Scoring student practice conversations with OpenAI

Goal:

Our goal for this project was to identify the ability of an AI model to analyze and grade student conversation data from the new “Mr. Kato” chatbot module and provide qualifying feedback.

Key points

- AI grading works better with lots of API calls on multiple specific rubrics.
- o3-mini-high was the best model available through OpenAI.
- AI appears to be a harsh evaluator.
- The model was highly consistent in evaluation passes, 85% – 100% reproducibility.
- Estimated cost per evaluation is $0.04 in API usage.

What We Used:

- Python for coding
- OpenAI API key as the grader (using o3-mini reasoning model)
- Student conversation logs for our data (15 students)
- University of Kentucky’s AI OSCE Grader as a template and inspiration

Process:

A trial of the process in the OSCE Grader template provided by the University of Kentucky revealed incompatibilities with the Kato output, as it does not result in a SOAP note. We also identified early that a single evaluation call would result in a high level of variability. We created a de novo evaluation rubric starting with the prompts used in the Mr. Kato module and including information from the AI OSCE Grader from the University of Kentucky, Royal College documentation, and OSCE materials provided by UBC Faculty Development. We then used a dataset of 15 unique student conversations from the Mr. Kato module chosen for their completeness to evaluate variability in the LLMs application of our rubric. Any identifiable information was removed from the conversations. The rubric is provided to the LLM through the chat completions API as a prompt. The model is provided the context of the student conversation including messages with the AI preceptor. Evaluations were performed multiple times and the variability in scores was measured for each student conversation using statistical methods. Initially Choen's Kappa was used to validate the difference between two AI evaluations but that was abandoned to use multiple reviews and create an Intraclass Correlation Score (ICC). The conversations with higher variability or lower agreement between raters were used to tune the prompts. Language in the prompt was adjusted, and the tests were rerun, comparing scores. The process was iterated until we were able to achieve a signal of high agreement between evaluations and ran the evaluations 10 times for each conversation. We were able to achieve a consistent output of 87% - 100% as indicated by the ICC score. This shows a high rate of consistency in LLM evaluations.

The evaluation rubrics were created for the following categories: History of Present Illness (HPI), Differential Diagnosis (DDX), Communication, and clinical reasoning. The context of the student interaction includes messages and responses between the student and the patient and the student and virtual preceptor. The data was organized using python scripts to be presented in an organized and segmented fashion for evaluation by the LLM.

The process was tried with a few models available through the OpenAI API and 03-mini-high was selected as the model that had the current best balance of accuracy, consistency, latency and cost. Doing an evaluation of 15 conversations 10 cost a total of $6, with the whole investigation costing $17.34 in API usage from OpenAI.

Limitations

Validation of the scores requires expert review. We are able to pursue a consistent score and justification but true valuation of them will require review by a clinical faculty member. Evaluation of AI clinical reasoning has been shown to be high and there is a high level of trust in the output.

There is a lack of DDX data. Students are not engaging with the differential diagnosis aspects of the module. This should be reviewed as a design flaw.

Lack of qualitative feedback from students. No students who have completed the module have signed up to volunteer an opinion or perspective on it.

Next Steps

Get expert review of the scores and justifications. We want to validate that the rubrics are evaluating are providing accurate scores. This requires expert insight and faculty must be engaged. To that end we will need to get the conversations and rubrics available for faculty review and scoring.

Revise the module so more learners are engaging with the DDX aspects. We need more data on differential diagnosis and it's justification to validate the rubrics. The sample size indicates effectiveness but needs to be larger to provide validation.

We also want to reduce the overall cost in our process. This is a preliminary result and has not be optimized. We will investigate methods for optimizing the usage of the OpenAI API to reduce processing required. This could include batch requests, alternative models, more efficient prompts using structured outputs.

Integrate scoring into the feedback process for the module. Once a student finishes their chat with Mr. Kato, they are given very general feedback from the preceptor with no way to quantify how they performed. Integrating this into the chatbot directly will solve this issue by providing students with a score along with detailed feedback on how they performed and what they can improve upon the next time around. Ideally, we expect to see students retrying the chatbot after their initial attempt to achieve higher scores.

We need to create a script that can be used to validate the output of different models against the experience. As models change we want an eval that can be run to validate our ability to upgrade without disrupting the learners experience.

Output monitoring on the grading tool needs to be put in place. The first step will be to define the guardrails and edge cases that could lead to policy and pedagogy risks.

Discussion

Our biggest challenge was ensuring students can get fair and accurate marks in 4 different categories: **HPI (History Taking), DDX (Differential Diagnosis), Communication, and Clinical Reasoning**. Since there are a near infinite way to ask the same question, we simply can’t just rely on using the student message history to grade these. We addressed this issue by wrangling the data in such a way that it grades the student’s message while looking at the responses from both the preceptor and patient as context to see what information they were able to extract from both the chatbots. This ensures that if a student asked a “poor” question but still had the right idea and extracted the information out of Mr. Kato they would be graded as such rather than getting a low score of 1.

We created the prompts based on the original prompt given to Mr. Kato and the preceptor for the chatbot. After breaking the prompt down, we were able to figure out information such as what the key clinical findings are etc. This helped us create a template to build our prompt from. After a multitude of tests costing ~$18, we finally landed on prompts which gave accurate feedback and had the lowest variance across multiple different runs. You can find the list of prompts here.

Results

We were able to achieve a high level of consistency in the prompt output using the o3-mini model set to a ‘high’ level of reasoning. To quantify the variance, we tested the ICC (Intraclass correlation coefficient) across 10 different runs using the same prompt and the same data. Our results are as follows:

- HPI Score:
  ICC: 0.8869
  - 95% Confidence Interval: (0.9353, 0.6748)
- Communication Score:
  - ICC: 0.8643
  - 95% Confidence Interval: (0.9204, 0.5997)
- DDX Score:
  - ICC: 1.0000
  - 95% Confidence Interval: (1.0000, 1.0000)
- Reasoning Score:
  - ICC: 0.9889
  - 95% Confidence Interval: (0.9943, 0.9713)

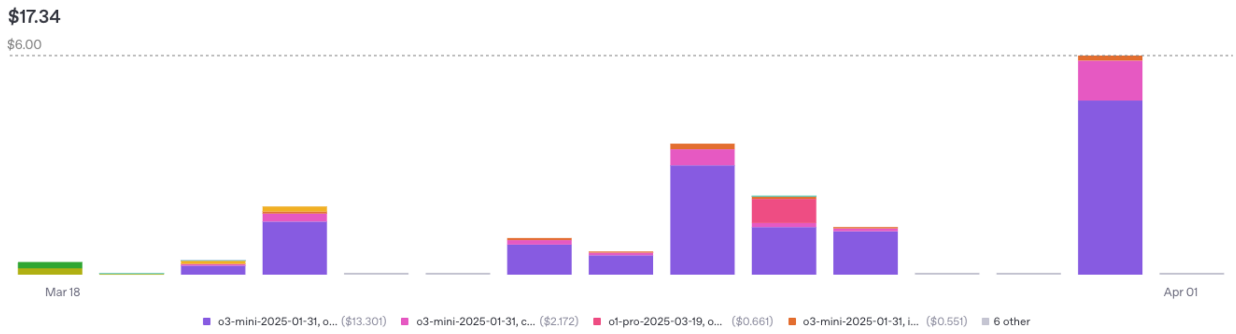
Overall, these are very good results. While HPI and Communication have the most variance, we are still achieving over 0.85 ICC score which is considered “excellent reliability”. DDX and Reasoning may be higher due to a smaller sample size.

Areas for improvement

Evaluation of the interview does not fully take into account information that has been voluntarily given by the patient. In instances where Mr. Kato responds with information that includes details, e.g. ‘I feel good today but have a problem with my right eye’, the student is being punished for not asking about which eye has the problem.

Student interaction is not showing a sufficient level of engagement to provide a thorough analysis. The module needs to have a structured output area that evidences the learners ability to evaluate the encounter, a PEP or SOAP note.

Model cost usage



Although we spent ~$18 to perform the tests, in practice this would cost roughly $0.04 for each student to receive full feedback in the 4 testable areas.

Flowchart showing the process of grading the conversations

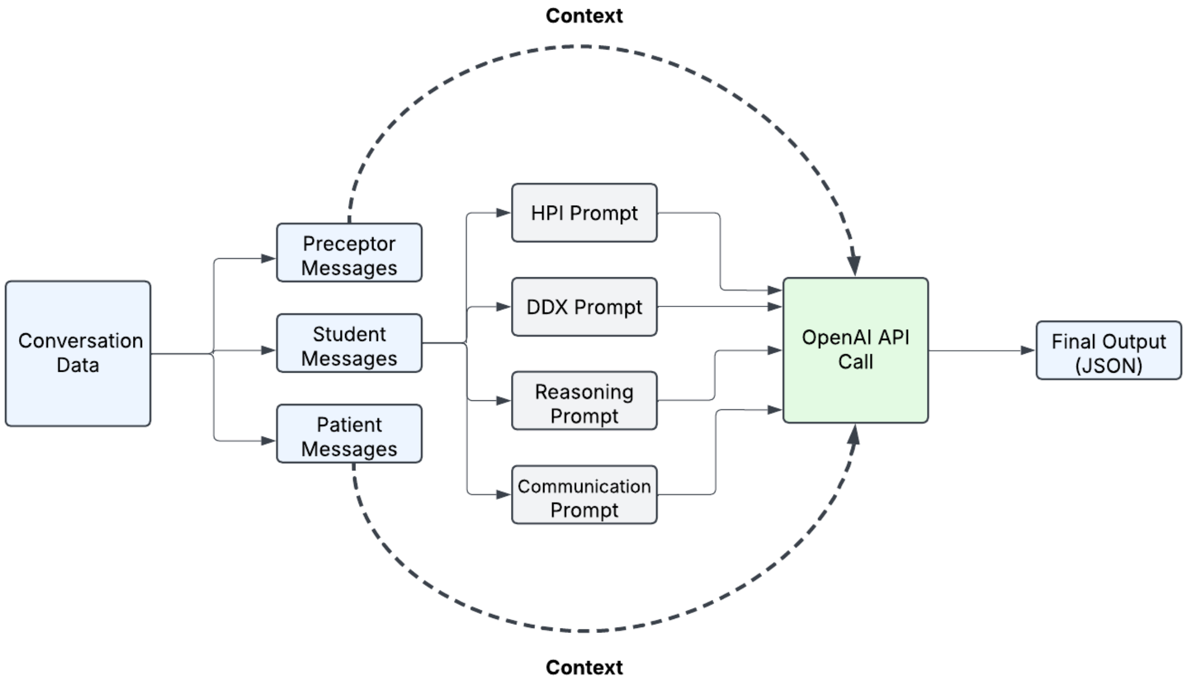


Table showing the scores for each conversation.

	HPI 0 - 5	Communication 0 - 5	DDX 0 - 5	Reasoning 0 - 5
Person 1	2	3	0	0
Person 4	2	3	2	1-2
Person 12	3-4	4	0	0
Person 14	1	5	0	0
Person 15	1	3	0	0
Person 16	1	3	0	0
Person 18	2	3-4	4	3
Person 19	1	1	0	0
Person 20	3	2-3	3-4	3
Person 22	2-3	4-5	0	0
Person 25	1	0	0	0
Person 27	2-3	2-3	0	0
Person 28	1-2	3	0	0
Person 32	1	1-2	0	0
Person 35	2-3	3	4	3

Example of justification output:

This is an example feedback snippet from a student who achieved a score of 2 on their history taking skills:

```json
"hpi_feedback": "Your HPI questioning captured only a few of the 10 essential elements. You asked about laterality (“is it only in your right eye?”), the impact on the patient’s life (“how has this impacted your life?”), and family history of eye disease. However, many critical aspects were not addressed. You did not ask about the duration of the vision loss (which should capture that it started 4–5 months ago with notable worsening in the last 2–3 weeks), the specific characteristics of the vision loss (i.e. full visual field loss), or if the patient has increased difficulty with tasks such as night driving or reading. You also did not explicitly inquire about his detailed diabetes history (15 years of type 2 diabetes) and cardiovascular history (minor heart attack 3 years ago), nor his smoking history. Focusing on these areas during the interview would improve your history taking to ensure all 10 key elements are systematically and explicitly addressed."
```