Issues Faced

Problem 1 :

OpenAPI : An **OpenAI API key** is required to use **OpenAI-hosted models** (like whisper-1, gpt-4o-mini). But OpenAI requires a billing setup after a free trial.

Took **Streamlit app** to use **Hugging Face free models** instead.

Used:

- openai/whisper-small → for **speech-to-text (transcription)**
- facebook/bart-large-cnn → for **summarization**
- google/flan-t5-large → for **action item extraction**

Problem 2 :

ModuleNotFoundError: No module named 'transformers'

- # Install transformers and dependencies
  pip install transformers datasets accelerate
- pip install git+https://github.com/openai/whisper.git

Problem 3 :

ERROR: Could not install packages due to an OSError: [Errno 28] No space left on device

(means your disk (or the container/virtual environment partition) ran out of space while installing **PyTorch + CUDA**, which are **huge** (several GB).)

- Install CPU-only PyTorch (much smaller)
  pip uninstall torch -y
  pip install torch --index-url https://download.pytorch.org/whl/cpu

- pip cache purge
- pip install transformers accelerate

Problem 4 :

ffmpeg is missing in your system. Hugging Face's Whisper ASR pipeline needs ffmpeg to read .mp3/.wav files.

- sudo apt update
- sudo apt install ffmpeg -y

Problem 5 :

Long Processing Time for Large Audio Files

- **Issue:**
  Processing long meeting recordings (>30 mins) took a lot of time and sometimes caused timeout/crash.

- **Solution:**
  - Pre-process audio into smaller **chunks (5–10 min segments)**.
  - Transcribe each chunk separately, then merge results.
  - Optionally use **faster models** (e.g., whisper-tiny or whisper-base).

Problem 6:

Out of Memory (RAM) During Summarization

Issue:

Large summarization models (like facebook/bart-large-cnn) consumed too much memory, leading to:

```
RuntimeError: CUDA out of memory
```

- **Solution:**
  - Switch to smaller distilled models (sshleifer/distilbart-cnn-12-6).
  - Use CPU-only mode if GPU RAM is limited.

- Enable gradient checkpointing in Hugging Face for memory efficiency.

Problem 6:

Streamlit Caching Issues

- **Issue:**
  Without caching, every interaction reloaded models → very slow app.

Solution:
Used @st.cache_resource to cache Hugging Face pipelines:

Problem 7:

Inconsistent Action Item Formatting

- **Issue:**
  flan-t5-large sometimes generated action items as plain text instead of a proper checklist.

Solution:

Added **prompt engineering**:

```
actions = extractor(action_prompt, max_new_tokens=400)[0]["generated_text"]
```