Issues Faced & Solutions – Multi-Modal Assistant

Problem 1:

Model Download Failures (Hugging Face / Large Models)

- **Issue:**
  Hugging Face models like google/flan-t5-small, Salesforce/blip-vqa-base, and sentence-transformers/all-MiniLM-L6-v2 require large downloads (hundreds of MB). In restricted environments, downloads failed or timed out.

- **Solution:**

    - Used **smaller models** for testing (flan-t5-small, MiniLM-L6-v2).

    - Pre-downloaded models locally and pointed Hugging Face cache via HF_HOME.

Added instructions to run:

huggingface-cli download google/flan-t5-small --local-dir ./models/flan-t5-small

    - then load from local path.

Problem 2:

Streamlit Cache Hashing Errors

Issue:
Passing model objects (non-hashable) into st.cache_data and st.cache_resource raised:

UnhashableTypeError: Cannot hash object of type 'AutoModel'

- **Solution:**

    - Used _embedder naming convention in get_embeddings to avoid hashing on the full model object.

    - Kept large models in st.session_state instead of st.cache_data.

Problem 3:

GPU Memory Exhaustion

- **Issue:**
  Running VQA (blip-vqa-base) and generation (flan-t5-small) on GPU together caused CUDA out of memory errors.

- **Solution:**

  Forced CPU execution for some models:

  model.to("cpu")

  - Reduced image size (640x640) before inference to save VRAM.
  - Allowed configurable max_new_tokens + chunk_size sliders in sidebar.

Problem 4:

PDF/Text Parsing Inconsistencies

- **Issue:**
  Some PDFs were scanned images (no extractable text), or chunking cut sentences mid-way, leading to poor retrieval.

- **Solution:**

  - Added **chunk_overlap** parameter to preserve context between chunks.
  - Implemented fallback to OCR (via pytesseract) for image-based PDFs.
  - Normalized whitespace and stripped metadata during chunking.

Problem 5:

Retrieval Returning Irrelevant Chunks

- **Issue:**
  retrieve_top_k sometimes surfaced irrelevant text due to short queries or embeddings mismatch.

- **Solution:**

- Added **Top-K slider** in sidebar for tuning retrieval depth.
- Used sentence-transformers/all-MiniLM-L6-v2 for better semantic retrieval.
- Included similarity score (sim=0.xxx) in UI so user can judge relevance.

Problem 6:

VQA Model Answering Irrelevantly

- **Issue:**
  BLIP sometimes gave generic answers unrelated to the question.
- **Solution:**
  - Forced explicit question context (resized input + normalized wording).
  - Combined VQA result with text retrieval inside a **joint prompt** to the generator.
  - Final multimodal answer crafted by LLM (generate_answer) instead of directly trusting VQA.

Problem 7:

Multimodal Prompt Construction

- **Issue:**
  Combining retrieved text chunks + VQA output into a single coherent prompt sometimes caused the generator to hallucinate.
- **Solution:**

  Designed structured prompt template:

  Image hint (from VQA): …
  Question: …
  Passages: ...
  - Instructed LLM: *"If uncertain, say you don't know"* to reduce hallucination.

Problem 8:

Performance Issues with Large PDFs

- **Issue:**
  Uploading big PDFs (50+ pages) slowed UI and memory.

- **Solution:**

    - Limited number of chunks processed at once.

    - Displayed progress (st.spinner) during chunking.

    - Suggested preprocessing documents offline for very large corpora.