Learning Management System (LMS) using gRPC and LLM

ACHINTHYA HEBBAR S (2021A7PS1457P)
UJJWAL AGGARWAL (2021A7PS2427P)
NISHANT SHUBHAM (2023H1120196P)

The **GPT-2 model** was integrated to serve as an AI tutor for students. The AI model provides context-aware responses to students' queries, enhancing the learning experience.

a. Tutoring Server

A new **Tutoring gRPC service** was created to handle AI-generated responses. This service runs on a separate port (`50052`) and responds to requests from the LMS server. The Tutoring Server performs the following actions:

1. Receives the student's query along with course details.
2. Sends the query to the GPT-2 model for response generation.
3. Returns the generated response back to the LMS server for storage and display.

b. Prompt Engineering and Response Generation

We improved the quality of the GPT-2 responses by:

- Crafting better prompts, including specific context like the course name and an instruction to provide simple answers (e.g., "Please explain in simple terms").
- Adjusting **GPT-2** parameters:
    - `max_length` was reduced to 200 to avoid verbose responses.
    - `temperature` was set to 0.5 for more coherent responses.
    - `top_p` was adjusted to 0.9 to ensure the selection of the most relevant tokens.
    - `repetition_penalty` was increased to 1.5 to avoid repetitive or incoherent text.

4. System Workflow

The integrated system works as follows:

1. **Student Query**: A student submits a question through the LMS client. If the query is marked for AI assistance (`is_ai=True`), the LMS server forwards the request to the Tutoring server.

2. **LLM Processing**: The Tutoring server processes the query using GPT-2, generating a context-aware response.
3. **Response Storage**: The generated response is saved in the LMS database under the `queries` table and is available as a reply to the student's original question.
4. **Error Handling**: If the Tutoring server is unavailable or encounters an issue, the LMS server returns an appropriate error message to the client.

5. Challenges and Optimizations

- **Random Responses**: The initial GPT-2 model generated overly verbose or off-topic responses. To mitigate this, prompt engineering was employed, and response generation parameters were fine-tuned.
- **gRPC Integration**: We ensured smooth integration between the LMS and Tutoring servers by setting up robust error handling and efficient communication using gRPC.