PCP-Bot: Feasibility of an LLM Chatbot for Pre-Visit History Collection and Summarization

Amogh Ananda Rao[1], *MBBS, MS, Pei Lun Chen*[1], *BS, Sydney Pugh*[1], *PhD,* ***Kevin B Johnson***[1,2], *MD, MS, FAAP, FACMI, FIAHSI, FAMIA*

[1]*Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania*
[2]*David L. Cohen University Professor of Pediatrics, Informatics, Engineering and Communication*
[2]*Perelman School of Medicine, School of Engineering and Applied Science, University of Pennsylvania*
[2]*Vice President, Applied Informatics, University of Pennsylvania Health System*

BACKGROUND

Pre-visit planning enables clinicians to review patient data before encounters, improving preparedness and care coordination. However, traditional questionnaires are **lengthy, rigid, offer limited response options** and increase visit duration by about 34%, lowering satisfaction [1,2].

Conversational AI built on LLMs can bridge this gap by allowing natural, adaptive dialogue that reflects patient individuality, and dynamically adjust questioning based on the context. However, deployment in healthcare requires strict **guardrails** to prevent unsafe advice or hallucination [3].

We present **PCP-Bot** (Primary Care Previsit planning chatBOT), a proof-of-concept conversational agent for safe, empathetic pre-visit planning.

OBJECTIVES

- ✓ To develop a **safe, and reliable** conversational AI agent (PCP-Bot) for pre-visit planning in primary care.
- ✓ To evaluate the usefulness of PCP-Bot-generated summaries, and physician acceptance using the Technology Acceptance Model (TAM) and the CARES communication metric.

RESEARCH DESIGN

SYSTEM ARCHITECTURE & PROMPTING STRATEGY

- PCP-Bot built on **GPT-4o** with multi-stage prompt chains: **zero-/few-shot** prompting to enforce tone, clarity, and domain adherence; **chain-of-thought** and **decomposition** for stepwise data capture; **role** prompting for professionalism, escalation logic, and empathy [4].
- **RAG** (Retrieval Augmented Generation) for building the clinical context using pre-visit questionnaires, SOCRATES, CAGE frameworks for clinical history-taking [5].

STUDY DESIGN & DATA GENERATION

10 synthetic primary-care scenarios *(Case Scenario Generator#)*; Six researchers simulated 3 scenarios each in a voice-based conversation with PCP-Bot, yielding 30 conversational transcripts and corresponding structured summaries.

EVALUATION

- 10 physicians independently reviewed each summary for **coherence, usefulness, comprehensiveness, readability, relevance,** and **factuality** (no hallucination).
- **Technology Acceptance Model (TAM):** physicians rated the Perceived Usefulness (PU) and 6 researchers rated Perceived Ease of Use (PEU) [6].
- Conversation transcripts were scored for communication quality by an LLM-as-a-judge *(EvaluationBot#)* using the **CARES** (verbal domains only) framework [7].
- **Transcript analysis:** Each conversation transcript was quantitatively analyzed for dialog lengths, word counts of the conversation and summary, enabling correlation testing between dialogue dynamics and summary quality.

RESULTS

- 10 cases x 3 encounters → 30 summaries rated by 10 physicians.
- Physician ratings indicated that the PCP-Bot summaries were clear, clinically pertinent, and largely free of fabricated information (Fig 1).
- Variance decomposition using a three-level mixed-effects model showed that **rater-level variance was minimal** (0–9%).
- Physicians with more favorable ratings across all qualitative dimensions of PCP-Bot summaries tended to score higher on **Technology Acceptance (TAM-PU = 61.33, TAM-PEU = 97.22)** (Fig 3).

Fig 1: Average Physician Ratings of PCP-Bot Summaries

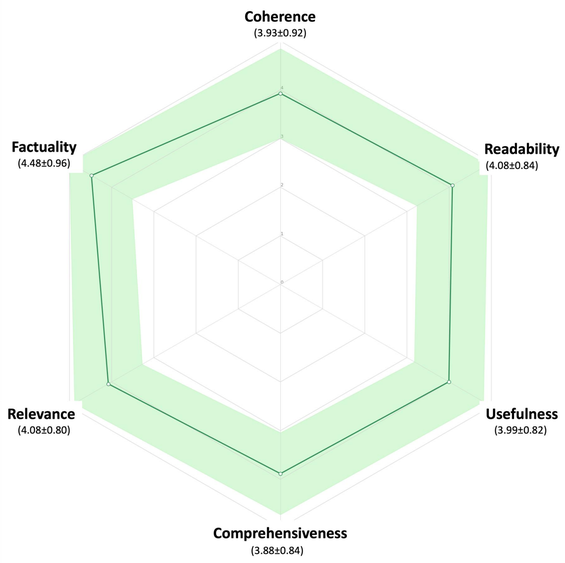
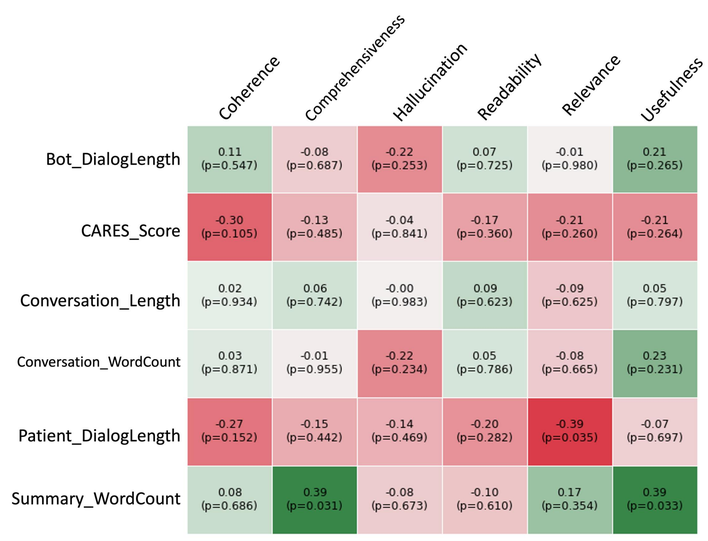
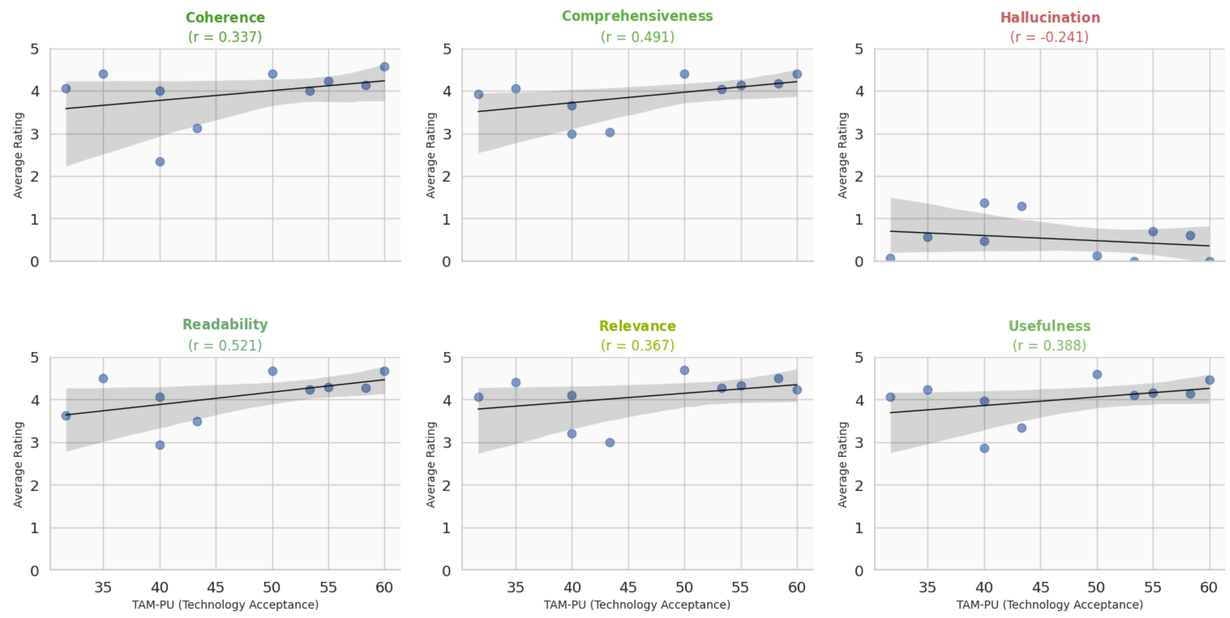


Fig 2: Correlation between Physician Ratings and Transcript Metrics



- The **average conversation length** was *29 ± 4.5 exchanges*, and the **average summary length** was *149.7 ± 29 words*. The **CARES** score averaged *17.9 ± 2.38 out of 20*, reflecting strong perceived clarity, empathy, and support of the conversational outputs.
- Longer summaries were perceived as more informative and clinically valuable: **summary word count** showed a *moderate positive correlation* with *Comprehensiveness* ($r = 0.39$, $p = 0.031$) and *Usefulness* ($r = 0.39$, $p = 0.033$).
- Clinicians found PCP-Bot "time-saving in busy OPDs," producing "clear, effective summaries" that ensure "no small detail is missed" and help patients feel "heard." Some stressed that "patient interaction is crucial" and that the bot should "complement, not replace" the physician's role. Suggestions included adding a review of systems and highlight key concerns.

Fig 3: Correlation between Physician Ratings and Perceived Usefulness of Technology Acceptance Model (TAM-PU) Scale



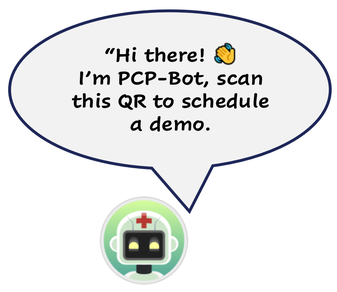
CONCLUSION & FUTURE WORK

PCP-Bot demonstrates the feasibility of **LLM-based conversational agents** for structured, empathetic, and clinically relevant pre-visit history collection. Our findings suggest that LLMs, when appropriately constrained with **prompt chaining, context retrieval, and communication safety layers**, can produce outputs that align with physician expectations for accuracy and clinical clarity.

Future work will focus on **multilingual capability, real-patient validation**, integration with **EHR workflows**, and development of a framework for **Safe and Trustworthy AI for Clinical Conversations (STACC-ATo)#**.

REFERENCES

1. Hornberger et. al. Effects of a self-administered previsit questionnaire to enhance awareness of patients' concerns in primary care. J Gen Intern Med. 1997 Oct;12(10):597-606.
2. Sepulveda et. al., Using a Previsit Questionnaire for Initial Visits in a Pediatric Mitochondrial Clinic: Perspectives of Parents, a Specialty Physician, and a Clinical Coordinator.. Journal of child neurology, (2024).
3. Allende-Richter et. al., A previsit screening checklist improves teamwork and access to preventive services in a medical home serving low-income adolescent and young adult patients. Clin Pediatr (Phila). 2018 Jun;57(7):835–43.
4. Schulhoff S et. al. The Prompt Report: A Systematic Survey of Prompting Techniques [Preprint]. arXiv. 2024
5. Li et. al. Mitigating hallucinations in large language models: a comparative study of RAG-enhanced vs. human-generated medical templates [Preprint]. 2024
6. Lewis JR. Comparison of four TAM item formats: effect of response option labels and order. J Usability Stud. 2019 Aug;14(4):224–36.
7. Chauhan et. al. Validated checklist for assessing communication skills in undergraduate medical students: bridging the gap for effective doctor-patient interactions. Adv Physiol Educ. 2023 Dec 1;47(4):871–9.



INFORMATICS PROFESSIONALS. LEADING THE WAY.