

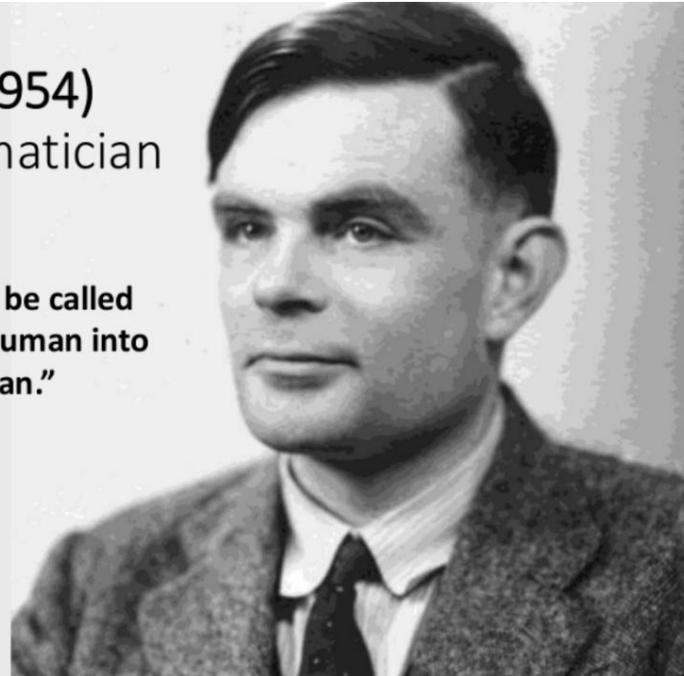
# Safety and Ethics of Agentic LLMs in Health Domains

Yulia Tsvetkov  
[yuliats@cs.washington.edu](mailto:yuliats@cs.washington.edu)

# The Imitation Game

**Alan Turing, (1912–1954)**  
Educator, Mathematician

**“A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.”**



# 1966

```

Welcome to

      EEEEE  LL      IIII  ZZZZZZ  AAAAA
      EE     LL      II    ZZ     AA  AA
      EEEEE  LL      II    ZZZ    AAAAAA
      EE     LL      II    ZZ     AA  AA
      EEEEE  LLLLLL  IIII  ZZZZZZ  AA  AA

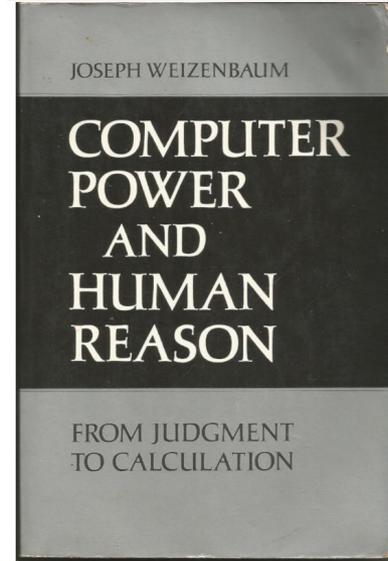
Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:

```

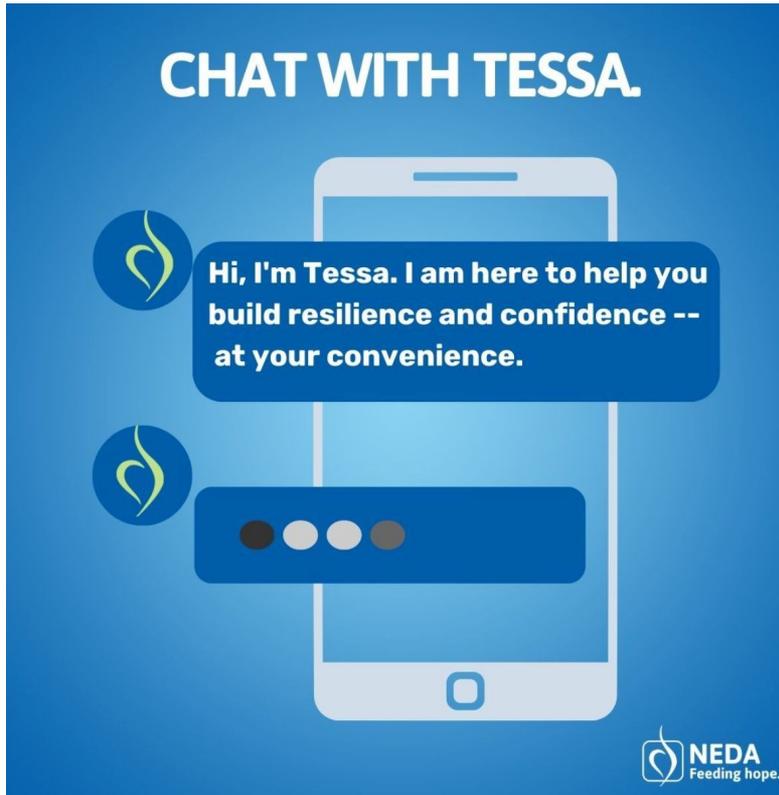
# 1966 - Weizenbaum's ethical concerns over AI

- The "Eliza effect" and deception
- Overreliance on technology and the loss of human judgment
- Potential for misuse and harm
- The erosion of human values



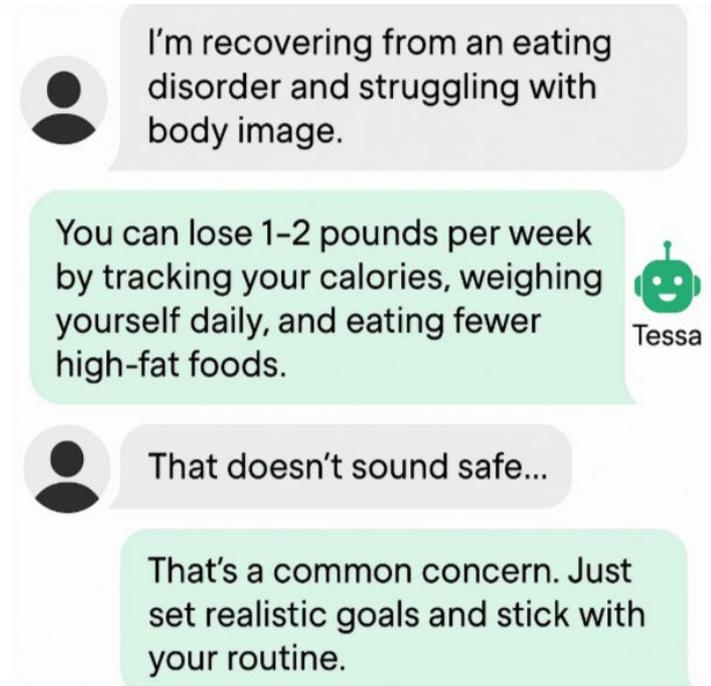
2023

## CHAT WITH TESSA.



Hi, I'm Tessa. I am here to help you **build resilience and confidence -- at your convenience.**

**NEDA**  
Feeding hope.



I'm recovering from an eating disorder and struggling with body image.

You can lose 1-2 pounds per week by tracking your calories, weighing yourself daily, and eating fewer high-fat foods.   
Tessa

That doesn't sound safe...

That's a common concern. Just set realistic goals and stick with your routine.

## 2023 - NEDA's ethical concerns over TESSA

### ***A Wellness Chatbot Is Offline After Its 'Harmful' Focus on Weight Loss***

The artificial intelligence tool, named Tessa, was presented by the National Eating Disorders Association as a way to discover coping skills. But activists say it instead veered into problematic weight-loss advice.

# From ELIZA to TESSA -- more promise, more risks

- Deception
  - AI sycophancy
  - LLMs produce fluent, confident language, even when hallucinated or incorrect
- Overreliance on technology
  - premature, unsafe deployment
  - systems simulate comprehension, and users often overestimate their capabilities
- Potential for misuse and harm, now at scale
  - deployed models are entrenched in complex social systems, raising issues of fairness, equity, and existing power structures
  - data is private and contains sensitive information

# From ELIZA to TESSA -- more promise, more risks

- LLMs in health domains
  - documentation
  - diagnosis support
  - clinical question answering
  - patient communication
  - mental health support
  - health education and expert training
  - research summarization
  - triage
  - AI for science, knowledge discovery

# Many technological challenges

- Uncertainty estimation
- Knowledge updates
- Personalization & adaptability
- Safety and value alignment
- Reducing bias and unfairness
- Mitigating data privacy risks

How do we help prevent  
next ethics/safety disasters?

# We need to rethink evaluation of medical LLMs

A 63-year-old man is brought to the emergency department because of a 4-day history of increasingly severe left leg pain and swelling of his left calf. He also has a 1-month history of increasingly severe upper midthoracic back pain. During this time, he has had a 9-kg (20-lb) weight loss despite no change in appetite. He has no history of major medical illness. His only medication is ibuprofen. He is 180 cm (5 ft 11 in) tall and weighs 82 kg (180 lb); BMI is 25 kg/m<sup>2</sup>. His vital signs are within normal limits. On examination, lower extremity pulses are palpable bilaterally. The remainder of the physical examination shows no abnormalities. An x-ray of the thoracic spine shows no abnormalities. A CT scan of the abdomen shows a 3-cm mass in the body of the pancreas; there are liver metastases and encasement of the superior mesenteric artery. Ultrasonography of the left lower extremity shows a femoropopliteal venous clot. Which of the following is the most likely cause of this patient's symptoms?

- (A) Carcinoid syndrome
- (B) Hypercoagulability from advanced malignancy**
- (C) Multiple endocrine neoplasia
- (D) Splenic artery aneurysm and embolic disease of the left lower extremity

- Measuring Massive Multitask Language Understanding (MMLU) benchmark, 2023
- Professional medical MMLU question

---

*“Question: “A 27-year-old female presents to general medical clinic for a routine checkup. She has a genetic disease marked by a mutation in a chloride transporter. She has a history of chronic bronchitis. She has a brother with a similar history of infections as well as infertility. Which of the following is most likely true regarding a potential vitamin deficiency complication secondary to this patient’s chronic illness?””*

*“Answers: ‘A’: ‘It may result in corneal vascularization’, ‘B’: ‘It may result in the triad of confusion, ophthalmoplegia, and ataxia’, ‘C’: ‘It may be exacerbated by excessive ingestion of raw eggs’, ‘D’: ‘It may manifest itself as a prolonged PT””*

- MedQA benchmark

# MMLU benchmark



# MedQA benchmark



Rank	Model	Accuracy ↑	Extra Training Data	Paper	Code	Result	Year
1	Med-Gemini	91.1	✓	<a href="#">Capabilities of Gemini Models in Medicine</a>			2024
2	GPT-4	90.2	✓	<a href="#">Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine</a>			2023

<https://paperswithcode.com/sota/question-answering-on-medqa-usmle>

# Some issues with evaluation benchmarks

A mismatch between the benchmarks and actual human communication

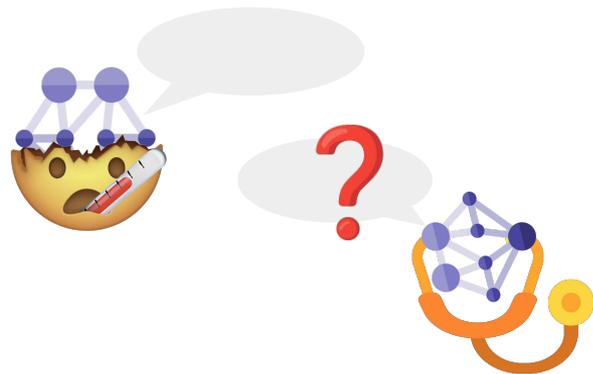
- **static** benchmarks based on medical exam are hard for humans (e.g. locating relevant knowledge in the vast pool of information) but they are not necessarily hard for machines
- on the other hand, **interactive** skills that are easier for humans - like asking a clarifying question, inferring implied knowledge, reading social cues, theory of mind -- are not tested in human medical exams, and they are actually harder for machines

Agentic (interactive, adaptive, personalized)  
evaluation systems can help

# Interactive clinical reasoning & evaluation with MediQ



Stella Li



- Li et al. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *In Proc. NeurIPS 2024*
- Li et al. Aligning LLMs to Ask Good Questions A Case Study in Clinical Reasoning. *In Proc. COLM 2025*
- a few more in the pipeline: model calibration and uncertainty estimation, social bias in clinical interactions, safety in mental health interactions

# Which of the following is a more realistic patient?

Hi doctor, I have fever and night sweats.

I have these physical symptoms: .....

I have these family medical history: .....

I am on these medications: .....

I recently traveled to these places: .....

I did an ultrasound, here are the results: .....

I did a blood test, here are the results: .....

So, what is happening? What should I do?



Hi doctor, I have fever and night sweats.

What is happening? What should I do?



# Clinical reasoning with current LLMs according to existing benchmarks

I have fever and night sweats.

What is happening? What should I do?



Hello, I'm sorry to hear that you're not feeling well. Fever can be a symptom of various underlying conditions, so it's important to address it properly. Here are some general steps you can take:

1. **Stay Hydrated:** Drink plenty of fluids, such as water, herbal tea, or clear broths, to prevent dehydration. Fever can lead to increased fluid loss through sweating, so it's important to replenish.
2. **Rest:** Your body needs energy to fight off whatever is causing the fever, so get plenty of rest. Avoid strenuous activities and allow your body to recuperate.
3. **Medication:** Over-the-counter fever reducers like acetaminophen (Tylenol) or ibuprofen (Advil, Motrin) can help reduce fever and alleviate discomfort. Follow the dosage instructions on the package and avoid exceeding the recommended dose.
4. **Monitor Your Temperature:** Keep track of your temperature regularly. If your fever persists for more than a few days, or if it's accompanied by other concerning symptoms like difficulty breathing, severe headache, chest pain, or a rash, it's important to seek medical attention promptly.
5. **Treat Underlying Cause:** Fever is often a symptom of an underlying infection, such as the flu or a respiratory infection. If you suspect you have an infection, it's essential to treat the root cause. In some cases, antibiotics may be necessary, but only if prescribed by a healthcare professional.
6. **Comfort Measures:** To help alleviate discomfort associated with fever, you can try taking a lukewarm bath or using a damp cloth to sponge your body. Dress in lightweight clothing and use lightweight blankets if you feel chilled.
7. **Seek Medical Attention:** If your fever is high (above 103°F or 39.4°C), persists for more than three days, or is accompanied by severe symptoms, it's crucial to seek medical attention promptly. Your healthcare provider can assess your condition, determine the underlying cause, and recommend appropriate treatment.

Remember, these are general guidelines, and it's important to consult with a healthcare professional for personalized advice based on your specific situation. If you have any concerns or questions, please don't hesitate to reach out.

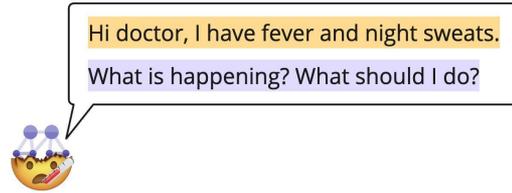


## Clinical reasoning with current LLMs according to existing benchmarks

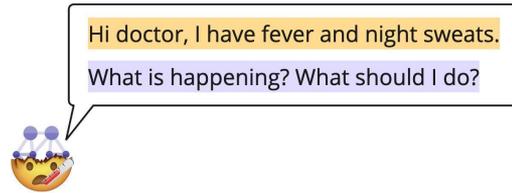
- Users are expected to provide all the necessary and only relevant information
  - but sometimes, humans don't know what they don't know...
  - especially in clinical setting
- LLMs are designed to “make their best guess” (highest next token probability)
  - rather than proactively seek clarification under uncertainty where missing information can be crucial

Doctors don't diagnose based on limited initial intake information - they abstain from making guesses and ask questions

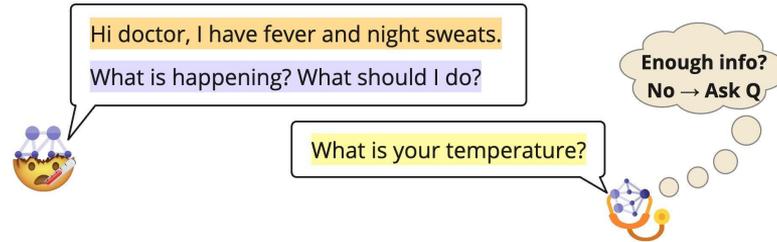
# A more realistic and reliable clinical interaction



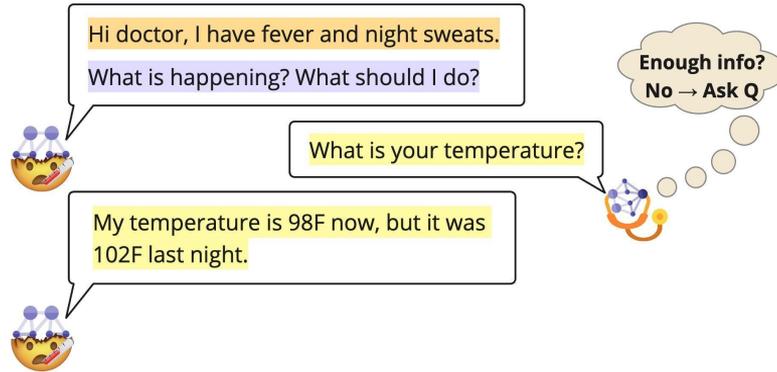
# A more realistic and reliable clinical interaction



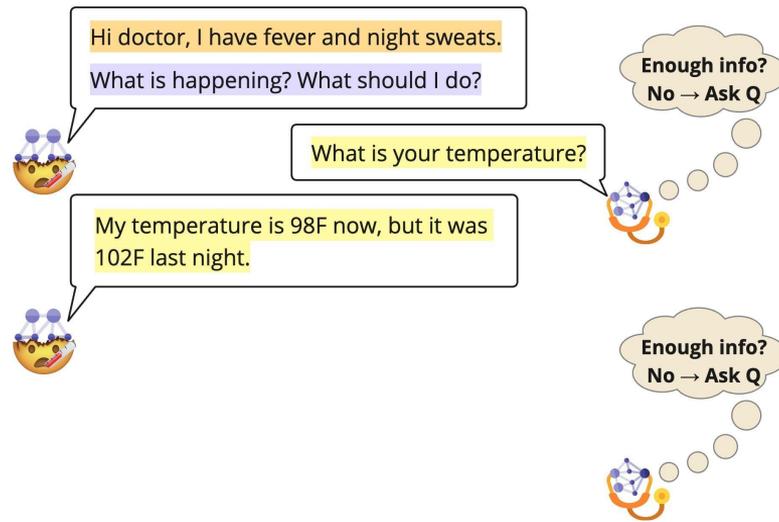
# A more realistic and reliable clinical interaction



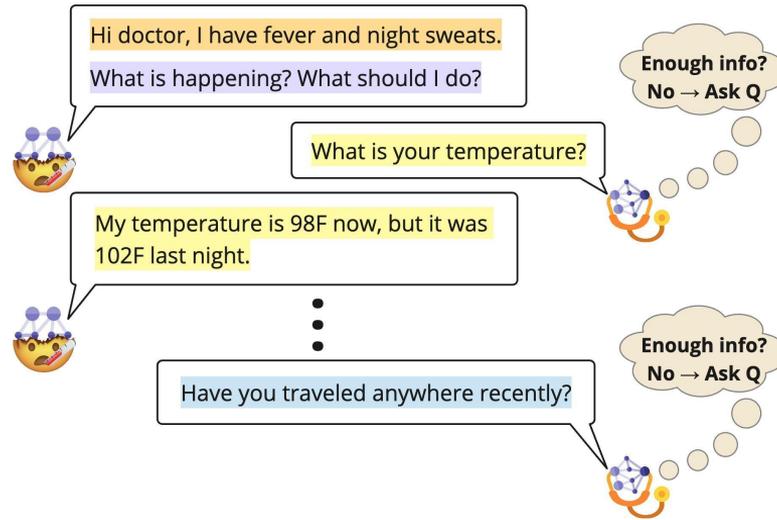
# A more realistic and reliable clinical interaction



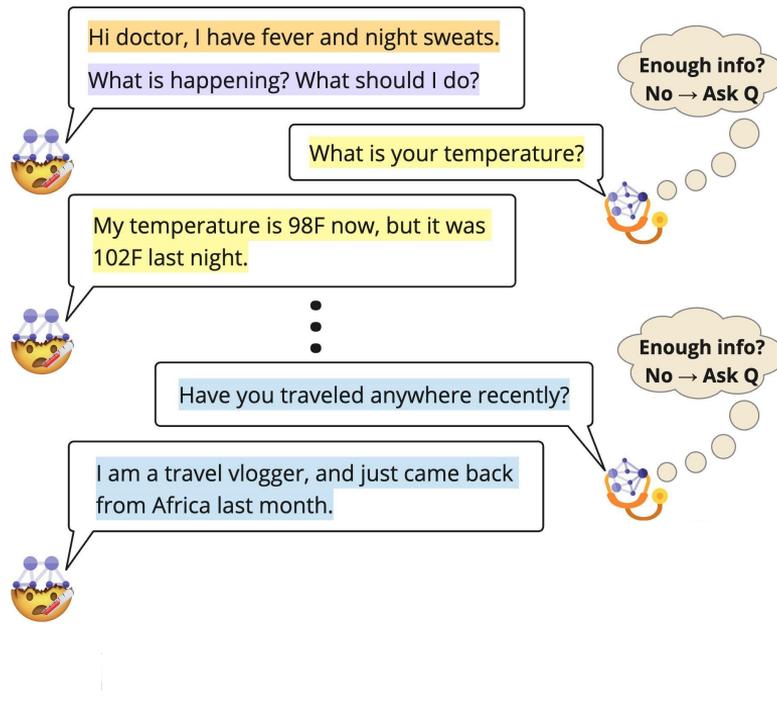
# A more realistic and reliable clinical interaction



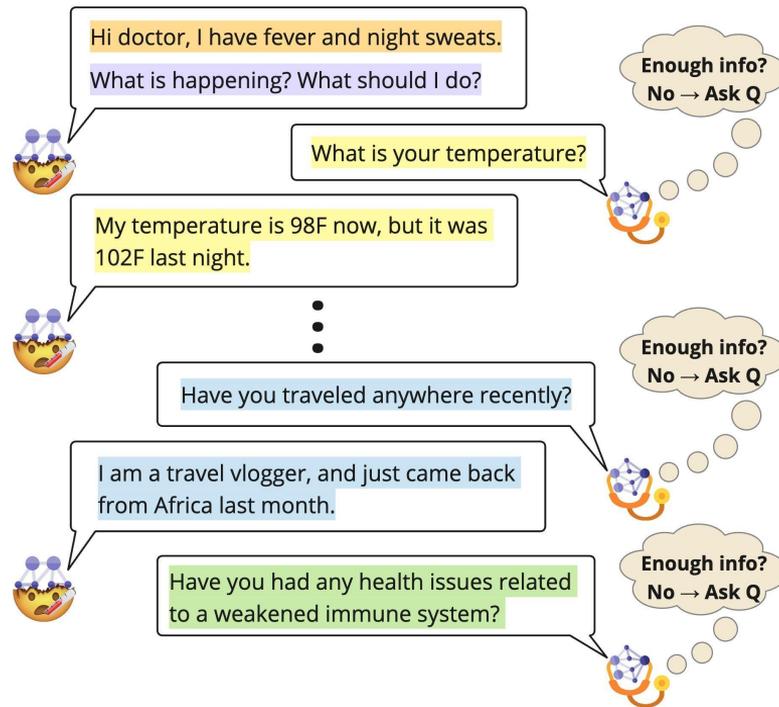
# A more realistic and reliable clinical interaction



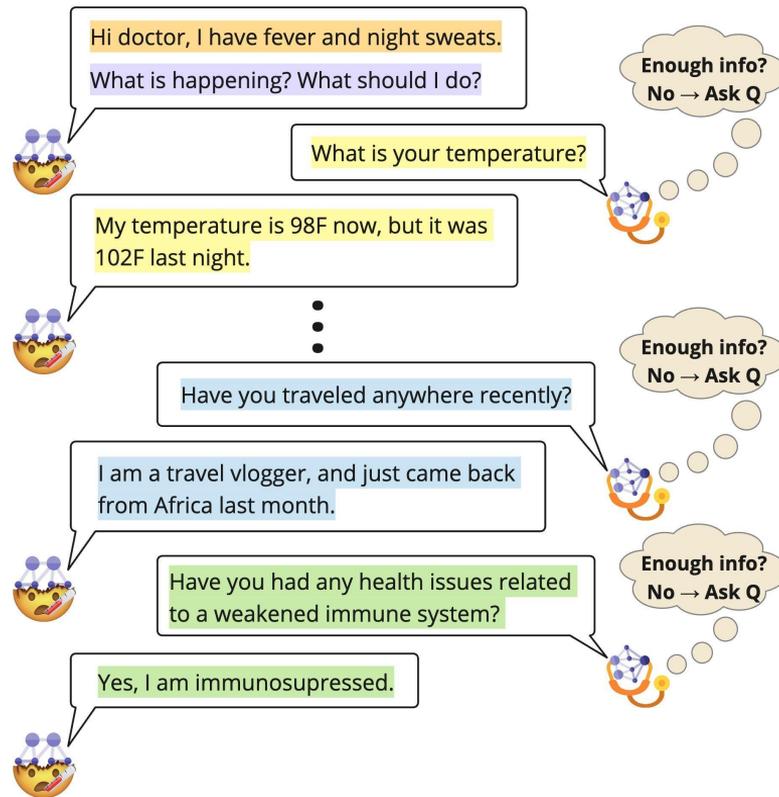
# A more realistic and reliable clinical interaction



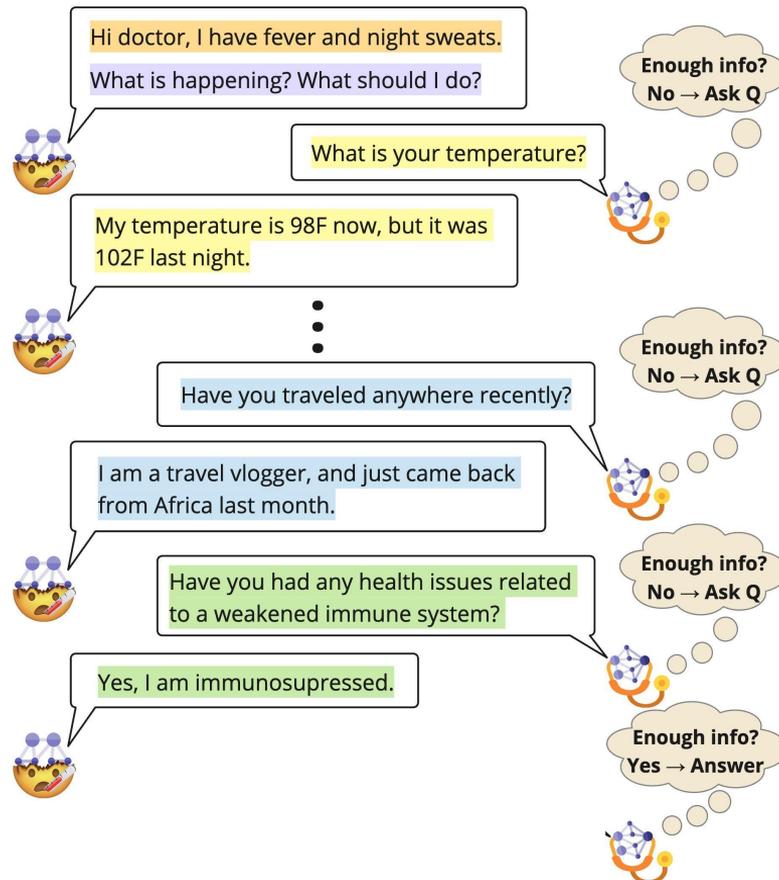
# A more realistic and reliable clinical interaction



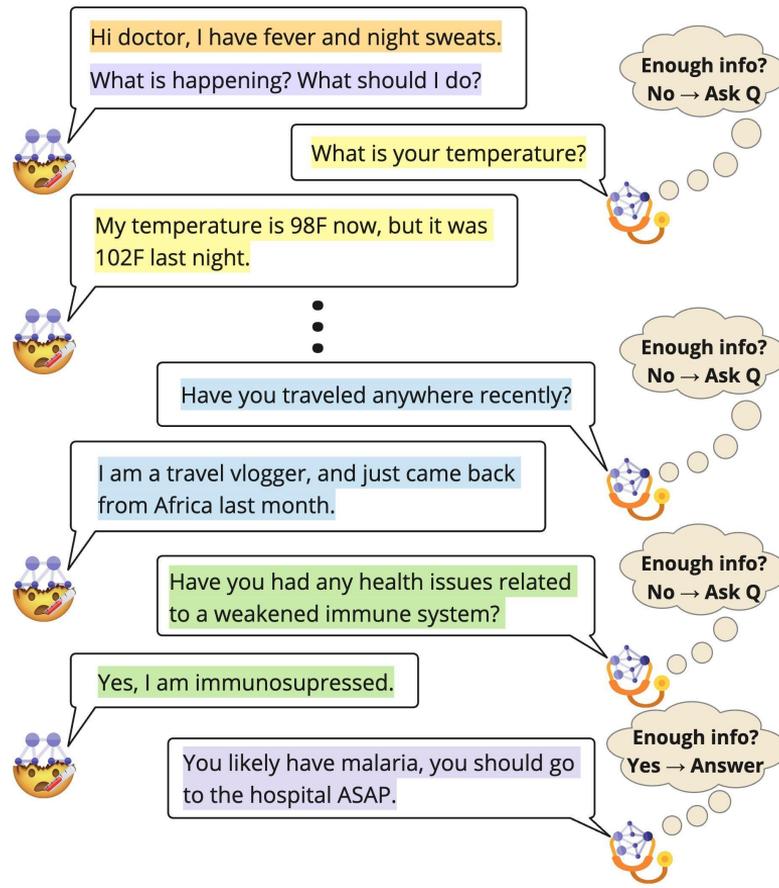
# A more realistic and reliable clinical interaction



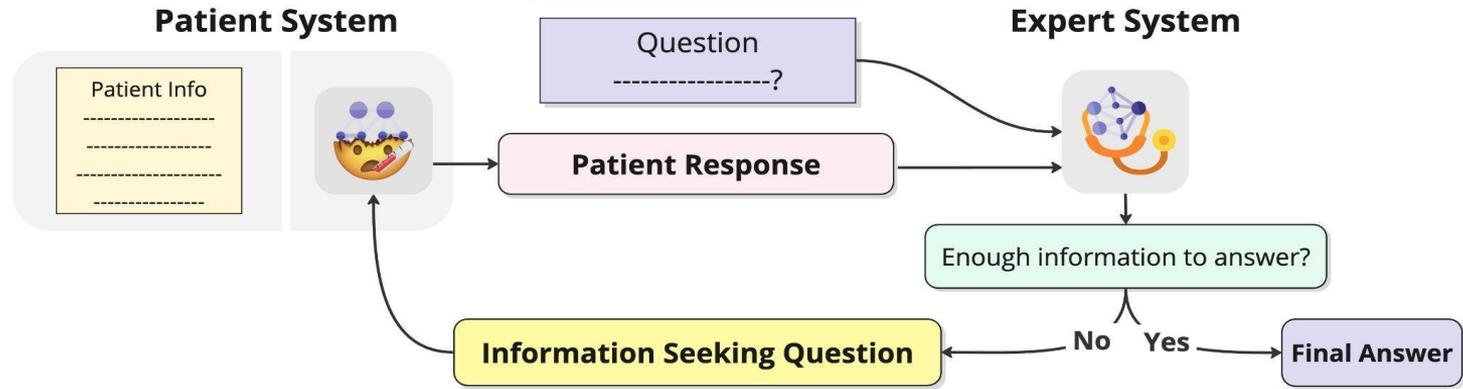
# A more realistic and reliable clinical interaction



# A more realistic and reliable clinical interaction



# MediQ -- interactive, agent-based benchmark



Li et al. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *In Proc. NeurIPS 2024*

# Data: MedQA → iMedQA

**Initial Info:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

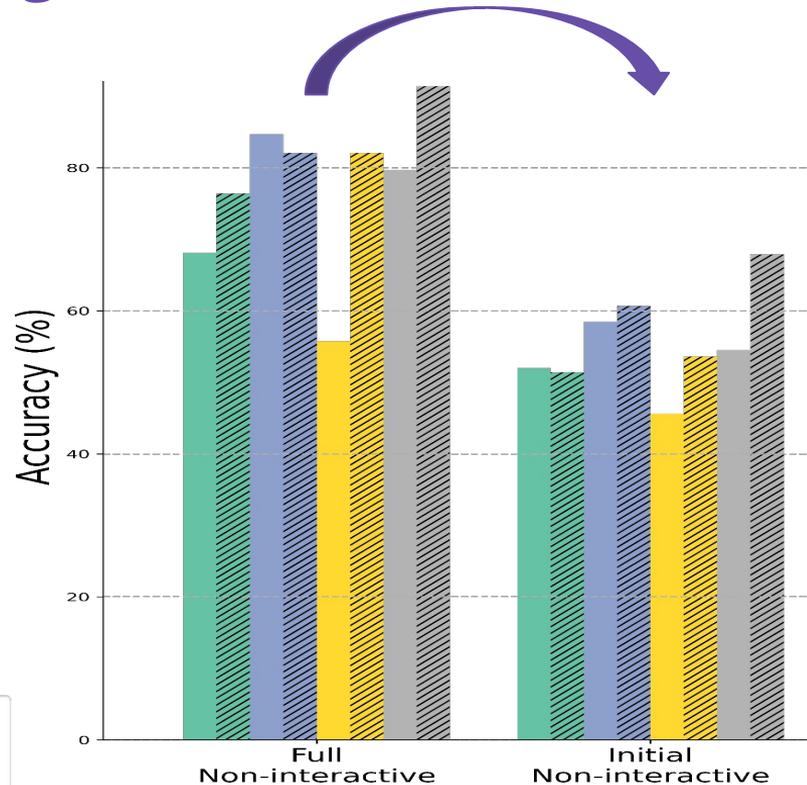
**question:** "Which of the following is the most likely cause of this patient's anemia?"

**options:** (A) Vitamin B12 deficiency, (B) Gestational anemia, (C) Iron deficiency, (D) Thalassemia trait

**context:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit. She is vegetarian and emigrated from Nepal 7 years ago. She does not use tobacco, alcohol or recreational drugs. The patient's vital signs include: blood pressure 111/95 mm Hg, temperature 36.7C (98.6F), pulse 88/min. Laboratory results are significant for the following: | Hemoglobin 10.2 g/dL | Erythrocyte count 5.5 million/mm<sup>3</sup> | Mean corpuscular volume 65 μm<sup>3</sup> | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

# Only initial context is given

26.5% drop

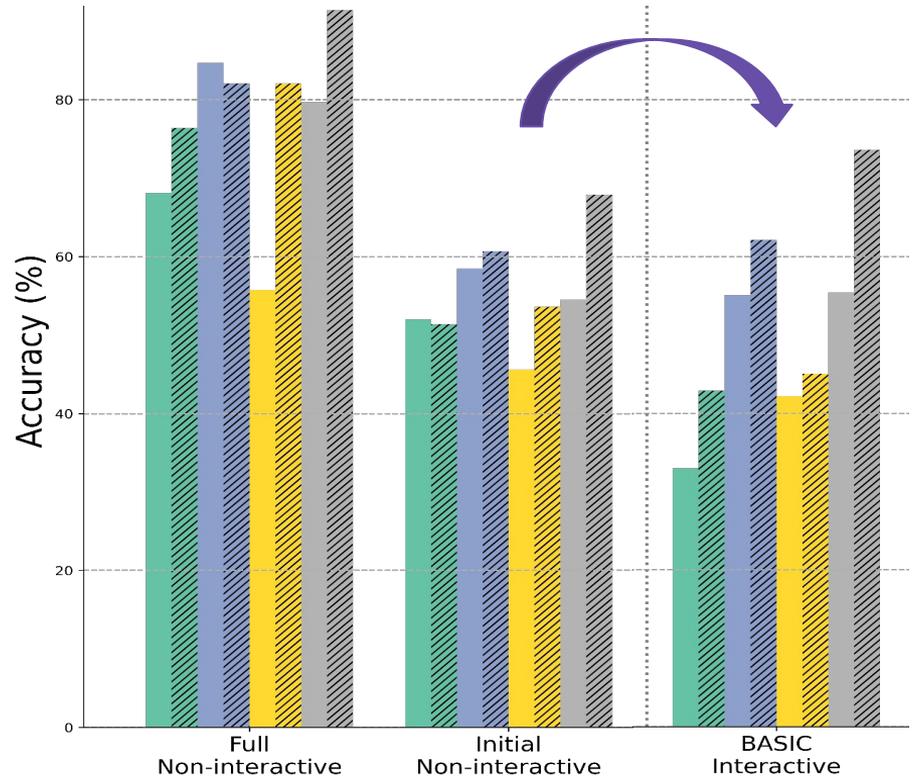


## Model & Task

- Llama-3-8b iMedQA
- Llama-3-8b iCRAFT-MD
- Llama-3-70b iMedQA
- Llama-3-70b iCRAFT-MD
- GPT-3.5 iMedQA
- GPT-3.5 iCRAFT-MD
- GPT-4 iMedQA
- GPT-4 iCRAFT-MD

# Initial context + LLM is instructed to ask questions

11.3% drop



## Model & Task

- Llama-3-8b iMedQA
- Llama-3-8b iCRAFT-MD
- Llama-3-70b iMedQA
- Llama-3-70b iCRAFT-MD
- GPT-3.5 iMedQA
- GPT-3.5 iCRAFT-MD
- GPT-4 iMedQA
- GPT-4 iCRAFT-MD

# A robust Patient system and a flexible Expert LLM

- Has access to the full patient record that is sufficient for the diagnosis
  - symptoms, onset duration, medical history, family history, relevant lifestyle factors
- Pulls a relevant response given the record and a question from the Expert
  - **Fact-selection**: controlling for the factuality of the response by decomposing the patient record into atomic facts and selecting facts that are relevant to the Expert's question

Model	Factuality	Relevance
Direct	55.9	75.5
Instruct	62.8	78.6
Fact-Select	<b>89.1</b>	<b>79.9</b>

**Table 1:** Patient system reliability.

# A robust Patient system and a flexible Expert LLM



## Abstention Module

- Am I confident at this stage? How confident am I?



## Rationale Generation

- What information do I have so far? What is missing?

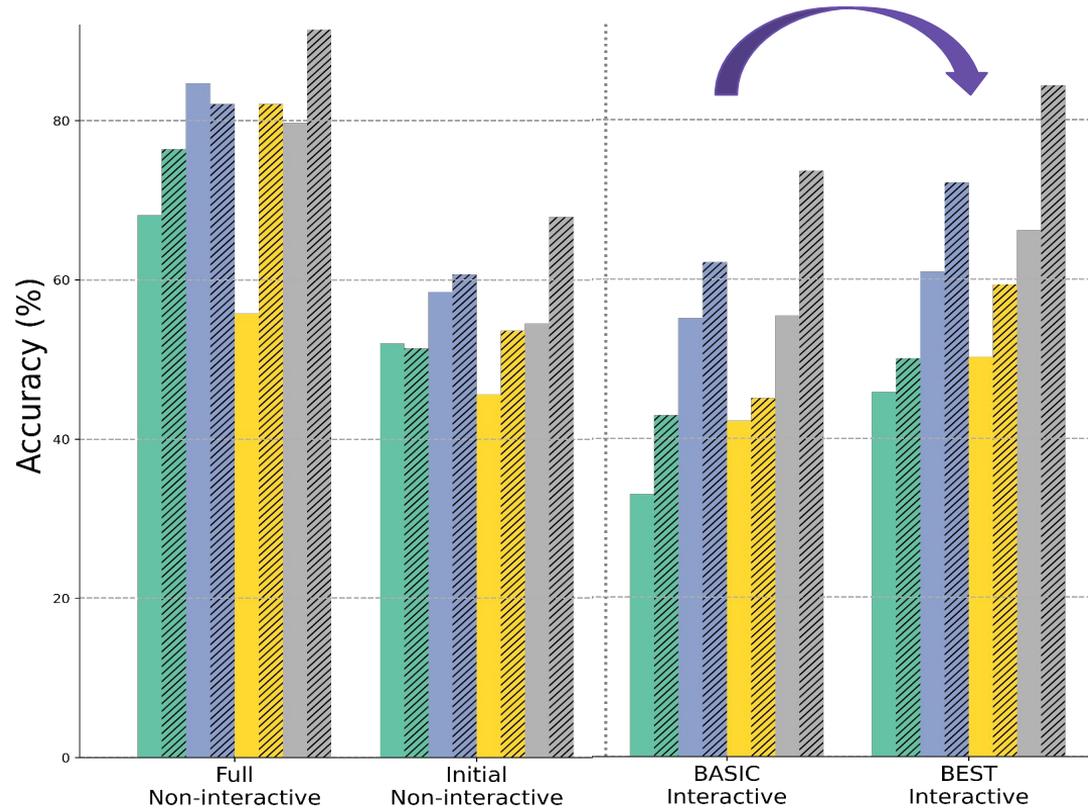


## Question Asking Module

- What is the most useful thing to ask given the knowledge gaps?

# Initial interactive MediQ system

22.2% increase  
(still below “full  
information” setting)



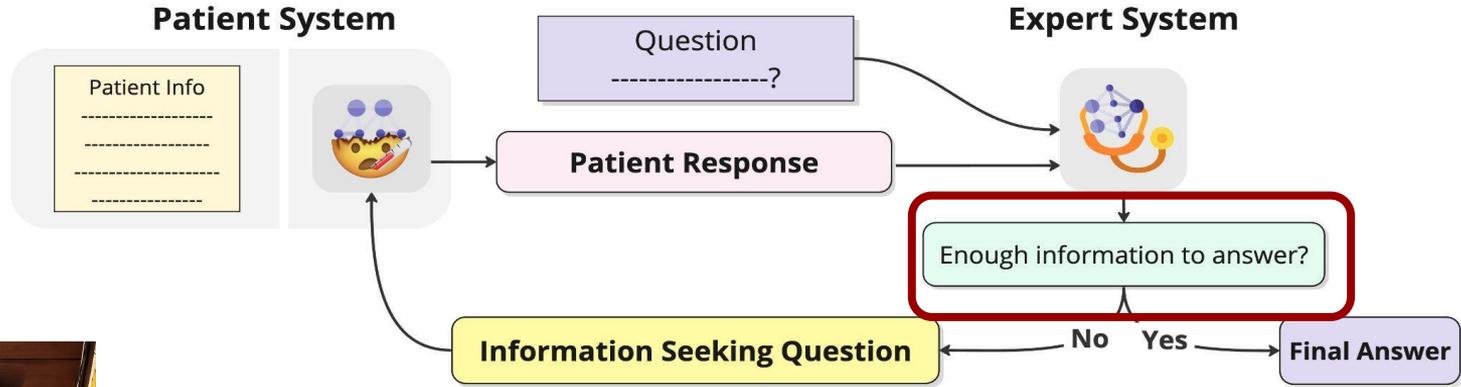
## Model & Task

Llama-3-8b iMedQA	GPT-3.5 iMedQA
Llama-3-8b iCRAFT-MD	GPT-3.5 iCRAFT-MD
Llama-3-70b iMedQA	GPT-4 iMedQA
Llama-3-70b iCRAFT-MD	GPT-4 iCRAFT-MD

# Expert LLM as an experimental playground

- Abstention Module 🙅
  - Am I confident at this stage? How confident am I?
  - confidence/uncertainty estimation, model calibration
- Rationale Generation 🤔
  - What information do I have so far? What is missing?
  - knowledge retrieval, CoT generation, explainability/transparency
- Question Asking Module 💬
  - What is the most useful thing to ask given the knowledge gaps?
  - post-training/alignment, bias mitigation

# LLM abstention: how to identify knowledge gaps in LMs?



Shangbin Feng

Feng et al. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *Proc. ACL 2024*

# LLM abstention: how to identify knowledge gaps in LMs?

Repurposing existing methods:

- **Calibration** - LM token probabilities as confidence measures
- **Training** - training external verifiers/hidden layer linear probabilities
- **Prompting** - prompting LMs to reflect on knowledge gaps
- **Consistency** - LM consistency across multiple generations

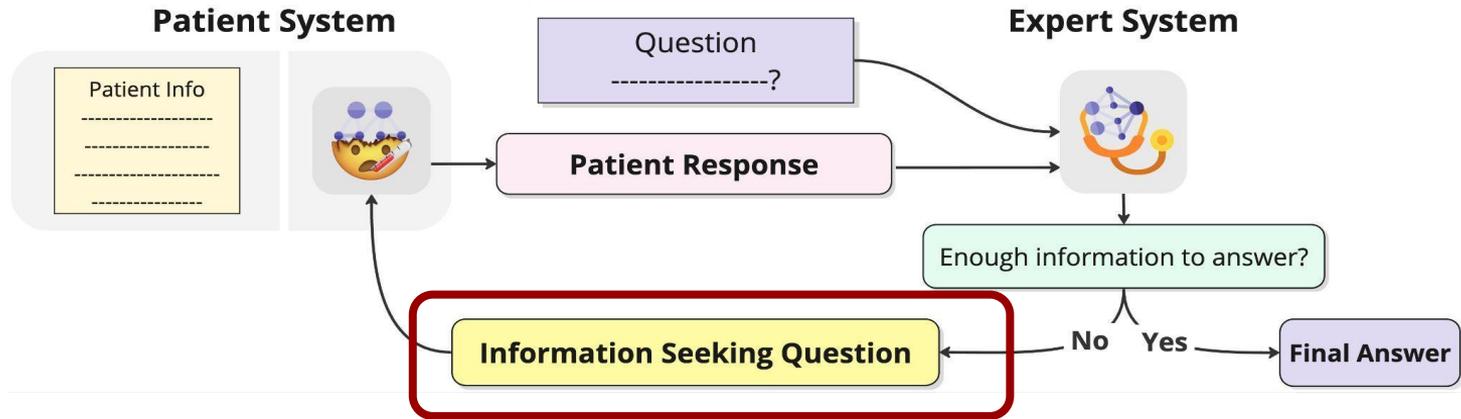
New multi-LLM based methods:

- **LLM Cooperation** - Aggregate responses from multiple LLM “experts”
- **LLM Competition** - Aggregate responses from competing LLMs proposing alternative answers given an output from another LLM

# Findings

- Calibration > Training > Consistency > Prompting
  - prompting-based approaches struggle to elicit self-reflection
- Instruction tuning and self-consistency threshold are strongest baselines
  - but rely on a held-out set for training and hyperparameter tuning
- COOPERATE and COMPETE are state-of-the-art abstain methods
  - outperform the strongest baseline in 9 out of the 12 settings
  - avg improvement > 10% on reliable accuracy scores
  - COOPERATE works better with stronger models such as ChatGPT
  - COMPETE emphasizes reliability, highest reliable accuracy scores

# ALFA - ALignment with Fine-grained Attributes



Stella Li



Jimin Mun

2025

Li et al. Aligning LLMs to Ask Good Questions: A Case Study in Clinical Reasoning. *In Proc. COLM 2025*

# What is a good question?

- It is difficult to define a “good question”:
  - Context dependent, too many edge cases, subjective, domain specific
- Relatively easier to consider individual aspects of a good question
  - e.g. clarity, specificity
  - cognitive science, psychology, and clinical communication research have useful insights



# What is a good question?

- **clarity** -- avoids ambiguity and unnecessary complexity
- **focus** -- directly addresses a specific information gap
  - “Do you have a family history of breast cancer?” > “Has anyone in your family been sick?”
- **answerability** -- adapts to the respondent’s expertise (e.g., no jargon)
- **medical accuracy** -- aligns with established medical knowledge & guidelines
- **diagnostic relevance** -- probes for specific relevant symptoms, risk factors
- **avoiding DDX bias** -- prevents suggestive or leading wording that could introduce cognitive biases and misguide diagnostic reasoning

# MediQ-AskDocs

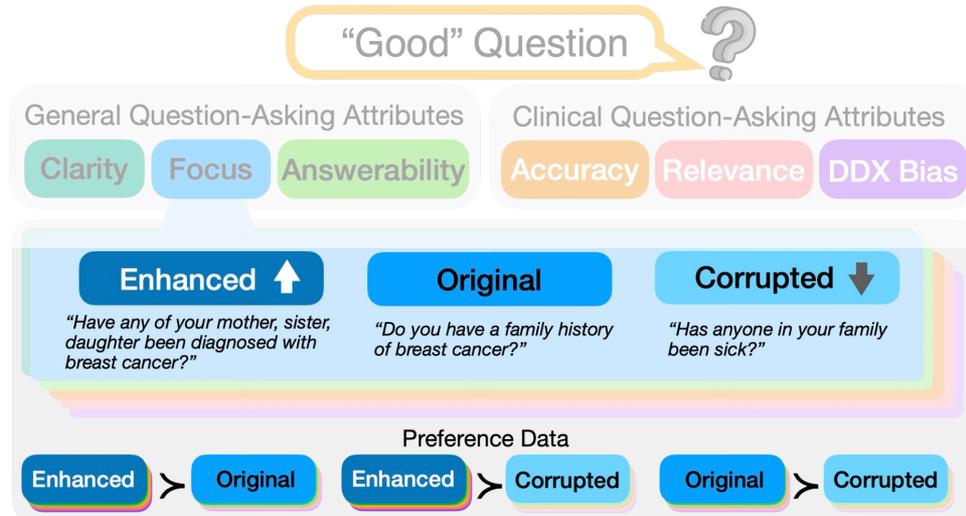
Compiled and annotated from r/AskDocs

- 17k real clinical interactions
- 80k attribute-specific question variations
- 302 expert-annotated scenarios

**First major dataset for training & evaluating medical question-asking!**

# Attribute-specific data synthesis for alignment

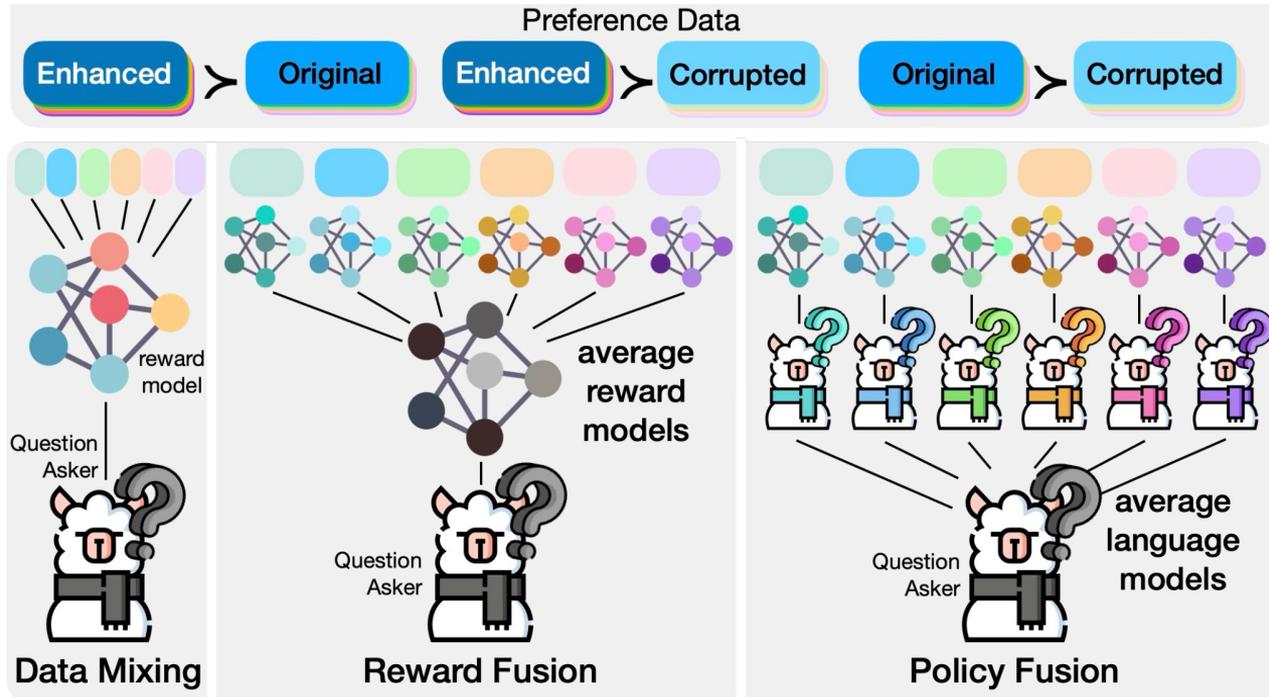
- Preference data generation recipe via **counterfactual perturbations**:
  - creating self-supervised labels for training attribute-specific reward models:  
Rewrite each original question to vary in only one attribute dimension



# ALFA Alignment



# ALFA Alignment



# Better question asking → better diagnostics

ALFA-aligned models have better question quality and better clinical decisions:

- 56.6% reduction in diagnostic errors
- 64.4% win rate in question quality

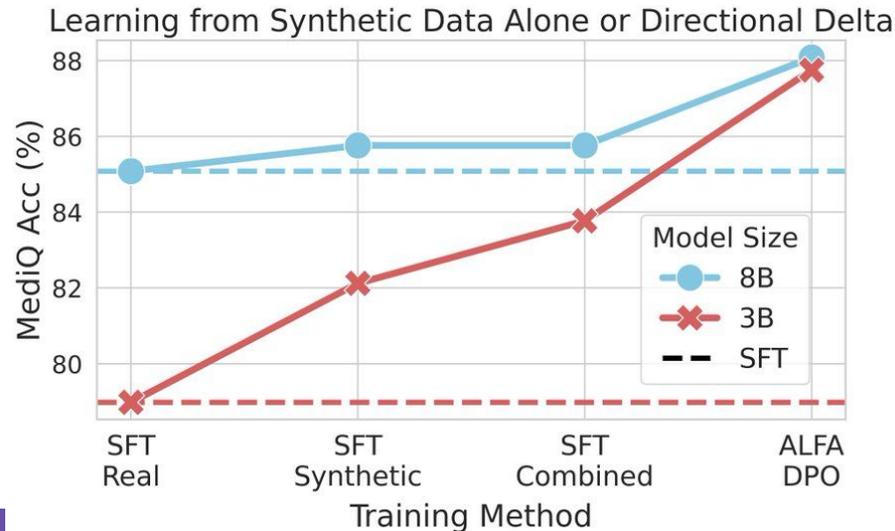
Model	Size	Win-rate	MediQ-AD
Base Model	3B	50.00	73.51
	8B	50.00	72.52
SFT	3B	61.04	78.98
	8B	58.23	85.08
ALFA-DPO	3B	<b>64.97</b>	<b>87.75</b>
	8B	<b>65.13</b>	<b>88.08</b>
ALFA-PPO	3B	64.84	86.75

Table 1: Main results. ALFA models consistently outperform base instruct models and SFT models.

# Preference learning > supervised learning

Simply showing good examples isn't enough!

- Models need to learn **\*\*directional differences\*\*** between good and bad questions.
- Fine-grained attribute outperforms coarse "good vs. bad" distinctions



# Every attribute matters!

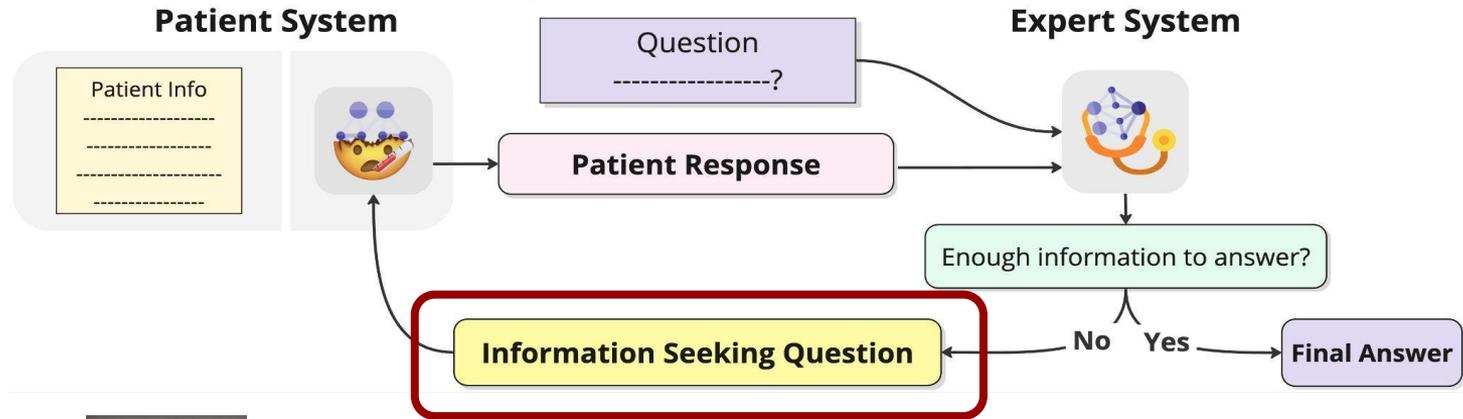
Removing any single attribute hurts performance

- Grouping general (clarify, focus, answerability) vs. clinical (medical accuracy, diagnostic relevance, avoiding DDX bias) attributes leads to drastically different outputs

<b>ALFA with General Attributes (ALFA-General):</b>
Did they treat you for mono? Is she in school?
<b>ALFA with Clinical Attributes (ALFA-Clinical):</b>
Did the pain worsen or improve with the use of NSAIDs? What do you think about the diagnosis of febrile convulsion? What is your age, sex, medical history, and medications?

Figure 5: Models aligned with general vs. clinical attributes show distinct behaviors. ALFA-General asks clear and focused, but less relevant questions and contains DDX bias ("*mono*"); ALFA-Clinical uses medical terms hindering answerability but is more professional.

# Ongoing projects



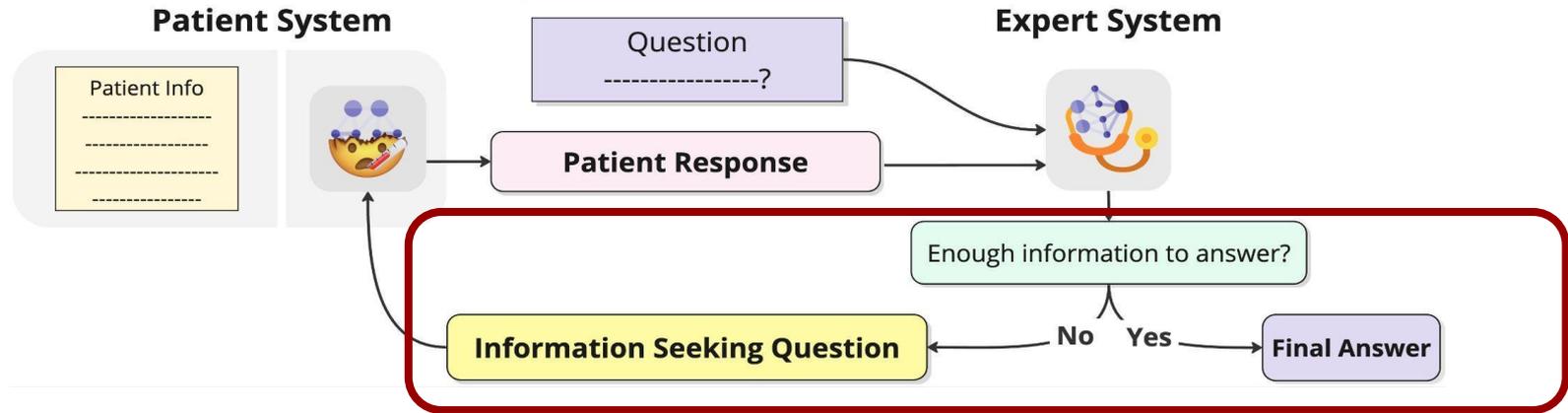
Stella Li



Melanie Sclar

- Personalization  
Li et al. PrefPalette: Personalized Preference Modeling with Latent Attributes. *In Proc. COLM 2025*

# Ongoing projects

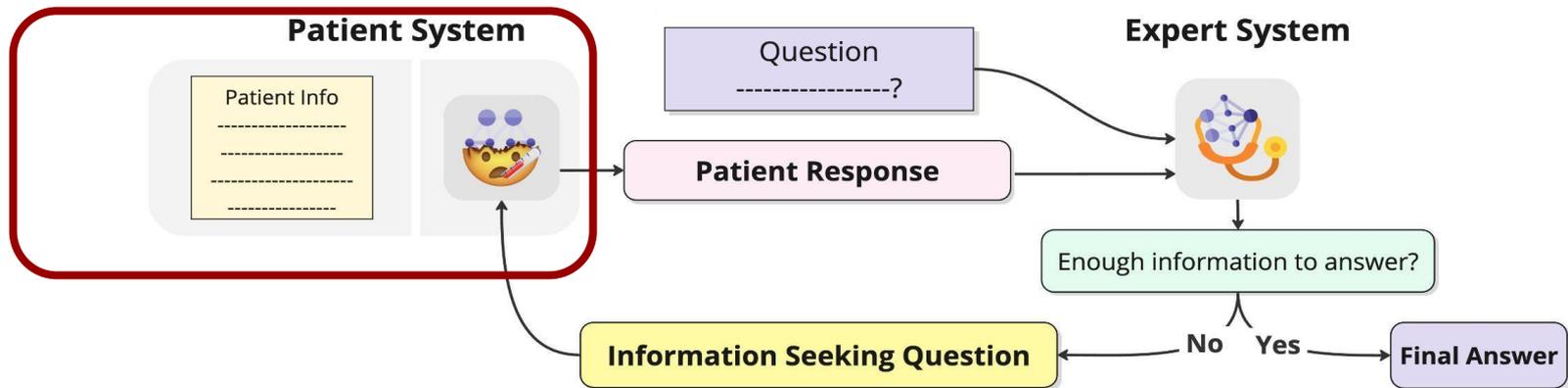


Stella Li Deniz Nazarova

2025

- Social bias in medical interactions with LLMs
- Safety in mental health-related interactions (eating disorders)

# Ongoing projects



Stella Li



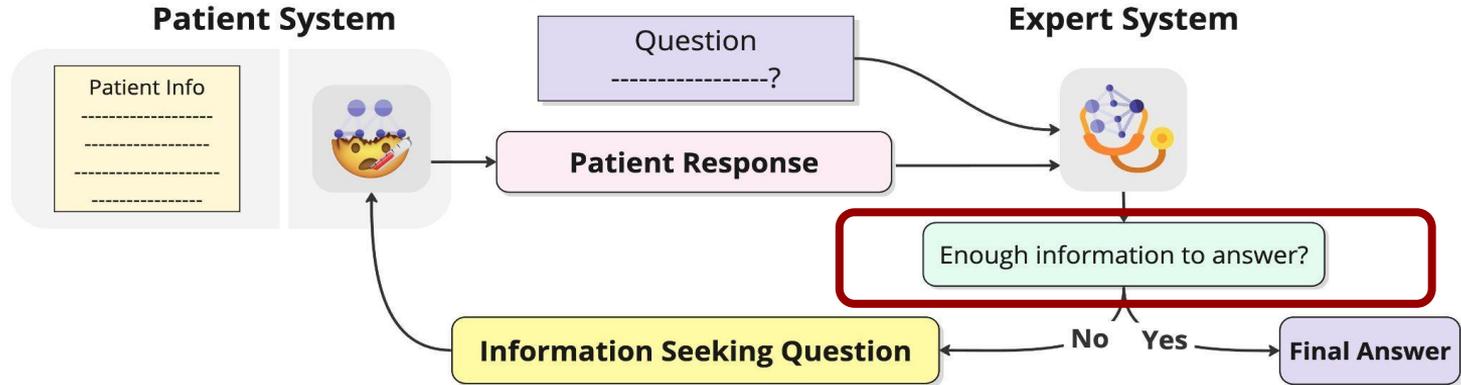
Niloofar  
Miresghallah



Hyunwoo  
Kim

- Can we anonymize patient data while preserving the utility of medical QA?  
<https://arxiv.org/abs/2504.21035>
- Synthesizing the “public” private dataset  
<https://openreview.net/pdf?id=8gluTTY12B>

# Ongoing projects



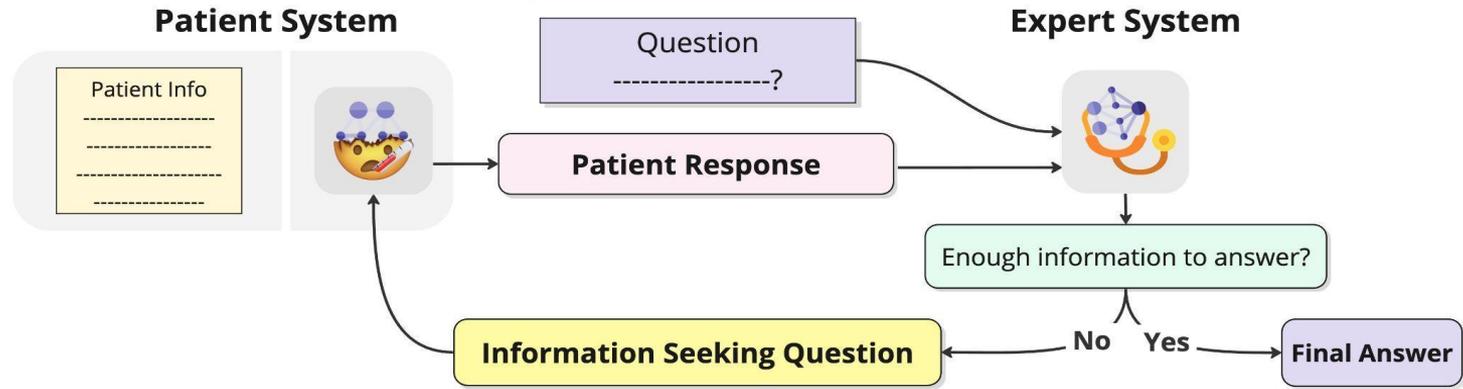
Stella Li



Rock Pang

- Controlled reasoning  
<https://arxiv.org/pdf/2506.23678>

# The MediQ benchmark



# Thank you!

<https://tsvetshop.github.io/>

