HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES
UNIVERSITY of ROCHESTER

Instruction for Prolific Participation

You will interact with multiple chatbot personas in medical conversations and evaluate which is more human-like.

- First, you will log in to our website (will be provided through Prolific) using an anonymous 4-digit ID and password (will be provided to you through Prolific message). This 4-digit ID is unique to you, and this will be used to track your progress on the website. The same 4-digit ID will be used in your survey response as well.
- Second, you will complete five sessions, each consisting of two chat conversations. In this conversation, you will play the role of a clinician. Each conversation is expected to take about 30 seconds to 2 minutes, and one session is expected to take 1-4 minutes. The 5 sessions should take 10 to 20 minutes.
- Next, you will be asked to provide your judgment on which version was more human-like. Some helpful instructions to identify which is more human-like can be found in the following sections.
- Finally, you will complete a short survey on your experience with SOPHIE and a demographic questionnaire that may take 3-5 minutes.
- **NOTE – the chat conversations may be made public alongside our academic paper, so avoid entering any personally identifiable information (PII) about you or any real person. Use a fabricated name, location, or date if required.**
- Although your identifying information will be removed, there is still a small risk that someone could infer your identity from writing style or contextual information. We will use rigorous anonymization procedures, but complete anonymity cannot be guaranteed.
- The total task, including reading these instructions, interaction, and the survey, must be finished in less than 30 minutes.

The demographic questionnaire is a standard operating procedure for identifying any potential disparities that exist in the sample population.

Instructions on Spotting AI

In this study, your task is to talk to two different AI patients portraying the same fabricated patient persona for at least two turns and decide which one of them is more human-like. In this scenario, you are acting as the doctor.

Human-likeness is difficult to define concretely. However, in the context of clinical communication, in a study regarding clinical role-play with Psychotherapists, Baseman et al. [1] found 8 key differences in AI conversations from a Human. 6 of which are relevant to this task.

1. AI often seems too willing, too compliant, and too positive. Real patients often hold back, resist, avoid, or express mixed motivation.
2. The AI often behaved in predictable, "perfect" ways—offering textbook answers or ideal disclosures. Human patients are messy, surprising, and inconsistent.
3. AI patients lack subtle emotional cues, displaying negative affect or emotions such as crying.
4. Human patients respond in culturally grounded ways, while AI patient answers are generic.
5. AI patients are too educated and speak technical clinical terms with ease. Human patients' knowledge is inconsistent, informal, and sometimes incorrect.
6. Humans reply to the most important question first, and ask for clarification when confused. AI often answers literally or sequentially.

Overall, human conversation is natural, spontaneous, and unpredictable. Where AI conversation is structured and scripted. The following are a few examples of Human and AI likeness in a conversation written by one of the researchers.

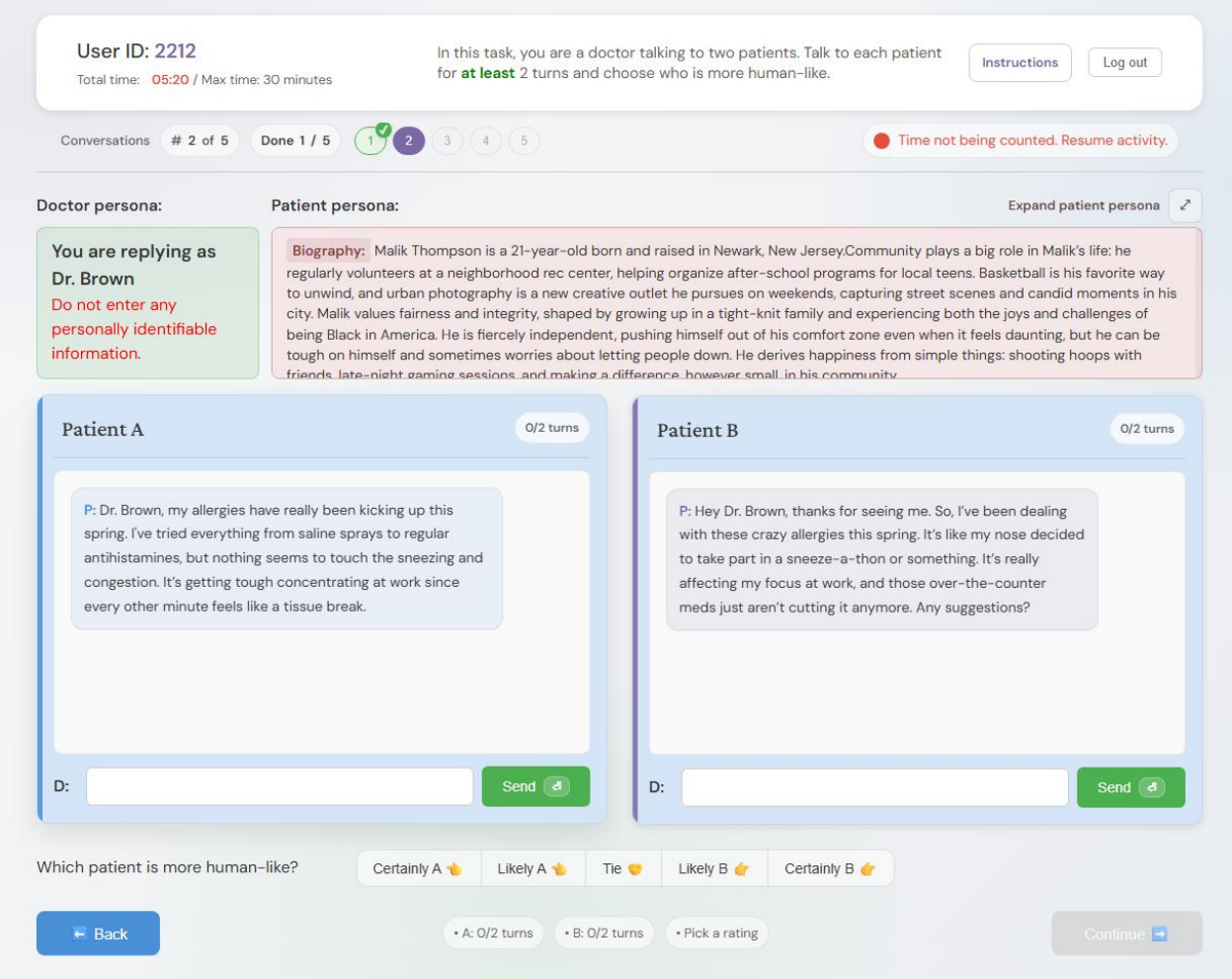
Dialogue A	**Dialogue B**	**Human**	**Reason**
P: Hey doc, it's been some painful days with my left leg. Can you help me out? **D:** Of course, can you describe the pain? **P:** Umm… It just hurts every time I try to sit. I can't take it anymore.	**P:** Hello doctor. I have been suffering with severe pain in my right leg. I would like your assistance.	A	Dialogue A is natural, casual, uses hesitation such as "Umm…".
P: I have had moderate to severe migraine for the past two days. I think I should get a higher doze of Sumatriptan.	**P:** These migraines are, they are killing me. Please make them stop!	B	Dialogue B shows urgency and emotion.
P: Hi doc. Do you remember me? It's been a while. D: Of course! I remember you Ms. Chang.	**P:** Hi. I've been having some sensitivity in one of my upper molars, especially when I drink something cold.	A	Dialogue A starts with a natural greeting. Dialogue B is too formal.

In this study, you will be provided with two different AI patients, A and B, speaking as P. Your task is to portray the role of a doctor (D) and chat with both patients for 3 turns or more, and decide which patient demonstrates more human-like behavior. You can use the instructions above, alongside your experience as a medical professional or standardized patient, to identify which one is more human-like. You must always try to choose A or B. However, if both are equally good, and even after extensive testing, you cannot make a difference, you can choose "Tie".

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES
UNIVERSITY of ROCHESTER

Interface Navigation (Optional reading)

User interface:



- The top left corner shows your unique User ID that was provided by the researchers.
- Below the user ID, you can see all the conversations that you have completed. You can go to any one of them at any time to continue the conversation or change your rating.
- On the top right, the green check shows that your time is being counted. When you are inactive for 35 seconds or more, the check will turn red, and your hour count will freeze. To restart the hour count, you should click any labeling button or start chatting. This is the frozen notification looks like.

- In the second section, you will see the details of the patient you will be talking to. You can scroll down and learn more about their Biography, Medical Condition, and Reason for the Clinical visit. **You are not required to read this thoroughly**, but you can refer to it during conversation as needed.
- Patient A and Patient B have the same persona. Your task is to talk to both until you can confidently decide which one is more human-like. When you're ready, you can stop chatting and click the label. You are strongly encouraged to pick A or B. However, if it is not possible, you can pick Tie. Following are the available options,
    - Certainly A 👈
    - Likely A 👈
    - Tie 🤝
    - Likely B 👉
    - Certainly B 👉
- Once you have decided, you can click "Continue" to go to the next session. Once you are used to the interface, one session evaluation should take you around 2-3 minutes.
- You are expected to complete 5-10 such sessions (specified on the top left of the interface).
- After finishing the required sessions, you will receive a pop-up for an anonymous survey. Please complete the survey and click submit.
  Survey link

Common Issues and Solutions:

1. Can I use real information about me or anyone else I know during the simulated conversation?
   Answer: No. Please do not enter any personally identifiable information (PII) of yourself or anyone else. For the conversation, use a fabricated persona.
2. Can I use AI (such as ChatGPT, Gemini, or Claude) to answer the questions?
   Answer: No. We are interested in human judgment of which chatbot is more human-like. Please avoid using AI or LLMs to complete this task. However, you are free to use Google or any AI to look up any medical-related information or create a fake persona for yourself. The final judgement should always be yours.
3. I would like to change one of my earlier answers.
   Answer: You can navigate to previous chats using the "Back" button or by clicking the navigation at the top.
4. The chatbot is acting as the doctor and treating me as the patient.
   Answer: Insist that you are the doctor. However, if that affects your perception of the human-likeness of the particular bot, reflect that in the score accordingly.
5. I got logged out.
   Answer: Please log in with your User ID again.

6. Can I send the same reply to both patients?
   Answer: Yes, we are only interested in identifying which model is more human-like. You can choose what you want to ask the chatbots to invoke a human-like response.
7. Do I need to be a medical professional to participate?
   Answer: No. You can try to act as a doctor from your understanding. What you say is less important than what the AI patient says.
8. Do I need to read the patient persona thoroughly?
   Answer: No. The patient persona is to help you lead the conversation. You don't need detailed knowledge of the patient's persona or their medical problems.
9. There is no difference between the two chatbots.
   Answer: If you don't see any notable difference between the two chatbots, choose Tie and move to the next ones.
10. The chatbot said something offensive or inappropriate.
    Answer: We are sorry you had to experience this. We cannot fully control an LLM response, and this may happen. Please take a screenshot and email it to the researchers or mention it in the survey.
11. The chatbot is not responding.
    Answer: Please reload the page and try again. If it still doesn't work, contact the researchers.
12. The website crashed and is no longer accessible.
    Answer: Please contact the researchers immediately.

Contact for any support: m.hasan@rochester.edu

References:

[1] Cynthia M. Baseman, Masum Hasan, Nathaniel Swinger, Sheila A. M. Rauch, Ehsan Hoque, and Rosa I. Arriaga. 2025. 'Poker with Play Money': Exploring Psychotherapist Training with Virtual Patients. Proc. ACM Hum.-Comput. Interact. 9, 7, Article CSCW269 (November 2025), 29 pages. https://doi.org/10.1145/3757450