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

Shuyue Stella Li[W], Vidhisha Balachandran[Wi-Fi], Shangbin Feng[W], Jonathan S. Ilgen[W]
Emma Pierson[Tree], Pang Wei Koh[W][Ai2], Yulia Tsvetkov[W] stelli@cs.washington.edu

NEURAL INFORMATION PROCESSING SYSTEMS

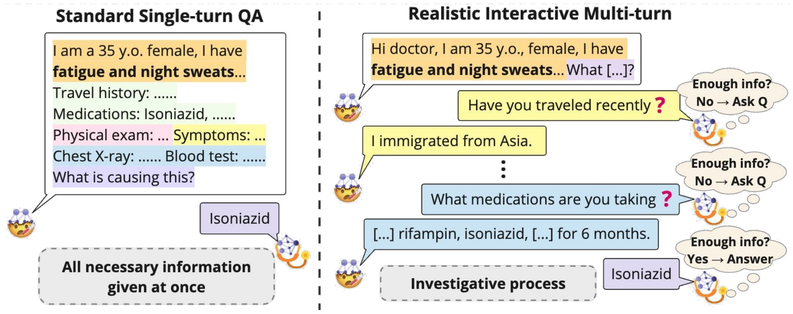
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING · UW Medicine · Carnegie Mellon University Language Technologies Institute · CORNELL TECH · Ai2 · UWNLP

TLDR

- We use LLMs interactively (e.g. ChatGPT), but most evals are static.
- Paradigm shift: LLMs should **proactively seek information** by interacting with users and **asking questions** when not confident, rather than making its general "best guess."
- SOTA LLMs are *really bad at information-seeking!* (11.3% drop)
- Better abstention decision improves performance (by 22.2%).
  - → More accurate confidence estimation
  - → More relevant follow-up questions

Novel Interactive Information-Seeking Task

The model is only provided initial information at start of the interaction.
Expected to ***reason*** and ***ask follow-up questions*** to elicit patient info.
Should decide to answer only when sufficiently confident.

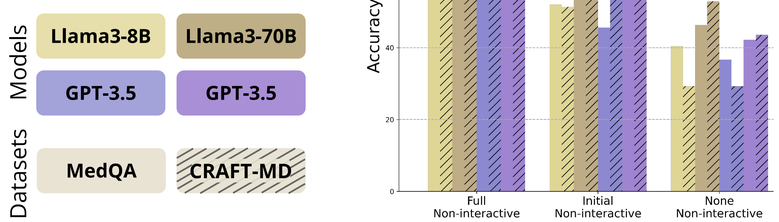


Sanity Check: Customizing to patient info is necessary

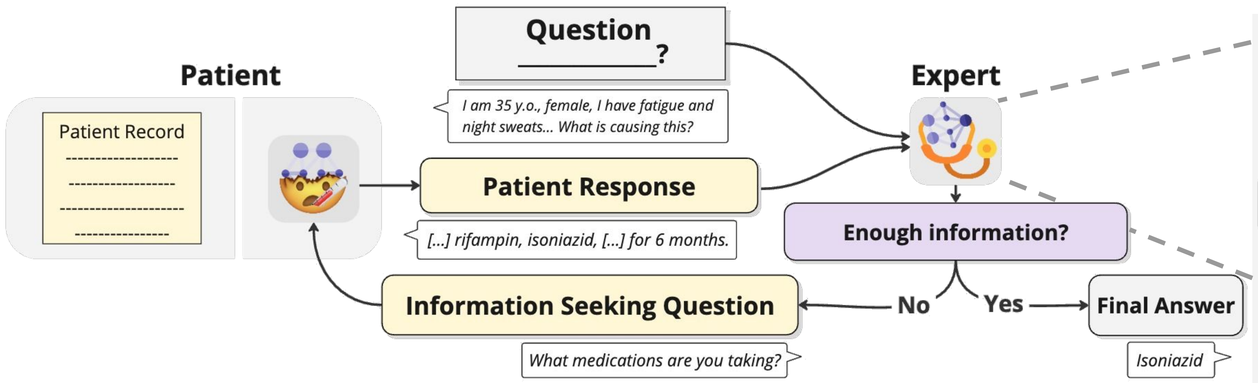
Parse patient records into question, initial info, and additional context.
Various information sufficiency levels:

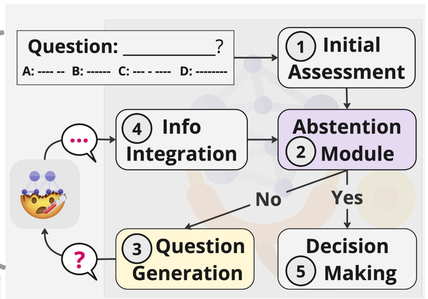
Full	Initial	None
A 40-year-old woman presents with difficulty falling asleep, diminished appetite, and tiredness. She has grown increasingly irritable and hopeless [...] diminished concentration, [...] lost 8.8 lb [...] drinks a glass of wine instead of eating [...] What is the best treatment for this patient?	A 40-year-old woman presents with difficulty falling asleep, diminished appetite, and tiredness. What is the best treatment for this patient?	What is the best treatment for this patient?
(A) Diazepam (B) Paroxetine (C) Zolpidem (D) Trazodone	Initial Info Details Question	

Performance *drops significantly* as *less information* is provided to the model in a static setting, across models and datasets.



The MediQ Interactive Information-Seeking Framework



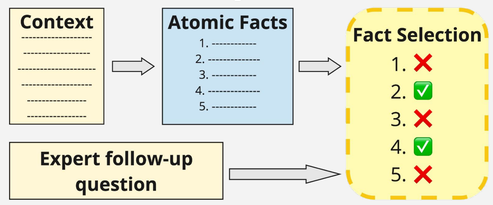


The Patient Agent:

- Uses full patient record to answer Expert questions.
- Evaluated on ***factuality*** and ***relevance***.

Patient Variant	Factuality	Relevance	Win-Rate
Direct	55.9	75.5	36.1
Instruct	62.8	78.6	37.4
Fact-Select	**89.1**	**79.9**	**63.8**

The Fact-Select Patient Agent:



The validated Patient Agent is fixed to **benchmark Expert agents**

The Expert Agent:

Cognitively inspired modular reasoning components.

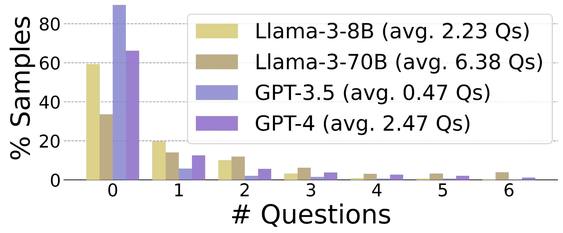
Focus on the **abstention module.**

Spoiler: accurate confidence estimation and rationale generation can significantly improve performance.

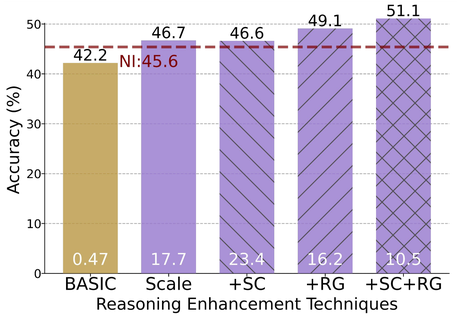
SOTA LLMs fail in interactive settings!

11.3% acc. drop from non-interactive setting!

Why? → they don't tend to ask questions.



Abstention improves Acc. 👑 BEST Expert := Scale+RG+SC → 22%⬆️

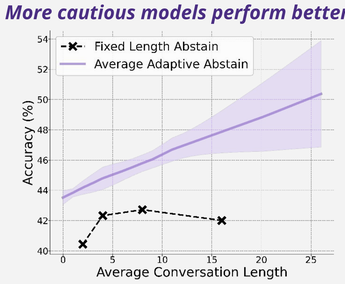


Key features of BEST Expert:

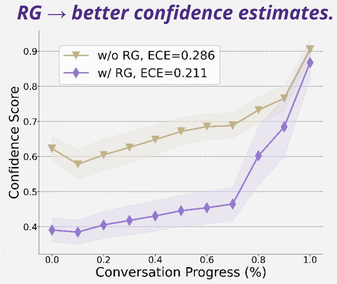
1. Confidence estimation output format.
   a. Binary vs. Numerical vs. **Likert Scale**
2. **Self-consistency** *with* **rationale generation**.
3. Confidence threshold.

Effect of confidence threshold and rationale generation (RG)

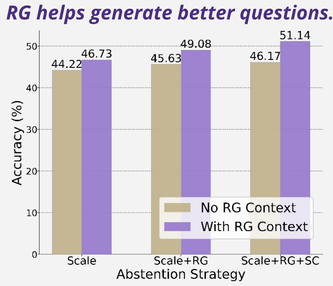
More cautious models perform better.



RG → better confidence estimates.



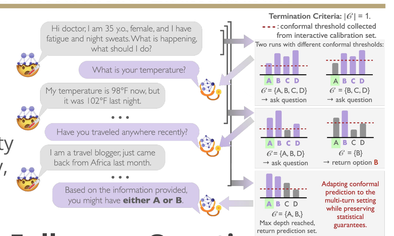
RG helps generate better questions.



Follow-up Work

Better Confidence Estimation

Conformal Prediction-inspired uncertainty quantification in interactive settings. Leverages uncertainty to improve efficiency, accuracy, and coverage.



Better Follow-up Questions

Structured decomposition of abstract goals into concrete features to generate counterfactual pairwise preference data. Combine **fine-grained rewards**.