Introduction
EHRTutor is a multi-component framework to help patients comprehend their Electronic Health Record (EHR) notes, more specifically discharge instruction. Moreover, EHRTutor also offers a framework for generating synthetic patient education dialogues that can be used for future in-house system training.

Workflow

Evaluation
There are four main aspects for evaluation. We want to evaluate each sub-metric in the following standards with 5 points each. Deduct one point for each unsatisfactory sentence.

Metrics
Question:
1. cover rate: Do the questions cover the following categories (if applicable)?
2. Verifiable: Whether question can be answered by the discharge instruction.
3. Relevance: Do the questions relevance to discharge instruction?
4. Fluent & Conciseness: A concise and clear syntax and vocabulary, devoid of unnecessary question

Doctor agent:
1. Coherence: Whether the agent can make right decision

Response:
1. Relevance: Do the questions relevance to discharge instruction?
2. Sufficient: Can the patient recall or comprehend the correct answers based on hints?
3. Factuality: Is the response align with medical fact?

Summary:
1. Cover rate: Does the summary cover all key points?

Overall Performance

Discharge Instruction

Reason Chain
Thoughts: This is beginning of the conversation, I should ask a question.
Action: get question
Observation: Why do you want to leave the hospital?

Thoughts: The patient answer is incorrect, correct or not known.
Action: get question
Observation: What changes were made to your medications?

Thoughts: The patient was correct for the first time, I will give a hint.
Action: hint
Chain
Observation: Your answer is incorrect. Here is a hint...

Thoughts: The patient’s answer is incorrect twice, So I need to give the answer directly and ask the next question.
Action: hint, get question
Observation: Your answer is partially correct. You also missed something.

Evaluation Results

(a) human evaluation
(b) EHR Tutor
(c) Microsoft Azure ChatGPT

Table 1: Human evaluation results for each feature.

<table>
<thead>
<tr>
<th>Question</th>
<th>Agent</th>
<th>Response</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cover rate</td>
<td>Relevance</td>
<td>Patient</td>
<td>Rationality</td>
</tr>
<tr>
<td>4.80</td>
<td>4.80</td>
<td>4.00</td>
<td>4.36</td>
</tr>
</tbody>
</table>

Table 2: LLM evaluation results for each feature.

Human Evaluation
Pros:
- All questions and most responses are helpful for the patient to understand the discharge instruction.
- All questions do not need medical background to answer.
- Most responses are factually correct.
- More natural way of communicating with patients.

Cons:
- There are some cases that do not cover all key points in the discharge instruction.
- May assume the patient only got the symptoms mentioned in the discharge instruction and consider the answer related to symptoms not shown in the instruction as incorrect.

Large Language Models Evaluation
We conduct the evaluation based on the metric described on the left, and found the result align with human evaluation.