**Model Choice for Study**

**1. Context**

We are building a study assistant that can answer questions from PDFs, audio, video, and lecture materials. The system retrieves relevant content using embeddings stored in pgvector and generates simplified, human-readable answers using an LLM.

**2. Gemini AI**

**Advantages:**

* Supports citation-backed responses, ideal for academic purposes.
* Can integrate public data directly to enhance answers.
* Better at complex reasoning and multimodal inputs.

**Considerations:**

* Cost is significantly higher, especially for frequent or high-volume queries.
* Integration is newer and may require more development effort.

**Use case fit:**

* Gemini fits perfectly for academic-quality answers, citations, and public data integration.

**3. GPT-3.5-Turbo**

**Advantages:**

* Cost-effective and scalable for high-volume queries.
* Mature API ecosystem (embeddings, LLM, fine-tuning).
* Fast and reliable for generating simplified answers from retrieved chunks.

**Considerations:**

* Does not automatically provide citations.
* Requires manual fetching or integration of public data if needed.

**Use case fit:**

* GPT-3.5-Turbo is ideal for the development phase when many tokens are needed, keeping costs manageable.

**4. Decision**

* **Development Phase:** We will use GPT-3.5-Turbo due to cost efficiency and ease of integration. It allows us to test and iterate on the system extensively without incurring high costs.
* **Future Consideration:** Gemini can be integrated later for advanced features like citation-backed answers and better public data synthesis once the system is stable and token usage is optimized.