

Authors: Cheng Kai-Kao, MD., Heather Himelhoch, PhD, MPH, Jincy Oommen

Problem

- UCM in partnership with Inference Analytics developed ChatUCM, an AI-powered virtual assistant, designed to help hospital staff and administrators obtain answers to questions quickly and efficiently. Unlike external Chatbots, (e.g. ChatGPT, Google Gemini), this platform complies with internal data and security policies ensuring sensitive information (including PHI) remains protected.
- During the development of the tool, it became clear that there was not a standard way (neither internally or externally) to evaluate the tool's readiness for deployment to end-users given the nascency of AI tools

Goal

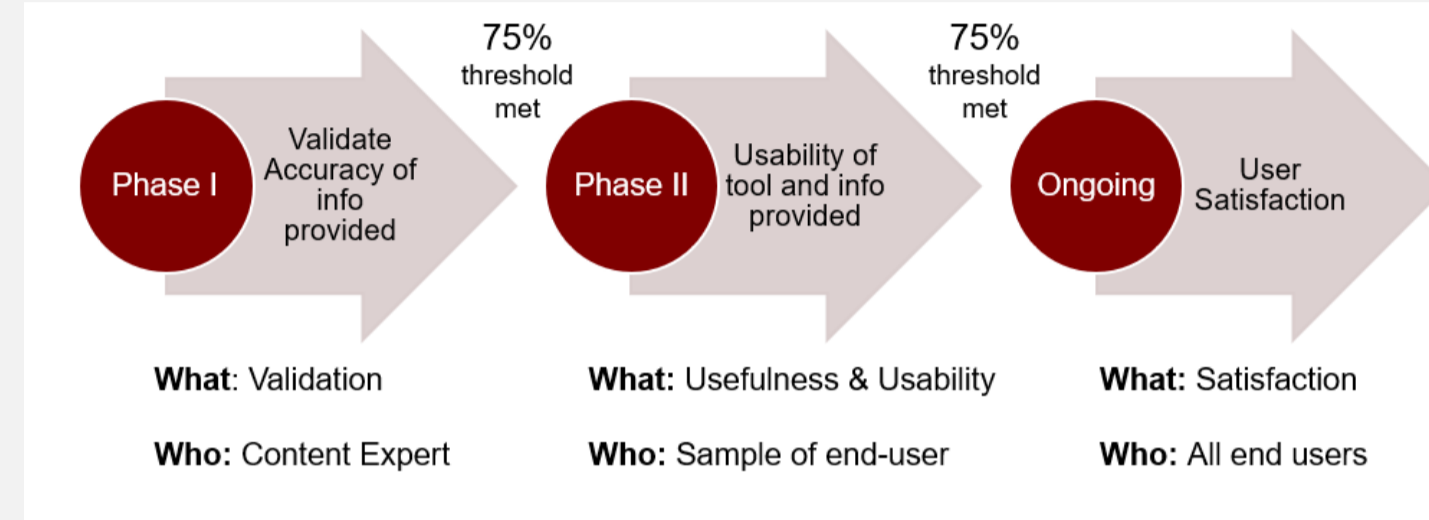
- Create an evaluation framework that will:
 - Ensure the tool is providing accurate information, particularly when being trained on internal data
 - Promote usability of the tool
 - Continuously iterate incorporating user feedback

Strategy

- Root in Published Science:** Ensure that our evaluation framework is rooted in established methods
 - Identification of Scientific Frameworks
 - Rooted in "Core Principles for Trustworthy Health AI" [Coalition for Health AI \(CHAI\) Assurance Standards Guide. 2024](#)
 - Usefulness, Usability, and Efficacy
 - Fairness and Equity
 - Safety and Reliability
 - Transparency, Intelligibility, and Accountability
 - Security and Privacy
 - Influenced by the Technology Acceptance Model (TAM)
 - Validated Tools and Methods
 - Adapted System Usability Scale (SUS)
 - Usability testing rule of thumb: 5 ppl reveal 80%-85% of issues (qualitative data)
 - Threshold to pass considerations:
 - 80% generally considered rule of thumb (survey research, usability testing).
 - However, with very small sample sizes 70% - 75% is reasonable.
- Adapt to be Pragmatic:** Adapt established methods to ensure our evaluation framework is quick, replicable, and rigorous without adding unnecessary barriers
 - Application of Scientific Frameworks
 - Assumptions:
 - Safety, Reliability, Security & Privacy are addressed through AI governance
 - Transparency, Intelligibility and Accountability are addressed through Technical Review.
 - Our approach focuses on remaining core principles:
 - Usefulness, Usability, and Efficacy
 - Post-launch monitoring: Fairness and Equity
 - Adaptation of Tools and Methods
 - Threshold:: 75% validation is a balance between consistency in achieving 'good' results and acknowledgement of very low sample size – a balance necessary to accelerate to production while ensuring accuracy
 - # Users: Request testing by at least 2 users per 'question' to get inter-rater reliability but reduce burden
 - Administration: Smartsheet allows process to be easily replicable for new modules and accessible to teams

Strategy Continued

- Articulate a Clearly Understandable Process:** Ensure that the process is simple and clear



Results

Overall Usage:

- Over 750 unique users and > 4000 queries since June 2025 launch across all modes of ChatUCM
- 83% of feedback provided through thumbs up/thumbs down function is positive¹

Infection Control Policy Mode Evaluation Results:

The following are the results of ChatUCM's performance with Infection Control policy related content. Based on the following results, ChatUCM's Infection Control Mode was deployed in June 2025.

Phase I: Validation of Accuracy

The chatbot understood the intent of the question	99%
The response helped me achieve my intended goal	94%
The chatbot response was complete	93%
The chatbot is easy to use	92%
I am confident the response I was provided is accurate	89%
Total	93%



Phase II: Usefulness and Usability

The chatbot provided output that matched my intent	94%
The response helped me achieve my intended goal	86%
The response fully answered my question	91%
The chatbot is easy to use	95%
I trust the output I was provide is accurate.	79%
Total	89%



Phase III: User Satisfaction (Ongoing)

- Data Collection Inside ChatUCM
 - Users can provide good (thumbs up) or bad (thumbs down) rating
 - Comments section: Allows for ongoing qualitative data collection



Next Steps

- Based on the success of the evaluation criteria to help determine readiness of deployment, we will continue to use this approach in the following scenarios
 - Development of a new module or use case
 - To date: Infection Control policies, Insurance Denial Management, General chatbot
 - Significant update
 - Update of underlying GPT module

Acknowledgements

- We would like to acknowledge the contributions of the following team and individuals who helped iterate and test the evaluation framework: Infection Prevention and Control (Vera Chu, Molly Steele, Palak Patel, MD, David Zhang, MD), Patient Financial Services (Anthony Gibbs, Lori Weber, Shjavon Griffin), Mark Connolly. Whitney Westphal, William Moser, MD, as well as our IT partners, Shariq Ata and Vatsal Patel