**How Retrieval and Generation Work in the Code:**

In this system, the **retrieval** process starts when a user uploads a PDF. The document is broken down into smaller chunks of text using a method that ensures each piece is easy to handle. These chunks are then turned into vector embeddings using a pre-trained model, which converts the text into numerical representations that capture its meaning. These embeddings are indexed using FAISS, which is a fast way to search and retrieve the most relevant pieces of text when a user asks a question.

When a query is made, the system **retrieves** the most relevant chunks by turning the question into its own vector and searching through the index for the best matches. The top chunks are returned, each with a score showing how closely it matches the query. These chunks are then passed to the **generation** step, where the Gemini AI model creates a response using the context from the retrieved text. The generated answer is tailored to the user’s question, while also citing the relevant pages from the original document. This combination of retrieving relevant information and generating a response based on it allows the system to answer specific questions accurately, making use of the uploaded document’s content.