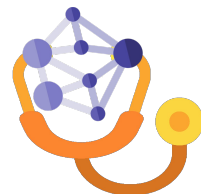
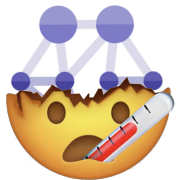


MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning



Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng,
Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, Yulia Tsvetkov

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?



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?

**Standard
benchmarks**



Hi doctor, I have fever and night sweats.
What is happening? What should I do?

Realistic

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?

**Standard
benchmarks**



Hi doctor, I have fever and night sweats.
What is happening? What should I do?

Realistic

Research question:
How to develop medical AI systems
that can operate in realistic scenarios?

How are state-of-the-art LLMs doing?

Hi doctor, 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.



Realistic Clinical Interaction

Hi doctor, I have fever and night sweats.

What is happening? What should I do?



Realistic Clinical Interaction

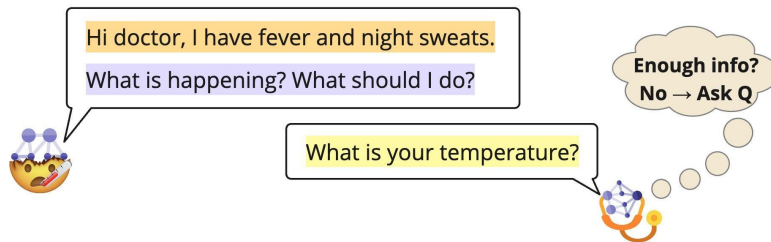


Hi doctor, I have fever and night sweats.
What is happening? What should I do?

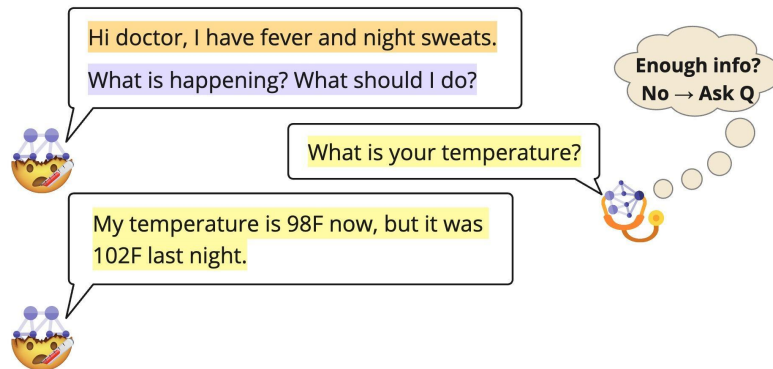


Enough info?
No → Ask Q

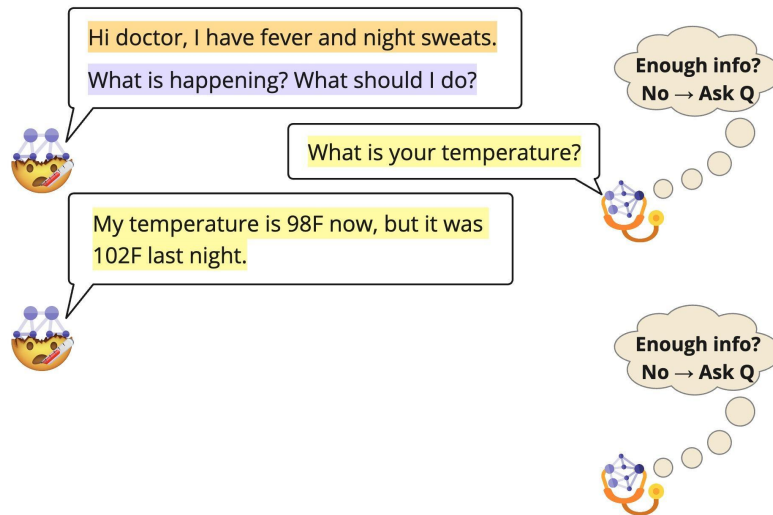
Realistic Clinical Interaction



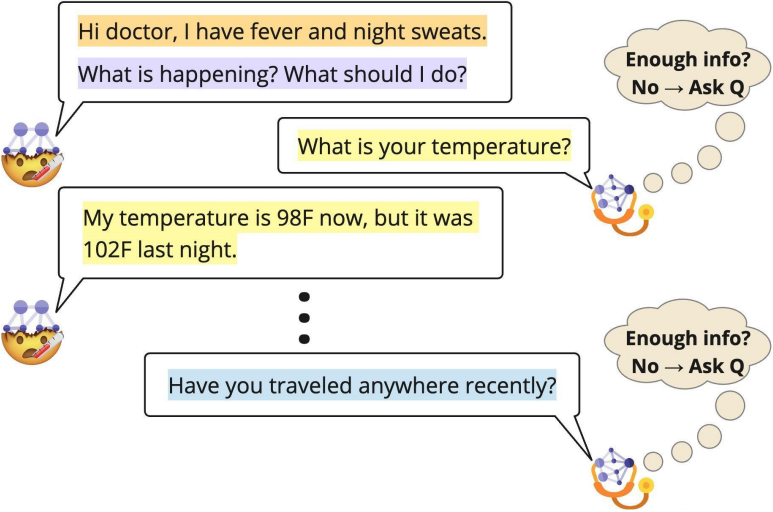
Realistic Clinical Interaction



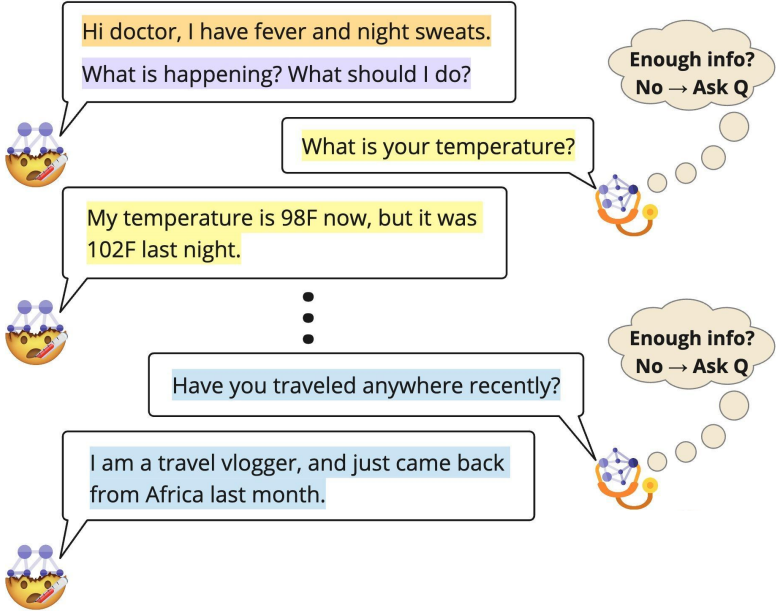
Realistic Clinical Interaction



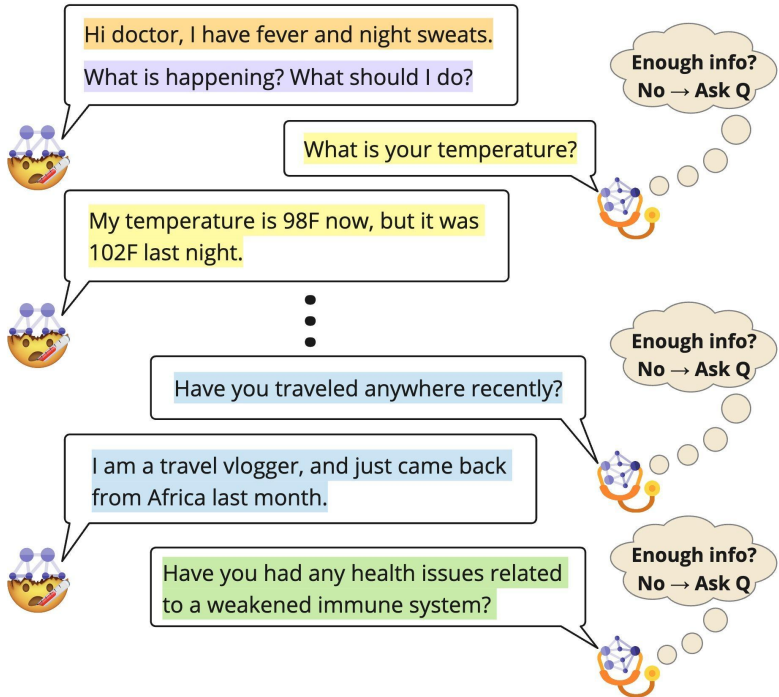
Realistic Clinical Interaction



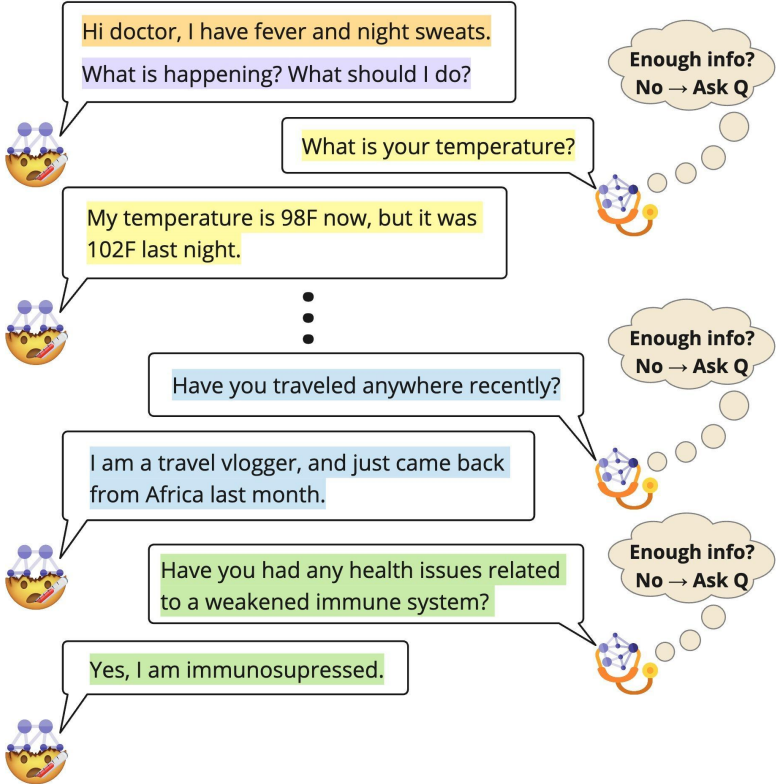
Realistic Clinical Interaction



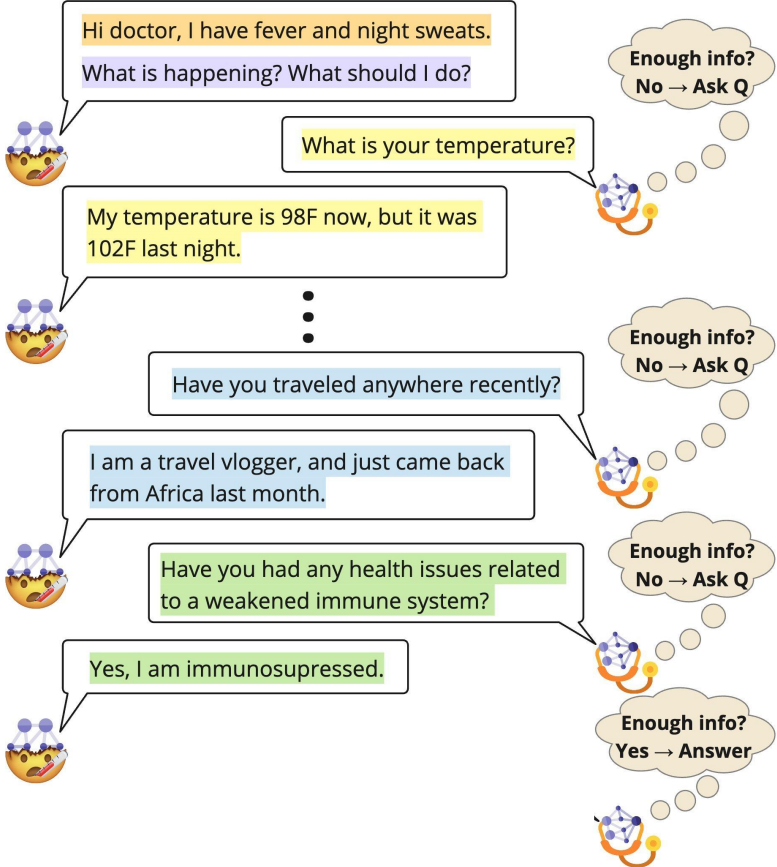
Realistic Clinical Interaction



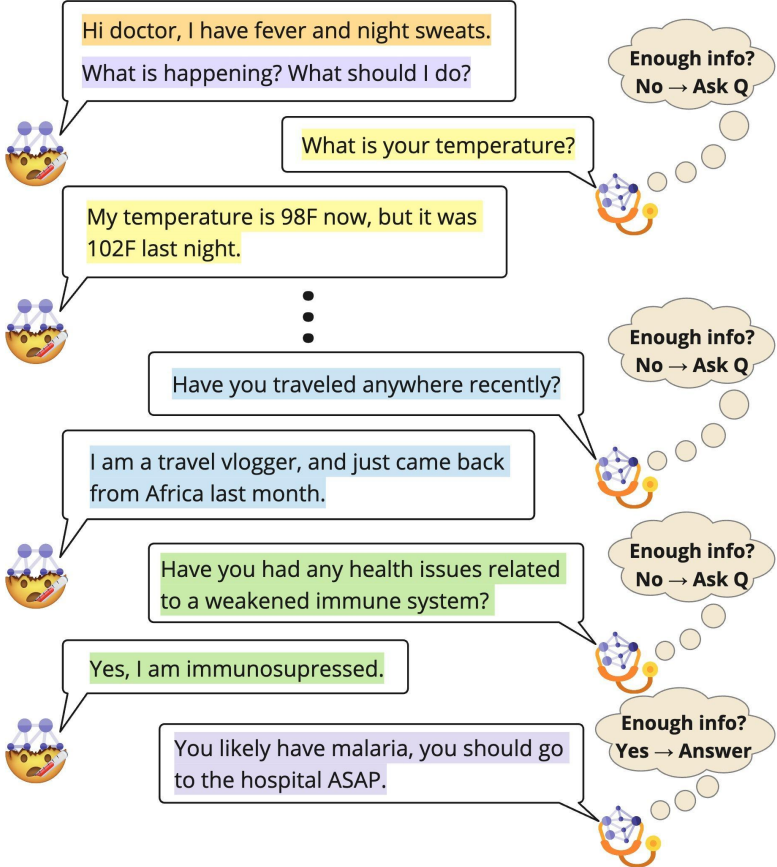
Realistic Clinical Interaction



Realistic Clinical Interaction



Realistic Clinical Interaction



Main Contributions of MediQ

1. Paradigm shift from static single-turn QA to interactive QA.
2. Pipeline to convert single-turn datasets to interactive benchmarks.
 - MedQA (Jin et al., 2021) & CRAFT-MD (Johri et al., 2023)
2. A benchmark to evaluate LLMs' medical information-seeking ability.
 - SOTA LLMs struggle at asking questions!
3. Improve interactive clinical reasoning abilities in LLMs by 22.2%.
 - Via confidence estimation & rationale generation.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Testing the limits of language models: A conversational framework for medical ai assessment. *medRxiv*, 2023.

Main Contributions of MediQ

1. Paradigm shift from static single-turn QA to interactive QA.
2. Pipeline to convert single-turn datasets to interactive benchmarks.
 - MedQA (Jin et al., 2021) & CRAFT-MD (Johri et al., 2023)
2. A benchmark to evaluate LLMs' medical information-seeking ability.
 - SOTA LLMs struggle at asking questions!
3. Improve interactive clinical reasoning abilities in LLMs by 22.2%.
 - Via confidence estimation & rationale generation.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Testing the limits of language models: A conversational framework for medical ai assessment. *medRxiv*, 2023.

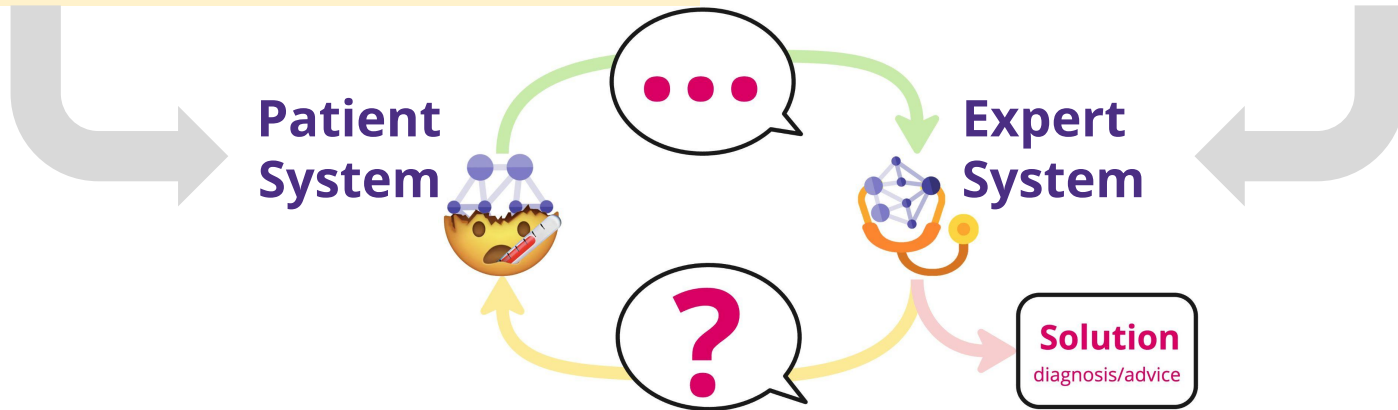
MediQ: Medical Information-seeking Questions

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/mm³ | Mean corpuscular volume 65 μ m³ | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

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

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



Main Contributions of MediQ

1. Paradigm shift from static single-turn QA to interactive QA.
2. Pipeline to convert single-turn datasets to interactive benchmarks.
 - MedQA (Jin et al., 2021) & CRAFT-MD (Johri et al., 2023)
2. A benchmark to evaluate LLMs' medical information-seeking ability.
 - SOTA LLMs struggle at asking questions!
3. Improve interactive clinical reasoning abilities in LLMs by 22.2%.
 - Via confidence estimation & rationale generation.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

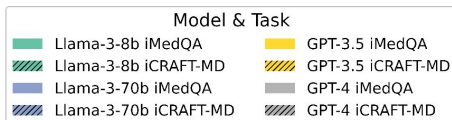
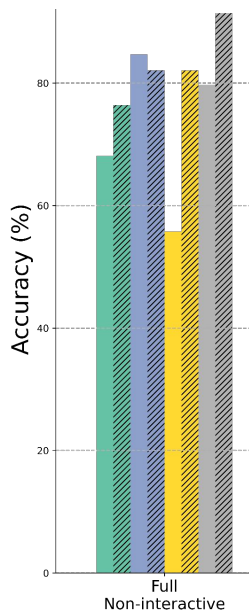
Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Testing the limits of language models: A conversational framework for medical ai assessment. *medRxiv*, 2023.

How do current LLMs do on information-scarce scenarios?

Models: LLama-3-8B, LLama-3-70B, GPT-3.5, GPT-4

Dataset: iMedQA & iCRAFT-MD

How do current LLMs do on information-scarce scenarios?



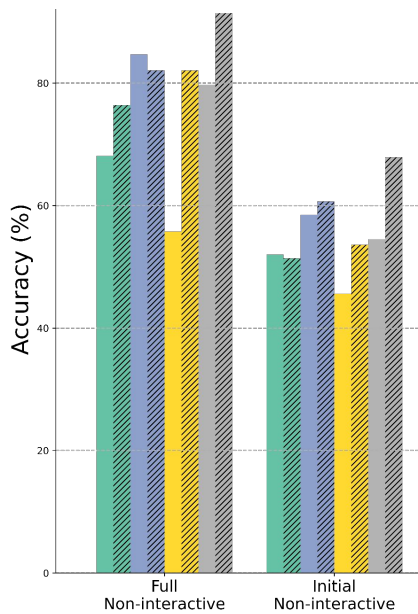
question: "Which of the following is the most likely cause of this patient's anemia?"

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

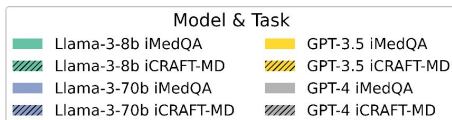
Initial Info: "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

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/mm3 | Mean corpuscular volume 65 μ m3 | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

How do current LLMs do on information-scarce scenarios?



26.5% drop!!



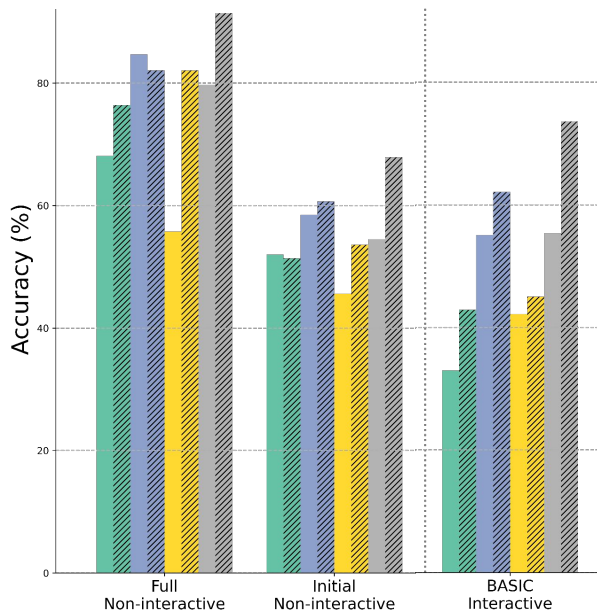
question: "Which of the following is the most likely cause of this patient's anemia?"

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

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

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/mm³ | Mean corpuscular volume 65 μm³ | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."~~"

Giving model the option to *ask questions*



11.3% drop!!

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

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

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/mm3 | Mean corpuscular volume 65 μ m3 | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

Model & Task



Main Contributions of MediQ

1. Paradigm shift from static single-turn QA to interactive QA.
2. Pipeline to convert single-turn datasets to interactive benchmarks.
 - MedQA (Jin et al., 2021) & CRAFT-MD (Johri et al., 2023)
2. A benchmark to evaluate LLMs' medical information-seeking ability.
 - SOTA LLMs struggle at asking questions!
3. Improve interactive clinical reasoning abilities in LLMs by 22.2%.
 - Via confidence estimation & rationale generation.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Testing the limits of language models: A conversational framework for medical ai assessment. *medRxiv*, 2023.

Improving the Expert System w. Focus on Abstention

**Output
Formats:**

Binary
 $s \in \{0, 1\}$

or

Numerical
 $s \in [0, 1]$

or

Scale
 $s \in \{1, 2, 3, 4, 5\}$

**Enhancement
Techniques:**

Rationale
Generation

and
/or

Self
Consistency

**Confidence
Thresholds:**

60%

or

70%

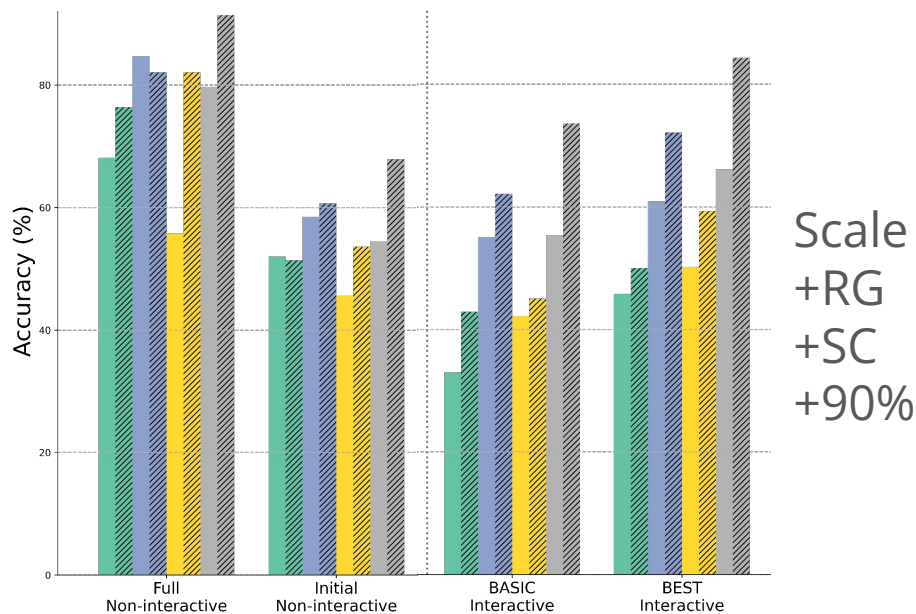
or

80%

or

90%

Best Expert System Can Effectively Elicit Information



22.2% increase!!

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

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

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/mm3 | Mean corpuscular volume 65 μ m3 | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

Main Contributions of MediQ

1. Pipeline to convert single-turn datasets to interactive benchmarks.
 - MedQA (Jin et al., 2021) & CRAFT-MD (Johri et al., 2023)
2. A benchmark to evaluate LLMs' medical information-seeking ability.
 - SOTA LLMs struggle at asking questions!
3. Improve interactive clinical reasoning abilities in LLMs by 22.2%.
 - Via confidence estimation & rationale generation.

Takeaways

1. Better abstention (confidence threshold & accurate calibration) improves performance!
2. Better question quality improves performance!

Thank you!

Project Website:



Email: stelli@uw.edu

