AI Agent for Academic PDF Summarization and Question Generation

Ankit Yadav

November 3, 2025

Abstract

This report presents the design and implementation of an AI agent that automates academic revision by summarizing PDF documents and generating theoretical questions. The system is built upon the fine-tuned `google/gemma-2-2b` model using the QLoRA approach. The project follows a complete pipeline—from data collection and preprocessing to fine-tuning, evaluation, and deployment as a fully functional planner–executor agent integrated with a Hugging Face model API and LaTeX-based PDF output rendering.

1. Introduction

Students preparing for examinations often struggle to quickly revise large volumes of study material. To address this challenge, an AI agent was developed to read PDF files, extract essential content, generate concise summaries, and produce relevant theoretical questions for rapid revision. The agent combines natural-language understanding, reasoning, and automation through a planner–executor architecture.

2. Stage 1: Dataset Creation and Preparation

2.1. Objective

The primary objective was to construct a dataset tailored to the dual task of summarization and question generation for academic texts, since no such dataset existed publicly.

2.2. Data Sources

Open-access educational and research repositories were scraped to collect PDF materials:

- arXiv.org
- ResearchGate

- Academia.edu
- NCERT Official Website
- IIT Kanpur OCW
- MIT OpenCourseWare

2.3. Pre-processing Pipeline

Implemented in `pdf_reader.ipynb`:

1. **Text extraction** using `PyPDF2` and `pdfminer.six`.
2. **Cleaning** of headers, footers, page numbers, and references.
3. **Chunking** texts into 300–500-token segments maintaining semantic continuity.
4. **Storage** of clean chunks in JSON for later prompt generation.

2.4. Synthetic Data Generation

Each text chunk was processed via the Gemini API with prompts requesting a concise summary and several conceptual questions. The generated pairs were saved as:

```
{"input": "<chunk_text>", "output": "<summary>\n<questions>"}
```

After removing duplicates and incomplete entries, the resulting JSONL dataset served as the fine-tuning corpus.

3. Stage 2: Model Fine-tuning Using QLoRA

3.1. Fine-tuning Target

The base model **google/gemma-2-2b** was chosen for:

- **Task specialization:** strong performance on summarization and educational text.
- **Improved reliability:** instruction-tuned weights yield consistent question formats.
- **Adapted style:** easily personalized output tone for academic clarity.
- **Resource efficiency:** compact enough for LoRA-based tuning on a single GPU.

3.2. QLoRA Configuration

The fine-tuning implemented in `model_summarizer.ipynb` employed:

- 4-bit quantization using `BitsAndBytes`.
- LoRA adapters on key attention layers.
- Learning rate $\approx 2 \times 10^{-4}$, batch size = 4, 3–5 epochs.
- Supervised fine-tuning objective on input–output pairs.

This setup preserved base model knowledge while efficiently adapting to the summarization + QG domain.

3.3. Training Results

The model converged stably with smooth loss decay and improved coherence in generated summaries. Qualitative inspection confirmed improved fluency, factual consistency, and relevance of questions compared to the base Gemma model.

4. Stage 3: AI Agent Architecture

4.1. Overall Design

The agent follows a modular **Planner + Executor** pattern:

- **Planner:** Interprets the user's goal (e.g., "Summarize this PDF"), decomposes it into subtasks—extract text, segment, summarize, question generation, and formatting.
- **Executor:** Executes each subtask—calls the fine-tuned model for summarization/QG, converts output to LaTeX for readability, and assembles the final PDF.

4.2. Pipeline Flow

1. **Input:** User uploads a PDF.
2. **Planner:** Decides required operations.
3. **Executor Modules:**
    - Text extraction.
    - Model inference via Hugging Face API.
    - LaTeX conversion and PDF rendering.
4. **Output:** Final summarized document with embedded questions for revision.

4.3. Integration and Deployment

The fine-tuned model was uploaded to Hugging Face Hub for public inference. The agent fetches the model through API calls, ensuring scalability and version control. Generated summaries and questions are formatted into LaTeX for professional-quality visualization.

5. Stage 4: Evaluation

5.1. Qualitative Evaluation

In the present implementation, evaluation was performed manually by observing generated outputs. The following aspects were assessed:

- **Summarization Clarity:** Summaries retained key ideas while compressing the text effectively.
- **Question Relevance:** Theoretical questions reflected the main concepts and terminology in the source text.
- **Formatting Quality:** LaTeX-based rendering improved readability and presentation.

5.2. Observations

The fine-tuned model produced contextually accurate and well-structured summaries. The generated questions were meaningful and academically appropriate. Outputs demonstrated a consistent tone aligned with educational material, validating the effectiveness of the fine-tuning process.

6. Conclusion

This project successfully demonstrates an end-to-end AI agent that automates academic revision through PDF summarization and question generation. By fine-tuning `google/gemma-2-2b` with QLoRA, the system achieved domain specialization while maintaining efficiency. The planner–executor architecture ensured modularity and adaptability, while deployment via Hugging Face enabled easy access and scalability. Future work includes integrating quantitative evaluation metrics, expanding the dataset to additional academic disciplines, and incorporating retrieval-augmented context for better factual grounding.