PROJECT 3

MULTI-MODAL AI CHATBOT FOR VIDEO

By Pedro Traba

PROJECT 3

INTRODUCTION

RAG from video transcription

PROJECT 3

01

THE PITCH

(Not perfect)

PROJECT 3

720,000

Hours of video are uploaded daily to Youtube

PROJECT 3

14,000,000,000

Videos on Youtube

11.7 min.

Average video length

311,643 years

To watch all Youtube videos

PROJECT 3

“Ain’t nobody got time for that.”

—THAT ONE LADY

PROJECT 3

WHY DO I NEED IT?

TIME

Summarizing, extracting specific parts of long videos.

ACCESSIBILITY

Language barriers, not being able to watch videos.

FUNCTIONALITIES

01 YOUTUBE AND LOCAL VIDEO

02 TRANSCRIPTION AND DIARIZATION

03 ASK ANYTHING

04 TWO MODES

- Fast
- Accurate

02

THE CODE

(Decoded)

PROJECT 3

APPROACH

FUNCTIONS AGENT EVALUATION DEPLOYMENT

WHAT I USED

LANGCHAIN OPENAI

CHROMADB

LANGSMITH

GETTING THE VIDEO

DOWNLOAD FROM YOUTUBE

Extracting Metadata from Youtube in the process

UPLOAD FROM DEVICE

Extracting video Metadata

PROCESSING

EXTRACTING THE AUDIO

Getting the .wav from the .mp4

WHISPER AND PYANNOTE

Transcriptions and Diarization

CREATING THE RETRIEVER

COMBINING DOCUMENTS

Metadata, Transcription, Diarization

VECTORSTORE AND RETRIEVER

Chromadb from documents

CREATING THE AGENT

TOOLS AND PROMPT

Started with several tools, ended up with one.

MODEL AND MEMORY

GPT-4o, REACT agent

from langgraph.checkpoint import MemorySaver

ERRORS AND ISSUES

PINECONE

Several problems when setting up and then seeming to be unavailable at some points.

WHISPER-FFMPG

Some libraries incompatibilities.

DATA FLOW HANDLING

Transforming and moving data around.

RESULTS

VERY IMPRESSIVE ACCURACY

Tested across several videos and 2 languages

INCONSISTENT LANGUAGE

Tried to get it to automatically answer in the user's language

EVALUATION

HELPFULNESS AND REFERENCE

Good results, as expected from my previous observations.

HALLUCINATION AND RELEVANCE

Could not finish.

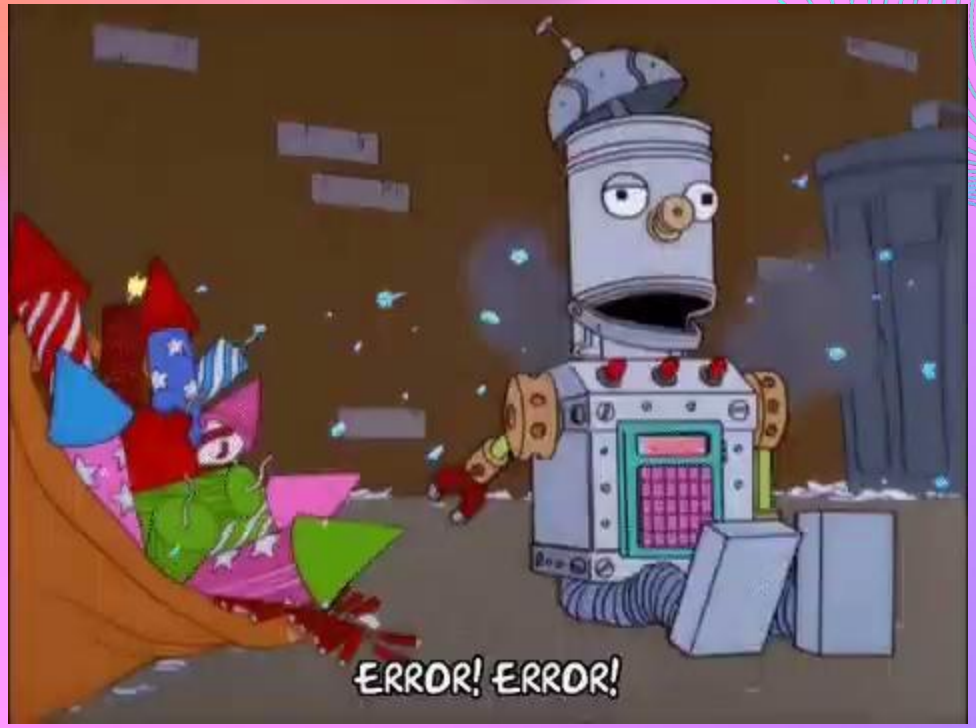
PROJECT 3

03

DEPLOYMENT

(Or not)

PROJECT 3



DEPLOYMENT

DOCKER

STREAMLIT

JUPYTER

NEXT STEPS

DEPLOYMENT

Finish and make it look nice

EVALUATION

Finish hallucination and document relevance, try another model

COMPUTER VISION

Add a second vectorstore and tool

SPEECH

Speech-to-text and text-to-speech with ElevenLabs

PERFORMANCE

Handle exceptions and userflow properly, adjust cost and time

RUNTIME CONFIG

Can be used to modify the agent's behaviour, like language, based on user actions during runtime

THANK YOU

Time for questions

PROJECT 3