A Human-AI Collaborative Approach to Generate Tailored Questions about Lab Test Results

Preetam Chamkura[1], Zhe He[1], Balu Bhasuran[1], Karim Hanna[2], Cindy Shavor[2], Lisbeth Garcia Arguello[2]

School of Information, Florida State University; [2]Morsani College of Medicine, University of South Florida

Introduction

Enhance patient engagement, medication adherence, and provider communication. Accurate lab results interpretation for preventing and managing chronic conditions among older adults. With wide adoption of electronic health records and patient portals, lab results are readily available but they are often difficult to interpret due to patients' limited health literacy and numeracy. In addition, lab results lack contextual explanations, leading to confusions. Previously, AI has been well explored and experimented in healthcare but limited research exists on AI-generated personalized lab-related questions.

In this study, we are experimenting with AI/large language models for generating tailored questions for patients to follow up with their doctors. We are iteratively developing LLM prompts and applying these promts on de-identified patient profiles consising of demographic information, lab test results, diagnosis, and medications. Then we are working closely with family doctors to evaluate the quality of the generated questions. By refining the prompt approach through iterative feedback, quantitative assessments, and NVIVO-based qualitative analysis, we aim to develop a strategy to generate questions improve patient comprehension, engagement, and shared decision-making, ultimately leading to better health outcomes.

Methodology

Round 1 LLM Prompt Engineering:

•Designed with three constraints:

- Rank questions by medical urgency.
- Focus on actionable patient recommendations.
- Ensure discussion within a 15-minute consultation.

Clinical Data Input:

- Used three de-identified cases from the OneFlorida Data Trust.
- Provided the LLM with patient demographics, lab results, medications, and diagnoses.

• **Example case:** 73-year-old Black male with Type 2 diabetes, hypertension, and reduced kidney function.

- Key lab values: Elevated creatinine (1.42 mg/dL), reduced eGFR (56 mL/min), A1c at 6.4%, and low albumin (18.1 ug/mL).
- Generated 40–60 questions per case.

Evaluation Process:

- Reviewed by three board-certified family physicians using:
  - Yes/No for clarity and clinical sense (**Figure 2**).
  - Likert scale ratings for relevance, significance, and willingness to answer (**Figure 3**).
- Conducted physician interviews for qualitative feedback.

NVIVO Analysis of Qualitative Feedback: Identified issues related to hallucinations, ambiguity, and areas for refinement.

Round 2 LLM Prompt EngineeringL

- Focused on 37 primary care-relevant lab tests.
- Adjusted prompts to emphasize abnormal results and proactive actions.
- Reduced question sets to 20 per case.
- Ensured questions were at a 6th-grade reading level.

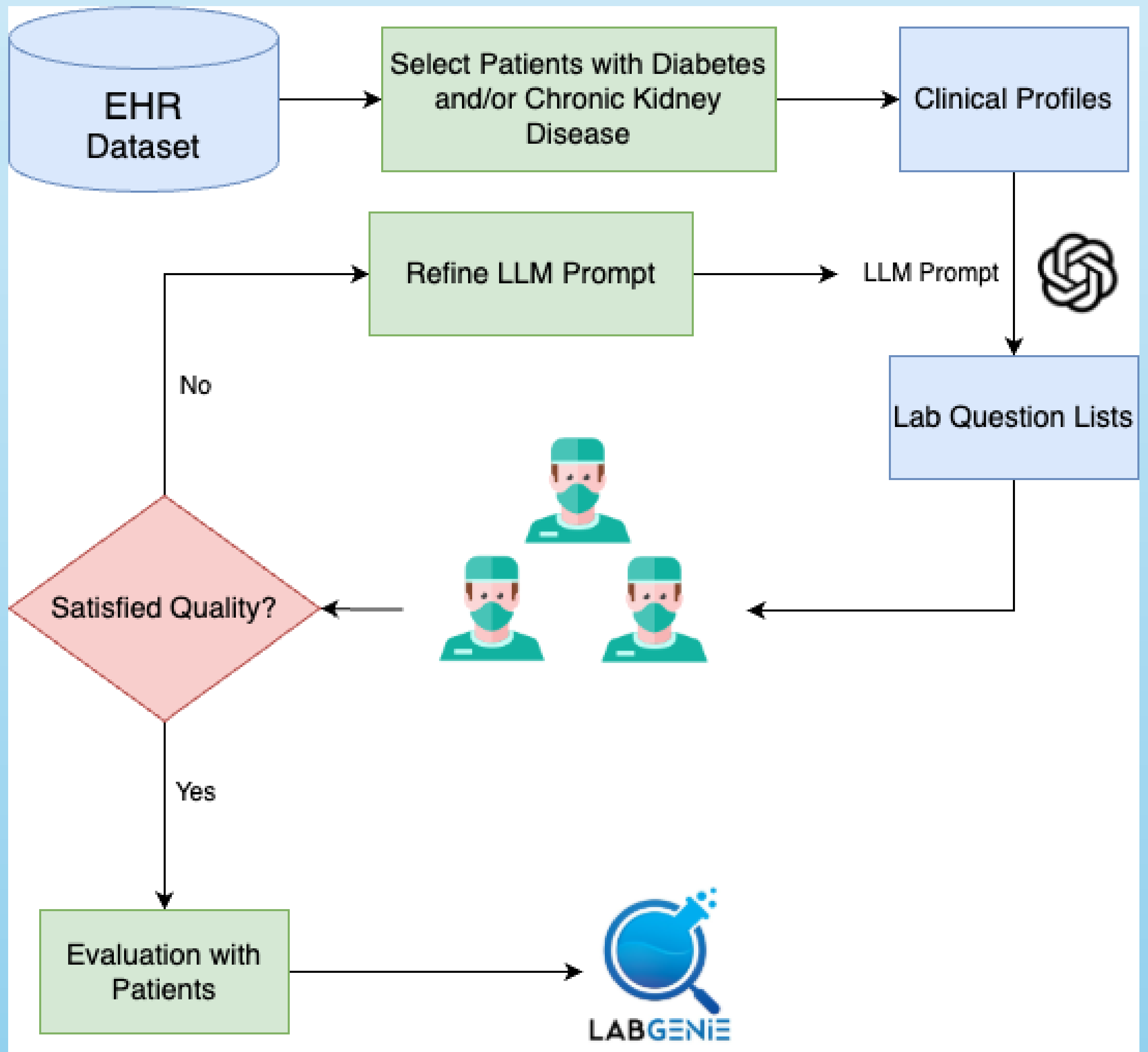

Figure 1: Study pipeline detailing the stages of AI-based question generation

Results

Sample generated questions:

"My eGFR is 56 mL/min. How can I prevent further decline, and should we review my medications?""

"Given my diabetes and hypertension, should we adjust my lisinopril dose or order additional kidney tests?"

Round 1 Findings: 96.7% of questions were clearly phrased; 89.6% made clinical sense (**Figure 2**); Questions were generally well-received (**Figure 3**).

•**Key Issues found in qualitative coding of interview transcripts:**

- Some questions contained hallucinations (e.g., incorrect medication-lab test correlations).
- Physicians would like to prioritize abnormal lab results for major chronic conditions

Round 2 Improvements:

- 100% of questions were clearly phrased and clinically sensible (**Figure 2**).
- Overall Likert scores improved (**Figure 3**).

• **Key takeaway:** Iterative refinement and clinician input enhanced question quality.

Next Steps (Round 3):

- Compare question generation across multiple LLMs (GPT-4, LLaMa 3.2-1B).
- Assess model-specific differences in generating clinically relevant and actionable questions.
- Findings will inform future AI-driven applications in healthcare.

Results

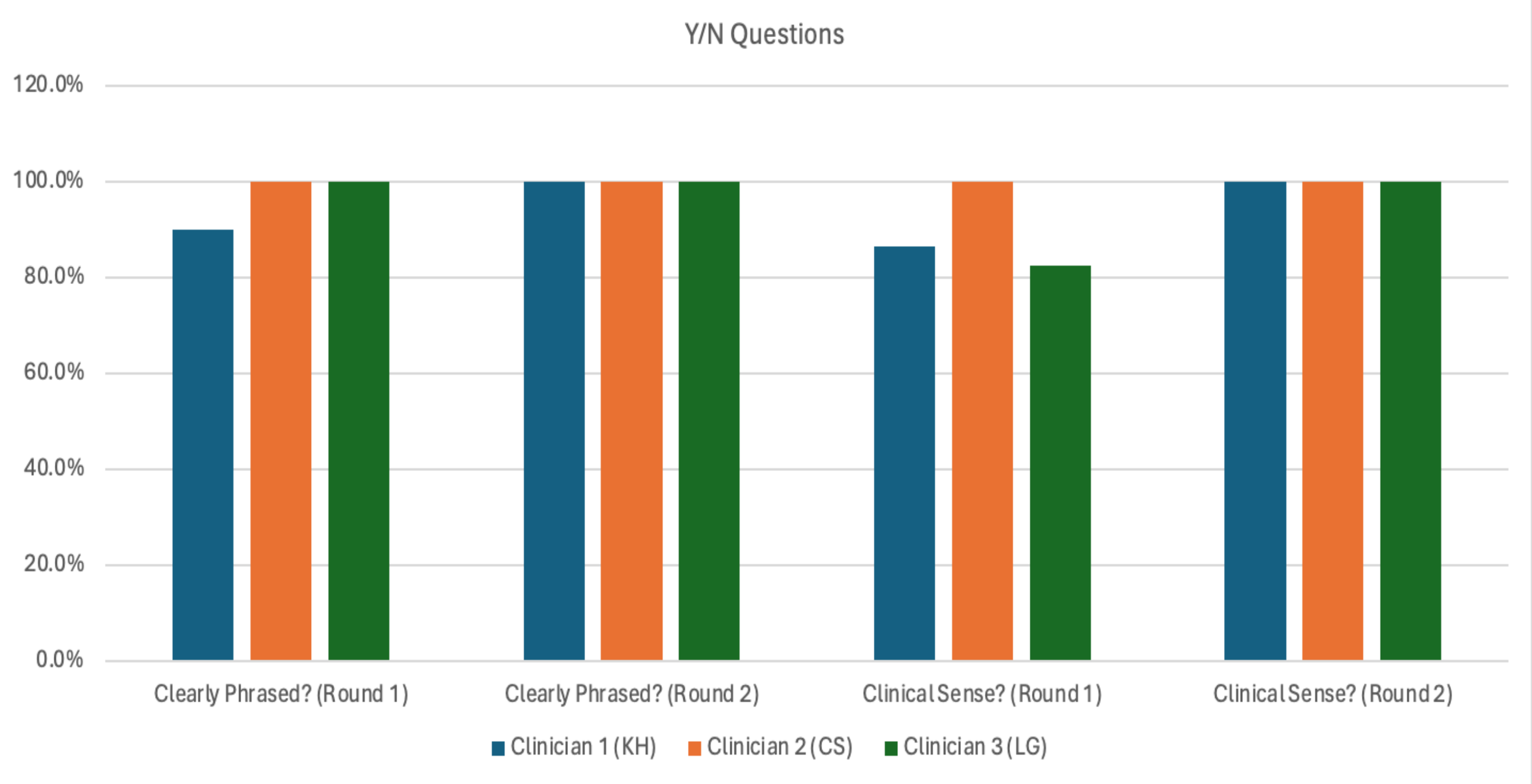

Figure 2: Y/N question answers from three family physicians regarding AI-generated questions

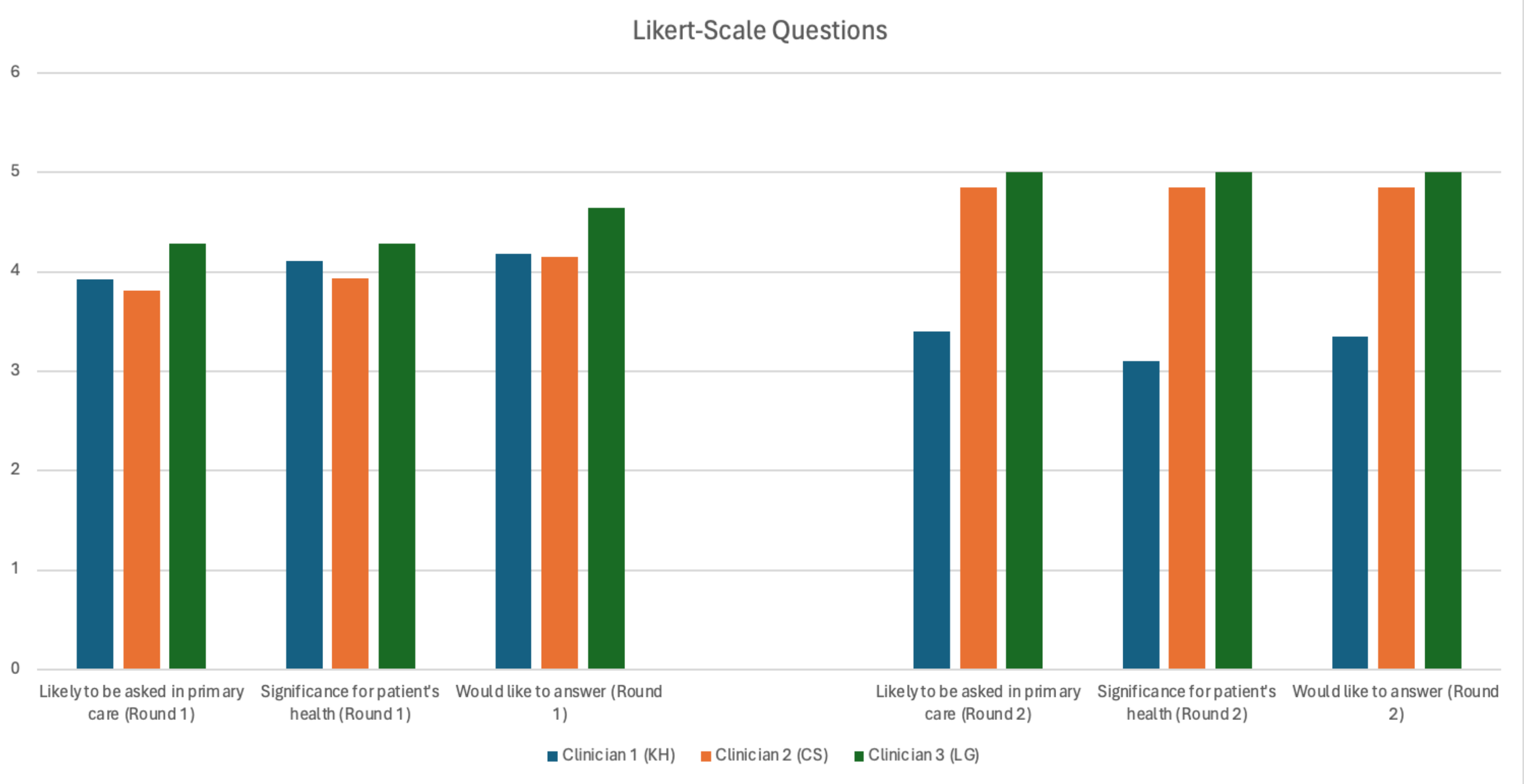

Figure 3: Likert scale evaluations from three family physicians regarding different attributes of AI-generated questions

Conclusion

Our study demonstrates that iterative refinement and clinician feedback significantly enhance the clarity, clinical relevance, and usability of LLM-generated questions, demonstrating AI's potential to improve patient-provider communication when properly fine-tuned with appropriate guidelines. However, a shortcoming of this approach is the small sample size of physicians and clinical cases limits generalizability. Future work should expand evaluation across diverse cases and medical professionals to validate these findings.

References

He Z, Bhasuran B, Jin Q, Tian S, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. Journal of medical Internet research. 2024;26:e56655.

Acknowledgements

The author expresses sincere gratitude to Dr. He and Dr. Bhasuran for the invaluable opportunity to contribute to this impactful project. The author also extends appreciation to the LabGenie team and FSU's CRE & UROP for their guidance and support. This project is funded by an AHRQ R21/R33 grant R21HS029969.