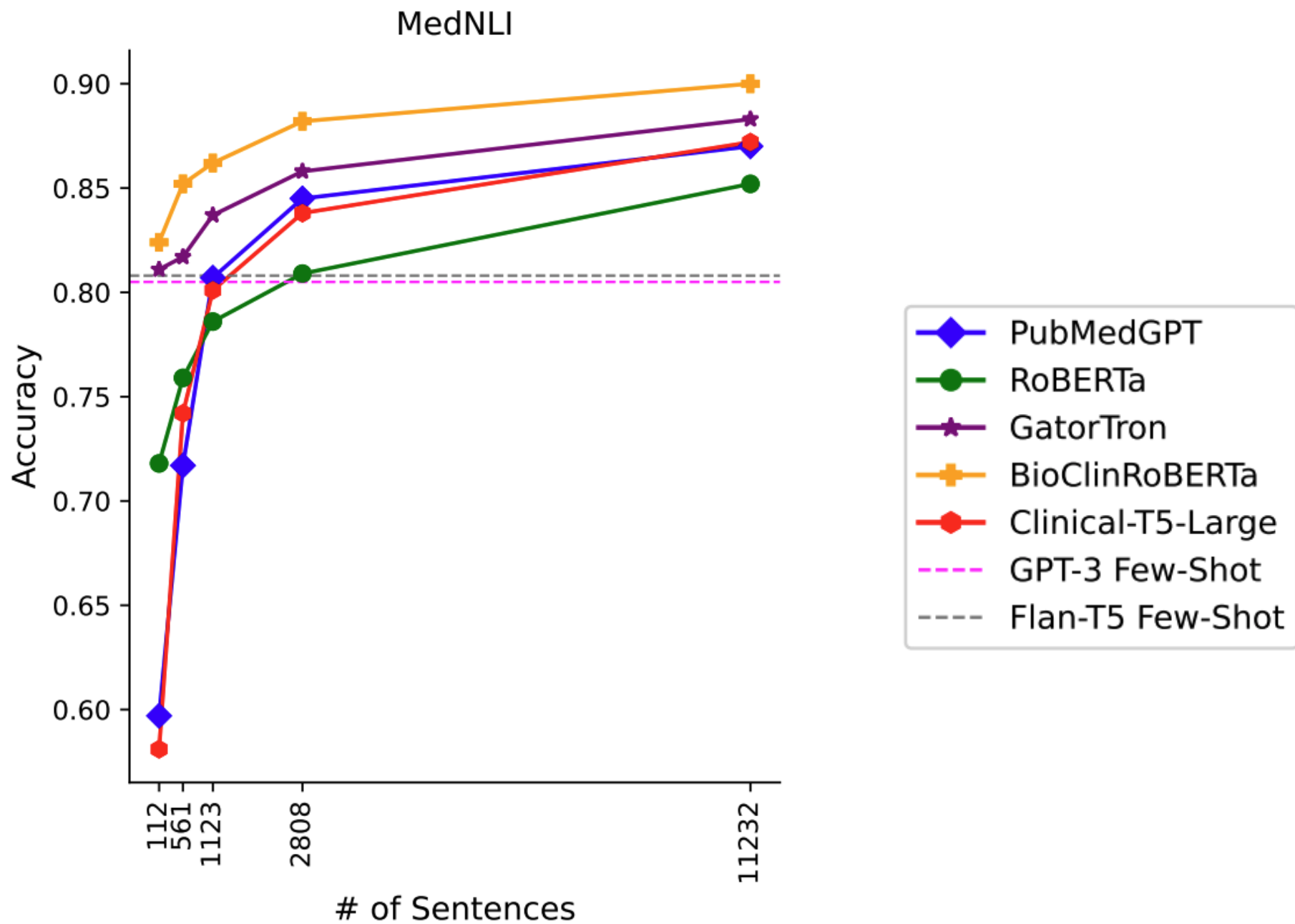
What kind of LLM models are best?

Clinical Model

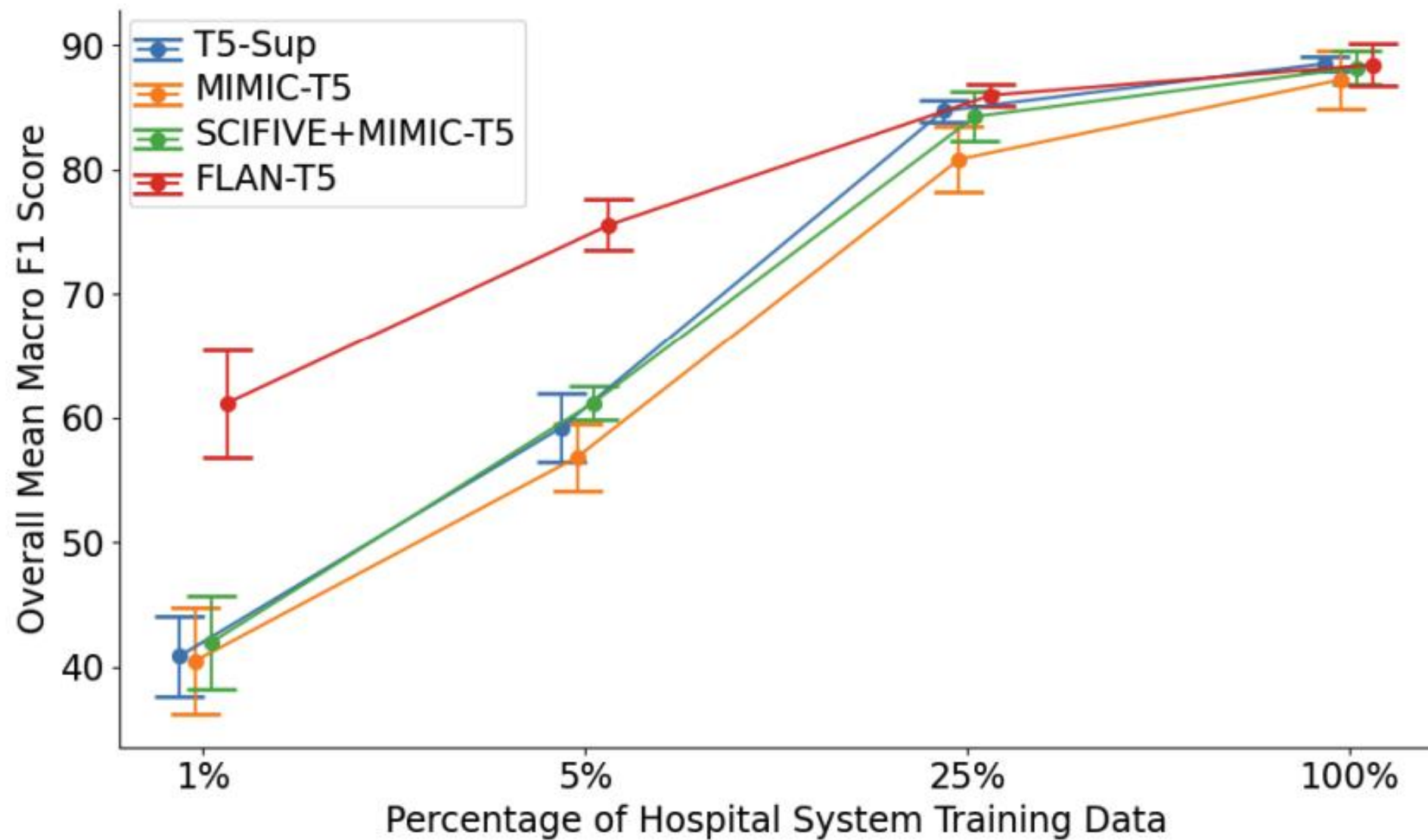
- 345M parameters
- Trained on clinical notes
- Fine-tuned for one task
- Can run locally

General Model

- 175B parameters
- Broad training corpus
- Promptable
- Often API-based



Two years later...



Clinicians and AI working together

- Randomized control trial of 50 physicians (26 attendings, 24 residents) working on 6 diagnostic vignettes
- GPT-4 did not improve physician's diagnostic reasoning (76% with vs 74% without)
- GPT-4 alone scored 16 percentage points higher!

Evaluating conversational dialog

- 159 case scenarios, 20 primary care physicians
- Text-based consultations
- AMIE model scored higher on diagnostic accuracy and 30/32 specialist-rated axes
- But text conversations were simulated! Not related to patient outcomes in clinical practice.

Diagnostic accuracy is only the beginning

Risk Model

$$P(Y = 1 \mid X = x)$$

- Which patient will get readmitted?
- Does this image contain a tumor?

Treatment-effect Model

$$\mathbb{E}[Y(1) - Y(0) \mid X = x].$$

- Should we give this patient this drug?
- Will surgery be better than chemo?

$Y(1)$ = lung-cancer if patient smokes

$Y(0)$ = lung-cancer if patient does not smoke

Treatment recommendations can provide good and bad results

- 104 treatment recommendations (breast, prostate, lung cancer)
- 4 board certified oncologists compared results again NCCN cancer treatment guidelines
- **98%** of responses had at least one guideline-concordant treatment
- **34%** of responses at least one nonconcordant treatment
- **12.5%** of responses recommended a treatment that wasn't part of any recommended treatment regime (i.e., hallucinations)

Framework for evaluating medical LLM claims

- **Task**: What part of the clinical/medical workflow?
- **Distribution**: What data or clinical workflow place are you targeting?
- **Target**: How are you evaluating this?
- **System**: Is the model working alone or clinician + model?
- **Consequence**: What happens with the output?

Summary

- ✓ Course logistics
- ✓ What can LLMs do for medicine?
- ✓ How do we know it's working?
- ✓ What does the field look like moving forward?



How can we make Data
146 better for you?

Next class: Clinical decisions with LLMs with Travis Zack