AI Teacher and Evaluator

Context

When a student answers a question, the AI gives feedback. Is that feedback appropriate? To assess this, we want to evaluate the AI's feedback.

Goal

Write a program (code + prompts) that assesses whether an AI teacher is giving appropriate feedback. In order to do this, you will need to use GPT to play 3 roles:

A teacher

who asks the student a set of questions, waits for the student's response, and gives feedback. Use this set of questions to ask the student.

A student

who responds to the teacher's question, but commits errors (as students in the real world do!). Here is a list of errors that a student can commit. If they commit any of these errors, their response is incorrect.

An evaluator

that looks at the teacher's feedback in response to the student and determines if it correctly identified the error that the student committed.

Function Decription

The sequence of API calls should be, for each question: ask the student a question, ask the teacher to respond with feedback to the student's answer, ask the evaluator to assess the teacher's feedback. The following functions can be found in `process.py`

- **ask_question(question, api_key)**: Gets the response from the student.

```
  bug_added_prompt = f"Have an incorrect response to question
              by change the response to include inaccurate facts, or to fail to
directly answer the question,
              or to use incomplete sentences."
```

- **give_feedback(question, response, api_key)**: Gives feedback on the response given the question.

```
  find_problem_prompt = f"Identify factual and grammatical errors or incomplete
sentence in the response
                          regarding a question"
  suggestion_prompt = f" After evaluating the response, suggest a response with
improvement."
```

- **evaluate_feedback(question, response, feedback, api_key)**: Evaluates the feedback given the response and question.

```
  evaluation_prompt = f"Assess the reasonability and fact-check on the feedback
                    to the following response to the following question"
  quality_evaluation_prompt = f"Assess the reasonability critifically and fact-
check of the feedback
                                on the reponse regarding the question"
```

```
  generic_evaluation_prompt = f"is the feedback appropriate and correct for the
response regarding question"
  suggestion_prompt = f" Give one improved suggestion for the feedback."
```

Evaluation

Handling evaluation on teacher's feedback is a challenging task because it not only has to account for the factual correctness and grammatical correctness, it also must be appropriate and concise as opposed to being verbose. While working on the process, I come to a realization that this task may or may not need a separate model trained that evaluates the appropriateness apart from correctness checking. Additionally, regarding any time-series based data such as history, the response and be more expansive to cover a broader historical perspective by suggesting connected events.