**PDF Chat API: Problems**

**1. Choosing Between Gemini 1.5M and RAG**

**Using Gemini Directly**

**Good Points**

* Easy to set up - just plug and play
* No extra databases needed
* AI sees the whole document at once
* Better answers about the whole document

**Problems**

* Uses more resources
* Costs more money
* Takes longer to process
* Limited to 1M tokens

**Using RAG (Retrieval-Augmented Generation)**

**Good Points**

* Can handle any size of document
* Cheaper to run
* Faster responses
* Works well with many documents

**Problems**

* Harder to set up
* Needs a vector database
* Might miss some connections in the text

**What to Choose?**

* For small PDFs (< 100KB): Use Gemini
* For big PDFs (> 100KB): Use RAG
* For one document: Use Gemini
* For many documents: Use RAG
* If money is tight: Use RAG
* If you need it working fast: Use Gemini

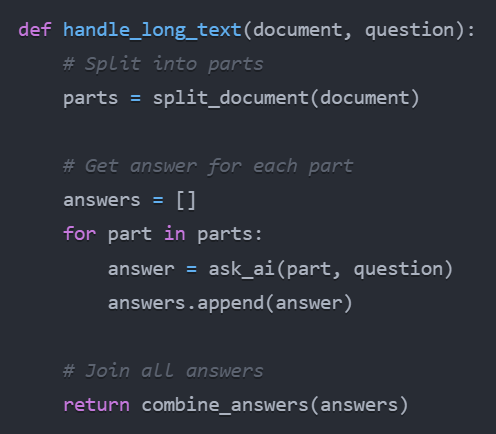
**2. Handling Token Limits**

**The Problem**

Gemini can only output 8,196 tokens at once. This means long answers get cut off.

**How to Fix It**

**Method 1: Split and Join**

****

**Method 2: Summary First**

1. Get a short summary
2. Ask about important parts
3. Put everything together

**3. Evaluate The LLM Performance**

**What to Test**

* How fast it answers
* If answers make sense
* If answers are correct
* How much it costs to run

**Test Cases Example**

****

**How to Test**

1. Make test PDFs with known answers
2. Ask AI questions
3. Compare AI answers with real answers
4. Keep track of scores

**What We Should Do**

1. Start with Gemini - it's easier
2. Add RAG later if needed
3. Test everything well
4. Ask users if answers are helpful
5. Fix problems as we find them