Problem Statement

Traditional AI assistants are limited to handling only one type of data — either text, or images, or documents. In the real world, however, users interact with multi-modal data (documents, text, and images simultaneously).

For example:

- A user may want to upload a PDF contract and ask questions about it.
- Another may want to upload an image and request a description or caption.
- Or they may want to chat directly with the assistant using natural language.

Current assistants struggle to combine these modalities into a single seamless workflow.

Solution Approach (Multi-Modal Assistant)

Build a Streamlit-based Multi-Modal Assistant that integrates:

1. **Document Understanding (RAG pipeline)**
   - Users upload PDFs, DOCX, TXT files.
   - Text is extracted, chunked, embedded with `SentenceTransformers`.
   - A vector database (in-memory via sklearn NearestNeighbors) is used for retrieval.
   - Relevant context is passed to a **Text Generation Model (Hugging Face Transformers)** for answering.
2. **Image Understanding**
   - Users upload an image.
   - Vision model (like BLIP, CLIP, or ViT) generates captions or embeddings.
   - Captions are used as text input for Q&A.

3. **Text Chat**
    - Users can directly ask questions.
    - The assistant leverages LLM (HuggingFace model, e.g., Flan-T5 or LLaMA-based) for natural conversation.
4. **User Interface (Streamlit)**
    - Single dashboard for uploading files, images, and chatting.
    - Multi-modal inputs (text, image, document) handled in a unified manner.

Key Benefits

- **Multi-Modal**: Handles text, images, and documents in one system.
- **Retrieval-Augmented**: Ensures accurate answers from uploaded files.
- **Extensible**: Can plug in different LLMs or vision models easily.
- **User-Friendly**: Simple Streamlit UI for non-technical users.

Solution Architecture Diagram

```
                     ┌────────────────────────┐
                     │     User Interface      │
                     │ (Streamlit Dashboard)   │
                     └───────────┬─────────────┘
                                 │
       ┌─────────────────────────┼─────────────────────────┐
       │                         │                         │
┌──────▼───────┐          ┌──────▼───────┐          ┌──────▼───────┐
│ Text Query   │          │ Document Q&A │          │ Image Q&A    │
│  (Chat)      │          │ (PDF, DOCX)  │          │ (Image→Text) │
└──────┬───────┘          └──────┬───────┘          └──────┬───────┘
       │                         │                         │
       │                  ┌──────▼──────────┐              │
       │                  │ Text Extractor  │              │
       │                  │ & Chunker       │              │
       │                  └──────┬──────────┘              │
       │                         │                         │
       │                  ┌──────▼──────────┐              │
       │                  │ Embedder (SBERT)│              │
       │                  └──────┬──────────┘              │
       │                         │                         │
┌──────▼────────────┐   ┌────────▼─────────┐   ┌───────────▼───────┐
│  LLM (HuggingFace│   │ Vector Index (NN)│   │ Vision Model (BLIP│
│   e.g. Flan-T5)   │   │ Context Retrieval│   │  / CLIP / ViT)     │
└──────────┬────────┘   └────────┬─────────┘   └───────────┬───────┘
           │                     │                         │
           └──────────────┬──────┴───────────┬─────────────┘
                          │                  │
                     ┌────▼──────────────────▼───────┐
                     │  Unified Response Generator    │
                     └──────────────┬─────────────────┘
                                    │
                          ┌─────────▼────────┐
                          │   Final Answer    │
                          └──────────────────┘
```