FABI‘S TUBE BUDDY

FABIAN HIEBER

NOW UNDERSTAND YOUTUBE VIDEOS

Tired of not understanding a video you watch on YouTube?

Your Options:

- Watch the video again
- Use the comment section to get dumb answers
- Ask someone smarter to explain it

Or:

Ask anything about the video Fabi's Tube Buddy

HOW IS THIS EVEN POSSIBLE?

Large Language Model (LLM):
ChatOpenAI model (gpt-3.5-turbo)

Embeddings:
OpenAIEmbeddings

Vector Store:
FAISS

Transcript Processing:
YouTubeTranscriptApi

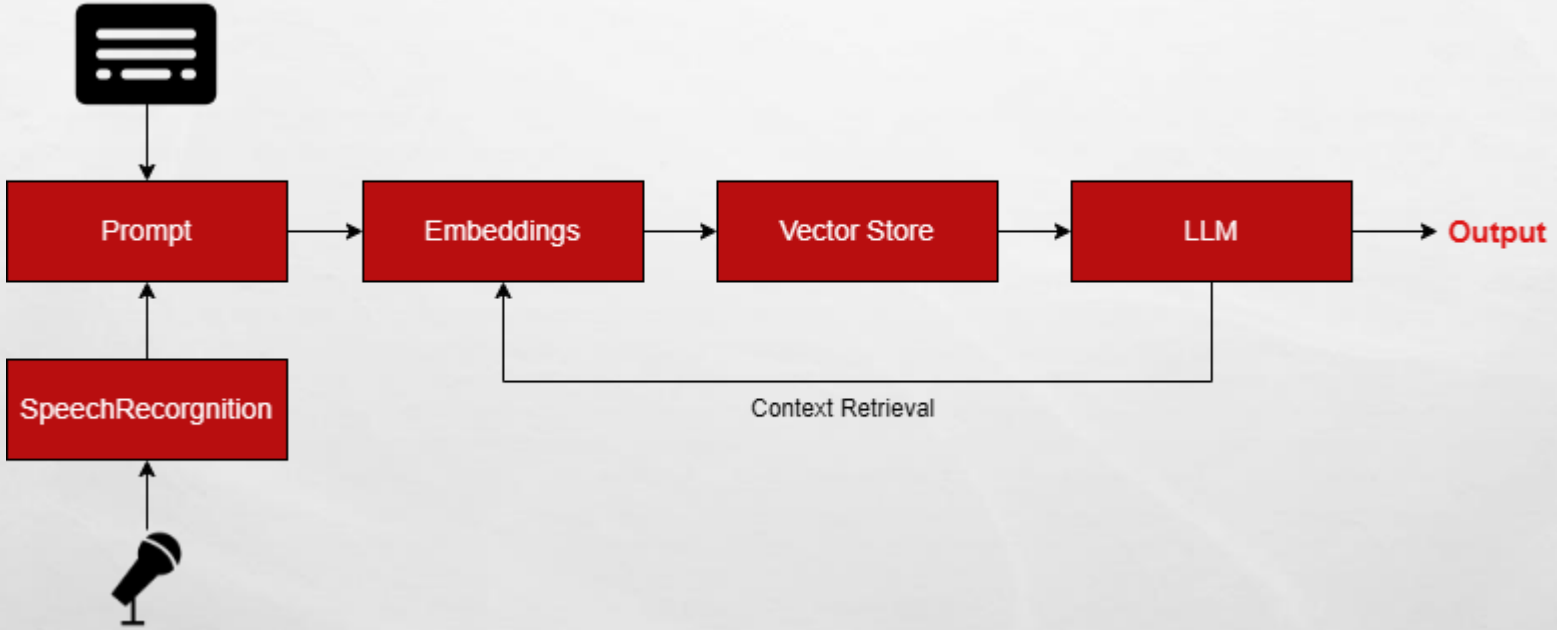
Chains:

- ConversationalRetrievalChain
  Used for user conversation with the bot
- RetrievalQA
  Used for generating initial example questions

Speech to Text:
Google Cloud Speech API

Deployment:
Flask Website

DATAFLOW


Prompt
Embeddings
Vector Store
LLM
Output
SpeechRecorgnition
Context Retrieval

DATASET

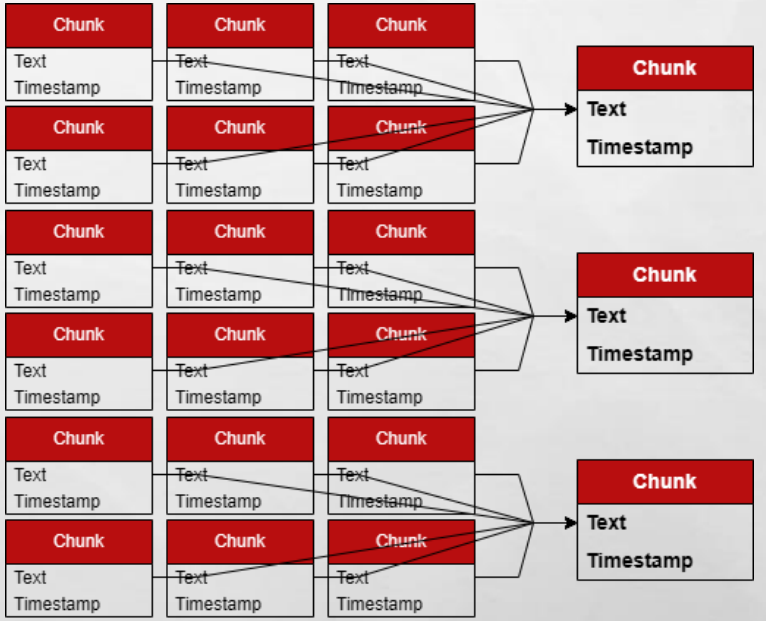
Your provided Video

YouTubeTranscriptApi:

- Transcription comes in little chunks with timestamps
- Combined in larger chunks with maximum of 1000 characters
- Preserved the start time of the chunk for metadata
- Texts cleaned

FAISS:

- OpenAIEmbeddings with texts and metadata



Q&A AGENT

LLM:

GPT-3.5-turbo

- Temperature 0.2

Memory:

ConversationBufferWindowMemory

- Save 3 messages

Chain:

ConversationalRetrievalChain

- LangSmith Callback

Vector Store:

FAISS

- Metadata: Timestamps
- Search Type: Similarity
- Documents: 4

Prompt:

- Persona of AI assistant and chatter
- Video name mentioned
- Attention drawn to Conversation
- Fallback response
- Question

EVALUATION

LangSmith:
Compared gpt-4o-mini and gpt-3.5-turbo
P50 Latency: 0.55s
P99 Latency: 2.98s

QAEvalChain:
Compared human answer to AI generated
Added false positive
8 / 8 correct ✓

Giskard:

- Performance ✓
- Robustness 2 X
- Calibration ✓
- Data Leakage ✓
- Stochasticity ✓
- Injection attacks ✓
- Hallucination & misinformation ✓
- Harmful content generation ✓
- Stereotypes ✓
- Information disclosure 1 X
- Output formatting ✓

DEPLOYMENT

Flask:

- Deployed on Heroku
- Push to deploy

Endpoints:

- Process video
- Process voice
- Questions

ChainManager:

Persists trained chains during runtime in memory

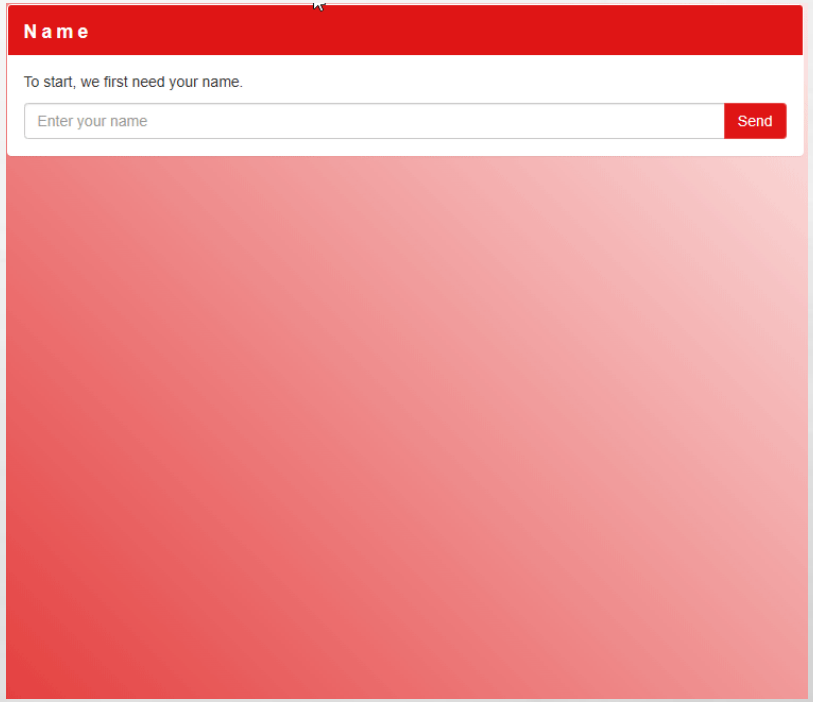
- Id
- Title
- QA_chain
- Example_chain
- Examples

DEMO

Deployment:
Flask Server on Heroku

Try it out:

https://fabis.tube-buddy.de

What you'll find?

- World fastest Chatbot
- Video Viewer
- Example questions