Test plan

Contents

> *"Chatbots are important because you won't feel stupid asking important questions. Sometimes talking to someone can be a bit intimidating. Talking to a chatbot makes that a lot easier!"*
> — Petter Bae Brandtzaeg, project leader, Social Health Bots project

1 Evaluation Objectives

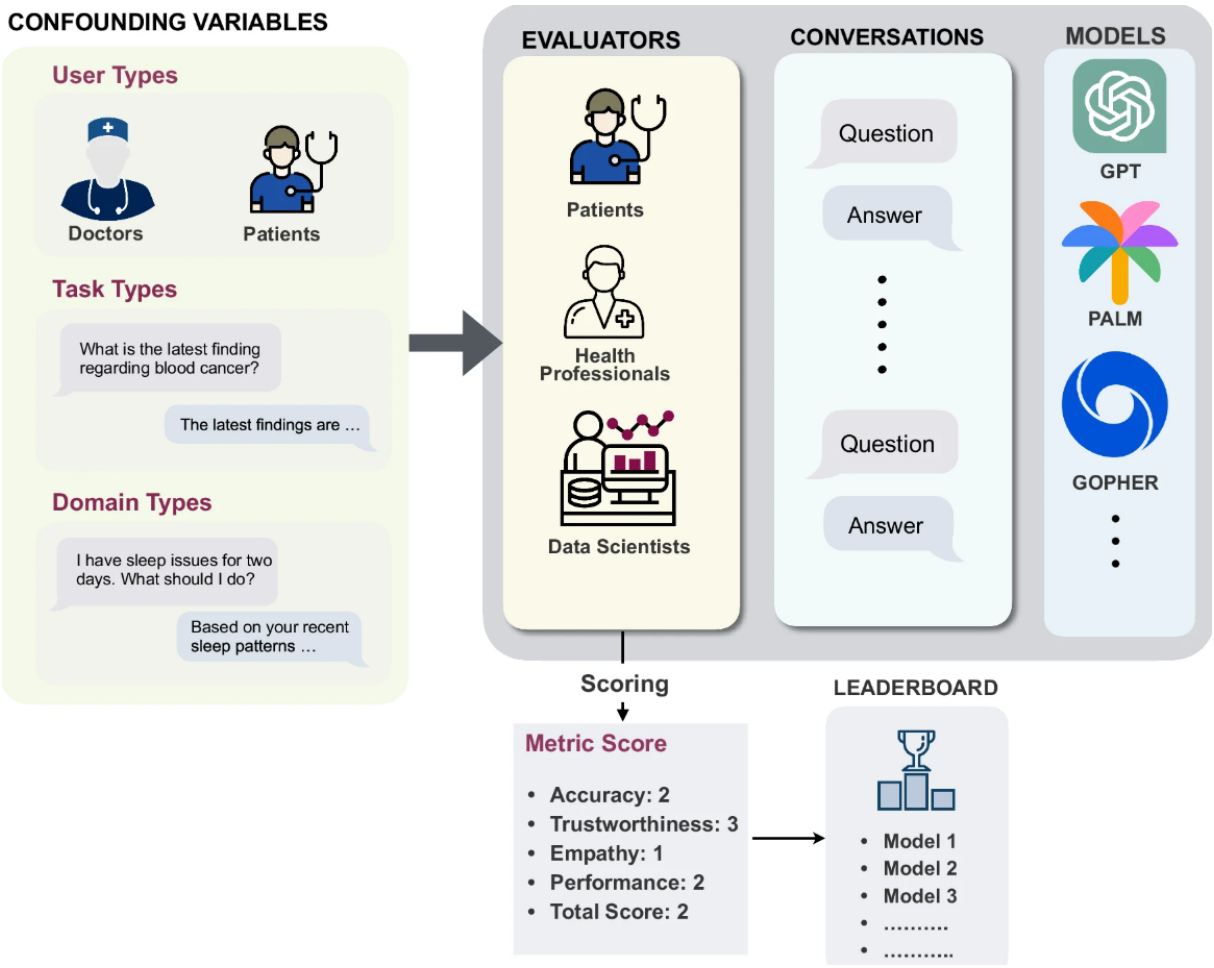

Figure 1: An example image

The main objective is to evaluate medical chatbots based on user interaction, considering the following confounding variables:

User Type

Users can be:

- **Patient**: An individual seeking health advice.
- **Doctor**: A healthcare professional looking for detailed information such as diagnoses, treatments, and prescriptions.
- **Nurse**: A healthcare professional involved in practical questions, including treatments to administer and daily care.

The evaluation should adapt to these user types based on security and data confidentiality requirements.

Domain Type

- **General**: A chatbot answering a range of health questions (symptoms, prevention tips, etc.).

- **Specific**: A chatbot dedicated to a particular domain, such as **mental health**, **chronic diseases**, or **cancer**. These chatbots must understand and respond with domain-specific and detailed information.

Task Type

The tasks the chatbot can accomplish vary, and metrics will need to be adjusted based on the task:

- **Medical Report Generation**: The chatbot generates a detailed medical report based on the symptoms and information provided by the user.
- **Medical Diagnosis**: The chatbot provides a diagnosis based on the user's information.
- **General Assistance**: The chatbot helps with daily tasks (medication management, appointment scheduling, etc.).

2. Essential Metrics

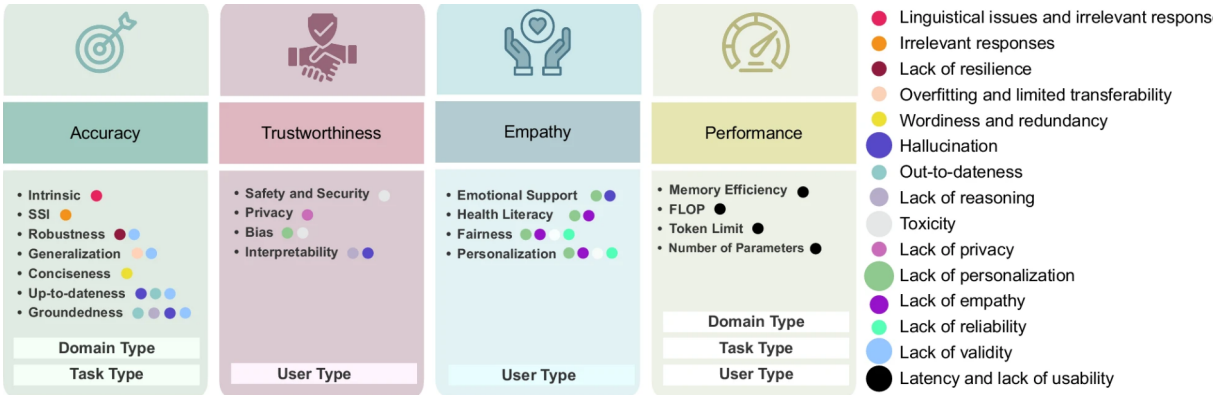

Figure 2: An example image

Evaluation metrics are grouped into four main categories: **Accuracy**, **Trustworthiness**, **Empathy**, and **Performance**.

A. Accuracy

Goal: To assess the chatbot's ability to provide accurate and relevant responses.

- **Metrics to evaluate**:
  - **Response accuracy rate**: Measure the percentage of correct responses compared to the total responses. This may include responses based on medical diagnoses, treatments, or health advice.
  - **Incorrect responses**: Analyze the error rate (incorrect or inconsistent responses) and their impact on the chatbot's reliability.

  - **Validation of medical information**: Check the validity of the provided information (e.g., medication prescription) according to the latest medical research and health guidelines.

Example test: Ask the chatbot questions about common symptoms (e.g., "What are the symptoms of the flu?") and verify the relevance and accuracy of the provided answers.

B. Trustworthiness

Goal: To measure the chatbot's credibility with users in terms of security, privacy, and the validity of the provided information.

- **Metrics to evaluate**:
  - **Perceived trust by the user**: This criterion measures how much the user trusts the chatbot, especially with sensitive issues (medications, diagnoses).
  - **Data protection**: Ensure the chatbot follows security standards and protects users' personal and medical data.
  - **Consistency of responses**: Verify that the chatbot provides consistent answers when the same question is asked multiple times.

Example test: Test the chatbot's ability to handle a sensitive query (e.g., "Is my information secure?") and provide a reassuring response while maintaining confidentiality.

C. Empathy

Goal: To evaluate the chatbot's ability to understand and respond appropriately to users' emotions in a human-like manner.

- **Metrics to evaluate**:
  - **Empathetic responses**: The chatbot's ability to emotionally respond to the user's concerns, especially in sensitive contexts like mental health or serious diagnoses.
  - **Tone and humanization**: Does the chatbot use an appropriate tone to reassure stressed or worried users?
  - **Emotion recognition**: Can the chatbot detect emotions in the user's queries (e.g., anxiety, anger) and adapt its responses accordingly?

Example test: Ask questions where the user may be worried or stressed (e.g., "I have chest pain, is it serious?") and evaluate whether the chatbot provides a reassuring and empathetic answer.

D. Performance

Goal: To measure the chatbot's responsiveness and efficiency in completing tasks.

- **Metrics to evaluate**:

  - **Response time**: How long does the chatbot take to respond to a question? A performant chatbot should respond quickly while providing accurate information.
  - **Request resolution rate**: How many questions or requests are effectively resolved by the chatbot? This includes providing useful responses and guiding the user to additional resources if needed.
  - **Technical stability**: Does the chatbot work properly without errors? This includes no bugs, disconnections, or other technical issues during interactions.

Example test: Ask a series of complex questions and measure how long the chatbot takes to give a complete and accurate response while ensuring that the system doesn't encounter bugs.

3. Confounding Variables

When evaluating the chatbot, it is crucial to consider the following **confounding variables** as they can influence the test results:

- **User Type**: Needs and expectations vary between a patient, doctor, or other healthcare professional. For example, a patient expects simpler and more reassuring answers, while a doctor looks for precise and technical information.
- **Domain Type**: A chatbot designed for **general health issues** will be evaluated mainly on its ability to provide simple and reliable answers. A chatbot designed for specific domains like **mental health** or **cancer** should be tested on its ability to understand more emotional contexts and provide relevant and specialized information.
- **Task Type**: The task the chatbot must perform influences the evaluation. For example, a chatbot generating **medical reports** should be evaluated on data accuracy, while a **general assistant** chatbot should be evaluated for its speed, ability to understand various requests, and provide useful solutions.

4. Scoring and Analysis of Results

The test results should be compiled and analyzed based on the following criteria:

- **Scoring**: Each interaction with the chatbot should be rated according to a predefined scale for each metric (e.g., 1 to 5, with 1 being poor performance and 5 being excellent performance).
- **Aggregation of Results**: After all tests are performed, the scores for each metric are aggregated to obtain an overall score for each type of chatbot evaluated. This allows for clear comparison of performance between different chatbots.
- **Dashboard**: The scores should be presented in a dashboard or report format to easily understand the chatbot's strengths and weaknesses.

2 Test Structure by Scenarios and Metrics

The tests will be organized into specific scenarios to evaluate each chatbot's performance based on the following metrics: **Accuracy, User Experience (UX), Performance, and Trustworthiness.** Each scenario will target a particular metric.

2.1 Accuracy Testing Scenarios

Objective: Test the chatbot's ability to provide accurate and verified medical information.

Scenario 1: Diagnosis of Disease

Description: The user asks the chatbot questions to evaluate the symptoms of someone suspected of having diabetes (or another chronic disease).
Metric: Accuracy of the chatbot's responses regarding symptoms and diagnosis.
Example Question: "What are the symptoms of type 2 diabetes?"

Scenario 2: Treatment Recommendations

Description: The user asks the chatbot for treatment recommendations for hypertension.
Metric: Accuracy of the treatment recommendations given by the chatbot.
Example Question: "What medications are recommended for hypertension?"

2.2 UX Testing Scenarios

Objective: Test the chatbot's usability, accessibility, and user engagement.

Scenario 1: Navigation and Interface

Description: The user interacts with the chatbot to find information about asthma (or another respiratory disease). The goal is to evaluate if the navigation is intuitive and clear.
Metric: Usability of the user interface and ease of use.
Example Task: "Find information about asthma treatments."

Scenario 2: Engagement and Response

Description: The user asks the chatbot about diabetes management. The goal is to evaluate the chatbot's engagement with the user and its response quality.
Metric: User engagement and the chatbot's response time.
Example Question: "What should I do if my blood sugar is too high?"

2.3 Performance Testing Scenarios

Objective: Test the chatbot's response time and stability under varying loads.

Scenario 1: Response Under Load

Description: The user asks multiple questions in a short period to test the chatbot's responsiveness.
Metric: Response time and performance under load.
Example Questions:
"What are the symptoms of diabetes?"
"What are the risk factors for hypertension?"

Scenario 2: Error Handling

Description: The user asks a series of medical questions, including poorly phrased or ambiguous ones, to test how well the chatbot handles errors.

Metric: Error management and robustness of the chatbot.
Example Question: "How to treat hypertension, or should I take this medication?"

2.4 Trustworthiness Testing Scenarios

Objective: Test the chatbot's ability to handle sensitive information and establish trust.

Scenario 1: Data Privacy
Description: The user asks the chatbot about how personal data is protected during interactions.
Metric: Trustworthiness and handling of sensitive information.
Example Question: "Is my personal data protected?"

Scenario 2: Accuracy of Medical Information
Description: The user asks about medications and their effectiveness in treating chronic diseases. The chatbot should provide validated and reliable answers.
Metric: Trust in the quality of the information provided.
Example Question: "Is this medication effective for treating hypertension?"

3 Test Execution Plan

3.1 Selection of Chatbots to Test

For each medical domain, select 3 to 5 chatbots focused on chronic disease management. For example, for diabetes, test 3 different chatbots with distinct characteristics.

3.2 Executing the Scenarios

For each chatbot, execute the scenarios as described above, with each team member responsible for testing specific scenarios for each metric (accuracy, UX, performance, trust).

3.3 Collecting Results

During the tests, each team member should systematically collect the results, focusing on:

- Response time (for performance testing)
- Errors detected (for error management testing)
- Quality of answers (for accuracy testing)
- User feedback (for UX and engagement testing)

3.4 Compiling Results

After completing the tests, compile the results and assign scores to each chatbot for each metric.