

ChatBot With Personality

Presented by:

Aishwarya Thorat

Dhruv Shetty

Leon Correia

Lincy Rebello

Sheetal Patnaik



Overview

- Project Background & Project Requirements
 - Technological Survey
 - Literature Survey
 - Resource Requirements
 - Project Methodology
- Project Development Workflow
- Project Plan
 - Data Process & Data Collection
 - Data Preprocessing & Data Transformation
 - Data Preparation & Big Data Statistics
 - Models & Model Prototyping
- Results
- Data Pipeline demonstration

Project Background

Growing Demand

Users want more human-like, engaging conversational AI.

Current Limits

Standard chatbots lack distinct character and feel generic.

Our Vision

Create a chatbot with multiple unique personalities.

Chosen Personalities

David Attenborough, Morgan Freeman, Elon Musk for unique styles.



Project Requirements

Substantial & Diverse Data

- **Sources:** Videos and Audios on various topics (documentaries, interviews, speeches) for Attenborough, Freeman, & Musk.
- **Quality:** High accuracy, minimal noise, rich in distinct linguistic traits.
- **Organization:** Secure, structured storage (e.g., Cloud), segmented by personality for raw data, processed text, and Q&A pairs.

Rich Contextual & Stylistic Metadata:

- Captured metadata including **tone, emotion, situational context, and speaking style** (inferred or explicitly noted during processing).
- **Personality:** Inferred and tagged (e.g., via filenames during collection, or implicitly through Q&A content source).
- *(Source/date metadata collection was secondary to capturing stylistic attributes).*

Technical Requirements

- **LLM & Embedding Models:** Access to a robust Large Language Model (LLM) for response generation.
- Effective sentence embedding model for RAG (e.g., Sentence-BERT).

Evaluation Requirements

- **Qualitative Evaluation:** Human assessment of response quality based on personality authenticity, coherence, relevance, and engagement.
- **Exploratory Data Analysis (EDA) Metrics:** Quantitative metrics from EDA (e.g., vocabulary analysis, tone distribution) to understand and verify personality characteristics in the data.

Technology Survey

Author	Title	Dataset	Model	Results
Kuhail et al. (2024)	Assessing the Impact of Chatbot-Human Personality Congruence on User Behavior: A Chatbot-Based Advising System Case	54 student interactions, 18 interviews)	Rule-based System	Trust and engagement increased, most significant for extroverted users.
Xing & Fernández (2018)	Automatic Evaluation of Neural Personality-Based Chatbots	TV Series Dialogue (Friends, Big Bang Theory)	SEQ2SEQ	Personality-based chatbots showed distinct response styles.
de Haan (2018)	Chatbot Personality and Customer Satisfaction	Survey responses	LSTM SEQ2SEQ	Customer satisfaction improved with personality-aligned responses.
Heppner et al. (2024)	Conveying Chatbot Personality through Conversational Cues in Social Media Messages	168 participants (online study)	Personality-Enhanced Chatbot	Social cues significantly influenced perceived chatbot personality

Technology Survey

Author	Title	Dataset	Model	Results
Polverini & Gregorcic (2024)	Evaluating Vision-Capable Chatbots in Interpreting Kinematics Graphs: A Comparative Study of Free and Subscription-Based Models	Kinematics graph test dataset	Large Multimodal Models (LMMs)	Best performing models: ChatGPT-4o (58.4), ChatGPT-4 (42.4), Gemini 1.5 Pro API (34.7)
Hsain & El Housni (2024)	Large Language Model-Powered Chatbots for Internationalizing Student Support in Higher Education	Scraped academic documents, PDFs	GPT-3.5-Turbo	Quality: 4.06 - 4.43, Relevance: 4.02 - 4.41, Formality: 4.24 - 4.54.
Li et al. (2024)	Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach	Public NLP datasets	RAG vs. Long-Context LLMs	Long-context models outperformed RAG but at a higher computational cost.
Sukhwal (2024)	Retrieval Augmented Generation: An Evaluation of RAG-based Chatbot for Customer Support	Support queries, knowledge base	GPT-4 with RAG	MRR: Simplified queries (With Keyword: 0.386, Without: 0.333), Original queries (With Keyword: 0.342, Without: 0.31).

Literature Survey

Author	Title	Methodology and Metrics	Results and Limitations
Raj et al.(2024)	Optimizing LLaMA models for persona-based Enterprise application	LoRA, LLaMA2-13B Accuracy: 81.2% (Legal), 77.5% (CS)	Persona improves credibility; fine-tuning causes knowledge decay.
Jain & Kumar (2024)	AI Unveiled Personalities: Profiling Optimistic and Pessimistic Attitudes in Hindi Dataset Using Transformer-Based Models	RAFT, IFT, LLaMA2-7B Accuracy: 86.86% (TF), 84.95% (Torch)	High accuracy but needs robustness for Top-K generalization.
Singh & Aarabhi (2023)	Emergent Personality Traits in Large Language Models	GPT-3.5, RAG Persona Embedding Accuracy: 82%	Personality reflected well, but neuroticism harder to control.
Kumar et al. (2024)	Examining Chatbot Role-Play and Human Perception in Learning Environments	GPT-4 Turbo, Educational Persona Engagement: 92%	Improves engagement but requires teacher oversight.
Fan et al. (2023)	How Well Can an AI Chatbot Infer Personality? Examining Psychometric Properties of Machine-Inferred Personality Scores	Juji AI Chatbot Reliability: 0.63-0.81	Strong validity but struggles with generalization.
Luo et al. (2022)	Towards a Personality AI for Robots: Potential Colony Capacity of a Goal-Shaped Generative Personality Model when Used for Expressing Personalities via Non-Verbal Behavior of Humanoid Robots	ROBOT Model Trait Capacity: 70%	Effective nonverbal personality but hardware limitations exist.
Wu et al. (2024)	Real-Time Persona Modulation in Counseling Chatbots	RAG + GPT-4 Persona Sentiment: 85%, Retention: 79%	Dynamic persona boosts trust but raises privacy risks

Literature Survey

Author	Title	Methodology and Metrics	Results and Limitations
Olabiyi et al. (2019)	A Persona-Based Multi-turn Conversation Model in an Adversarial Learning Framework	phredGAN F1 Score: 0.1692, DISTINCT-2: 24.53%	Handles multi-turn dialogues, but human evaluation needed.
Ruane et al. (2021)	User Perception of Chatbot Personality in Social Interactions	Transformer Sentiment & Behavioral Cues Trust: 91%	Chatbot friendliness enhanced, but repetitive phrasing an issue.
AlGhozali et al. (2024)	Natural Language Processing of Gemini Artificial Intelligence Powered Chatbot	Integrated NLP + Gemini AI	Context-aware responses but needs further validation
Kuhail et al. (2024)	Assessing the Impact of Chatbot-Human Personality Congruence on User Behavior: A Chatbot-Based Advising System Case	FFM Model Trust: 0.97 (Conscientious), 0.96 (Extroverted)	Trust & engagement improved but needs a larger sample size.
Hasan et al. (2023)	SAPIEN: Affective Virtual Agents Powered by Large Language Models	TTS & Human Evaluation	Realistic responses, but ethical concerns on aggression.
Zhao et al. (2024)	Retrieval-Augmented Generation (RAG) for Persona Consistency	RAG + GPT-4 Persona Embedding Stability: 90%	Persona stability improved, but retrieval overhead increases latency.

Resource Requirements

1 Key Software & Model Ecosystem

- Critical Libraries & Frameworks
- Language & Embedding Models
- Development & Collaboration Suite
- Transcription Tools

2 Computing & Infrastructure Needs

- GPUs
- **Cloud Platform (Essential): Google Cloud Platform (GCP)** for storage (e.g., Google Cloud Storage) and pipeline

3 Potential Project Risks

- Data Integrity & Bias
- Personality Emulation & Response Quality
- Technical & Dependency Risks

Mitigation Strategies

- Data Management & Validation
- Iterative Refinement & Evaluation
- Robust Technical Implementation & Monitoring

4 Project Timeline & Effort

- Thorough data collection, vetting, and cleaning for three distinct personalities.
- Preprocessing and meticulous Q&A pair generation.
- RAG pipeline construction and model integration.
- Iterative prompt refinement and extensive testing.
- Comprehensive evaluation and analysis.

Resource Requirements

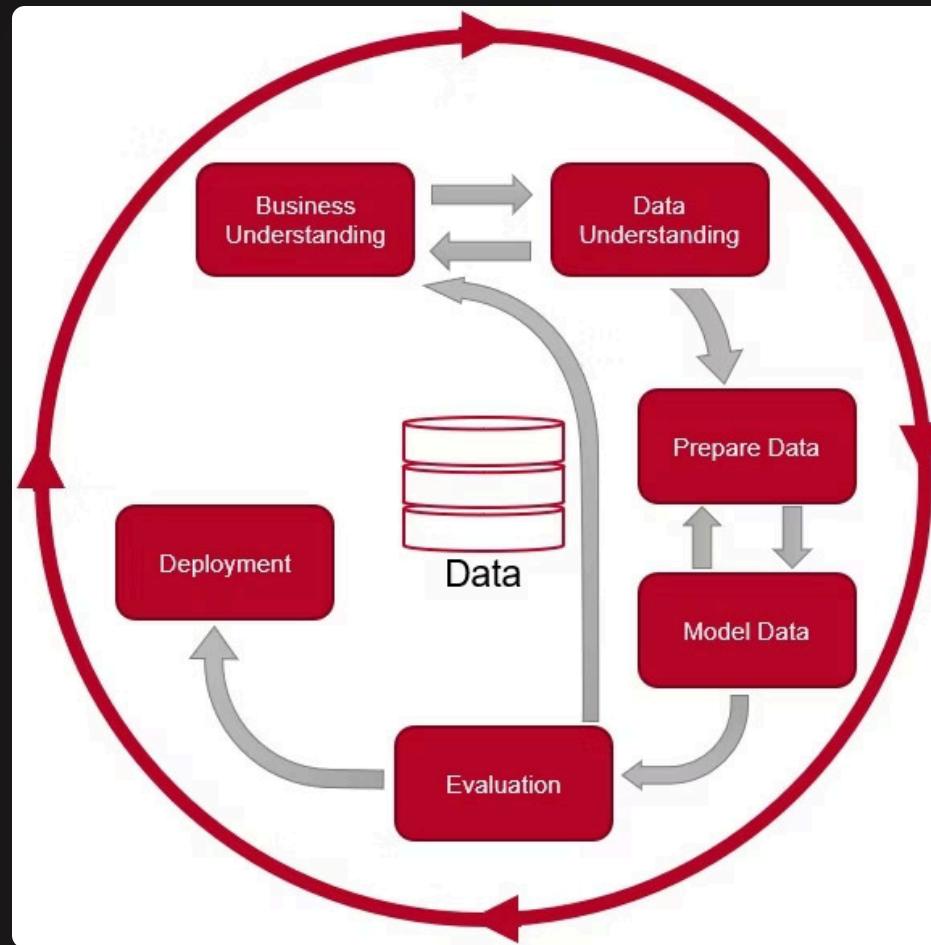
Function	Resource Type	Resource	Time Duration	Cost Estimate in USD
Model Training and RAG system deployment	Hardware	GCP Compute Engine (n1-standard-8)	2 months	\$554.8 (\$277.40 per month)
Data Storage	Hardware	GCP Cloud Storage - Standard-120 GB	3 months	\$8.28 (\$0.023 GB per month)
Data Storage for Backup	Hardware	GCP Cloud Storage - Archival-120 GB	3 months	\$0.9 (\$0.0025 per GB per month)
Google Colab Pro	Software	Access to GPU (for faster training of GPT-4, LLaMA, etc)	3 months	\$30 (\$10/month)
ML/RAG Model Development	Software	Python, Pandas, NumPy, Scikit-Learn	2 months	Free (Open Source)

Resource Requirements

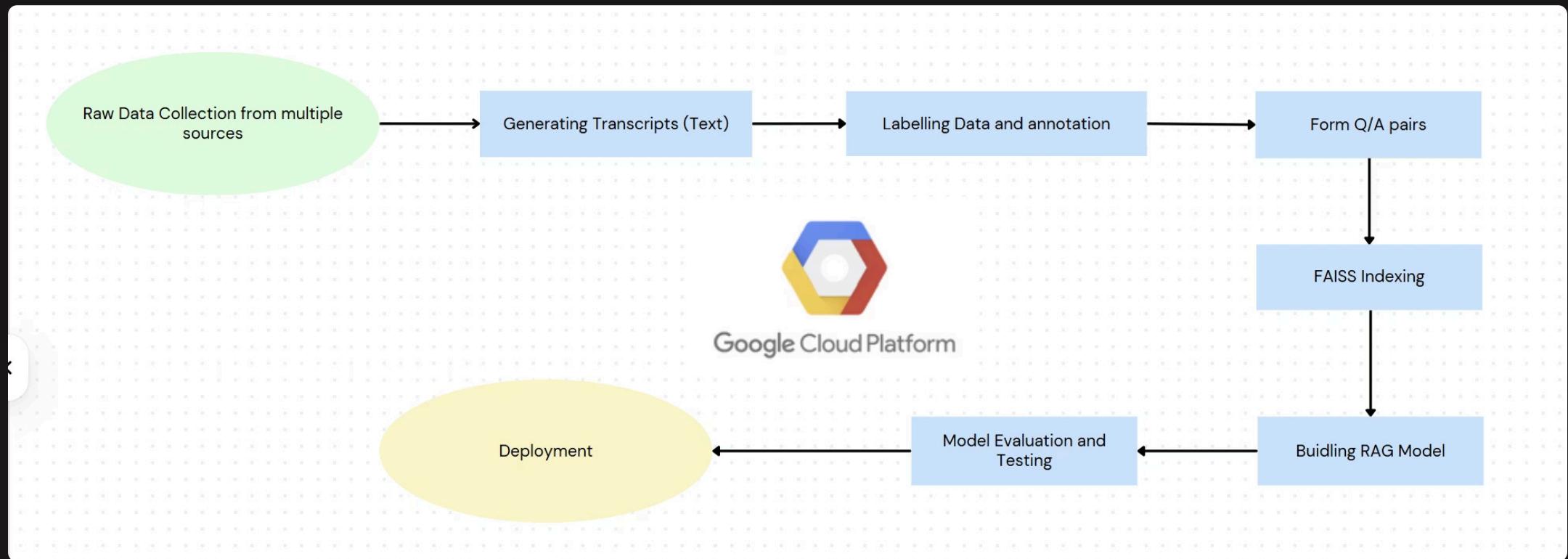
Function	Resource Type	Resource	Time Duration	Cost Estimate in USD
Miscellaneous Software Licenses (e.g., Whisper OpenAI, TensorFlow)	Software	TensorFlow, Keras, PyTorch	2 months	Free (Open Source)
Cloud SDK Management	Software	Google Cloud SDK	3 months	Free (Open Source)
Version Control	Tools	Git	3 months	Free (Open Source)
Task Management and Documentation	Tools	JIRA, Google Docs	4 months	Free (Educational Use)
Cloud-based Code Execution	Tools	Google Colab, JupyterLab	3 months	Free with GCP (limitations may apply)

Project Development Methodology

CRISP-DM

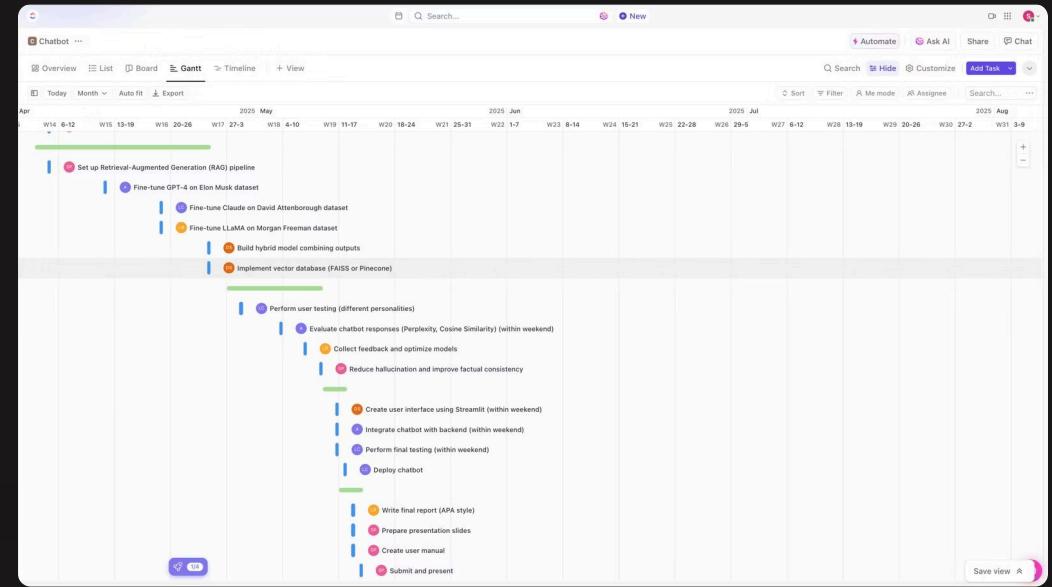
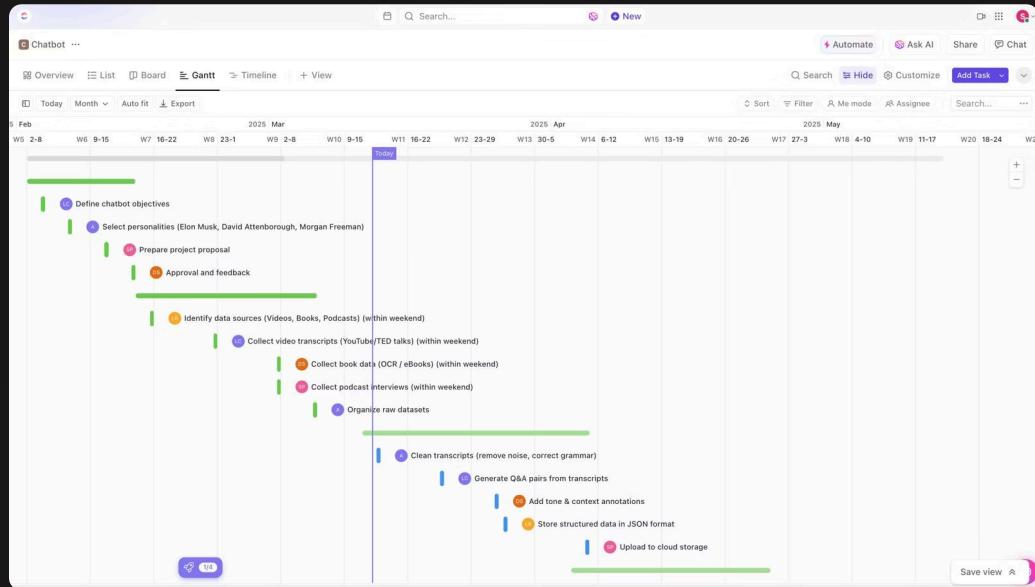


Project Development Workflow



Project Plan

Gantt chart



Project Plan

Chatbot with Personality Workspace visible Board Table

Epic Cards Gantt Power-Ups Automation Filters

SP LR A

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation and Optimization	Documentation
Scrum 1 Completed ✓ Develop Project Objective Feb 6 0/5	Scrum 5 Completed ✓ Data Collection and Video Sources Feb 28 0/5	Scrum 9 Transcript Preprocessing Feb 0/5	Scrum 13 Video Data + Model Development Feb 0/5	Scrum 17 Model Evaluation Feb 1 0/5	Scrum 22 Project Report Feb 1 0/5
Scrum 2 Completed ✓ Planning and Milestone Feb 12 0/5	Scrum 6 In Progress ✓ Data Collection from books and interviews Feb 3 0/5	Scrum 10 Q&A Generation and Synthesis Feb 0/5	Scrum 14 Book Data + Claude Model Development Feb 0/5	Scrum 18 Fine-Tuning Feb 1 0/4	Scrum 23 Final Presentation Feb 1 0/4
Scrum 3 Completed ✓ Define goals and Success Feb 18 0/4	Scrum 7 In Progress Data Exploration and Analysis Mar 12 0/4	Scrum 11 Data Preparation for RAG Feb 0/3	Scrum 15 Interview Data + LLaMA Model Development Feb 1 0/5	Scrum 19 Error handling and debugging Feb 1 0/4	+ Add a card
Scrum 4 Completed ✓ Communication and Feedback Feb 24 0/4	Scrum 8 Data Quality Assurance Feb 0/4	Scrum 12 Additional preprocessing Tasks Feb 0/4	Scrum 16 Hybrid Model Development Feb 2 0/5	Scrum 20 User Feedback Feb 1 0/7	Scrum 21 Final Evaluation Feb 0/3
+ Add a card	+ Add a card	+ Add a card	+ Add a card	+ Add a card	+ Add a card

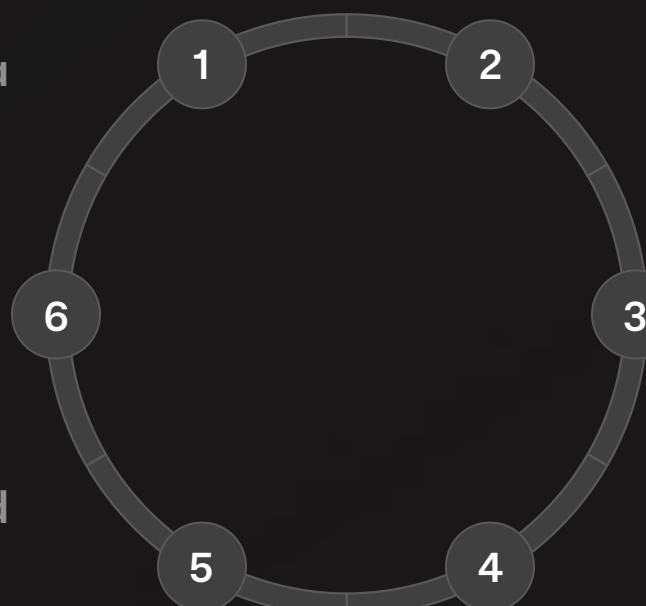
Data Process

High-quality, deeply representative data is paramount for authentically emulating David Attenborough, Morgan Freeman, and Elon Musk.

CRISP-DM methodology for data preparation

Personality-focused knowledge base for chatbot

Logical structure for streamlined retrieval



Convert video to audio for transcription

Generate transcripts; store by type

Chunking: Split files for efficiency

Data Collection

Dataset Source:

The project collects data from Videos and Audios including Interviews/Podcasts to capture the communication styles of Elon Musk, David Attenborough, and Morgan Freeman.

All datasets are stored in Google Cloud Storage, structured for seamless integration with the RAG chatbot system.

Dataset Details:

Data Scale: ~40,000 to 100,000 words per personality from video transcripts, books, and podcasts.

Total Size: 10-15 GB, including text files and structured documents

Storage: Organized in Google Cloud Storage with separate folders for Video, Audio and Transcripts

Context: Derived from publicly available sources to ensure diverse and natural chatbot interactions.

Metadata: Includes Tone, emotion, situation and speaking style.

Data Collection

Data Type	Source	Tools/Packages	Focus
Video Data	YouTube, TED Talks, documentaries, public speeches	OpenAI Whisper, YouTube DL	Conversational tone, spontaneous speech, intonation
Book Data	Books on/by Musk, Attenborough, Freeman	OCR (Tesseract), PDF parsers	Formal and structured language
Interview/Podcast Data	Podcasts, Q&A panels, published interviews	OpenAI Whisper, YouTube DL	Conversational exchanges, interactive Q&A dynamics

Once collected, the data will be processed using APIs to generate structured Question-Answer pairs with Context from interviews, books, and videos in an automated pipeline. This ensures realistic, context-aware responses while preserving each personality's unique style. The chatbot will use this structured format to deliver natural and engaging conversations.

Data Preprocessing

Transcription Review

Correcting errors from automated speech-to-text.

Noise Removal

Attenborough: informative, wondrous;
Freeman: calm, authoritative; Musk: direct, assertive.

Speaker Identification

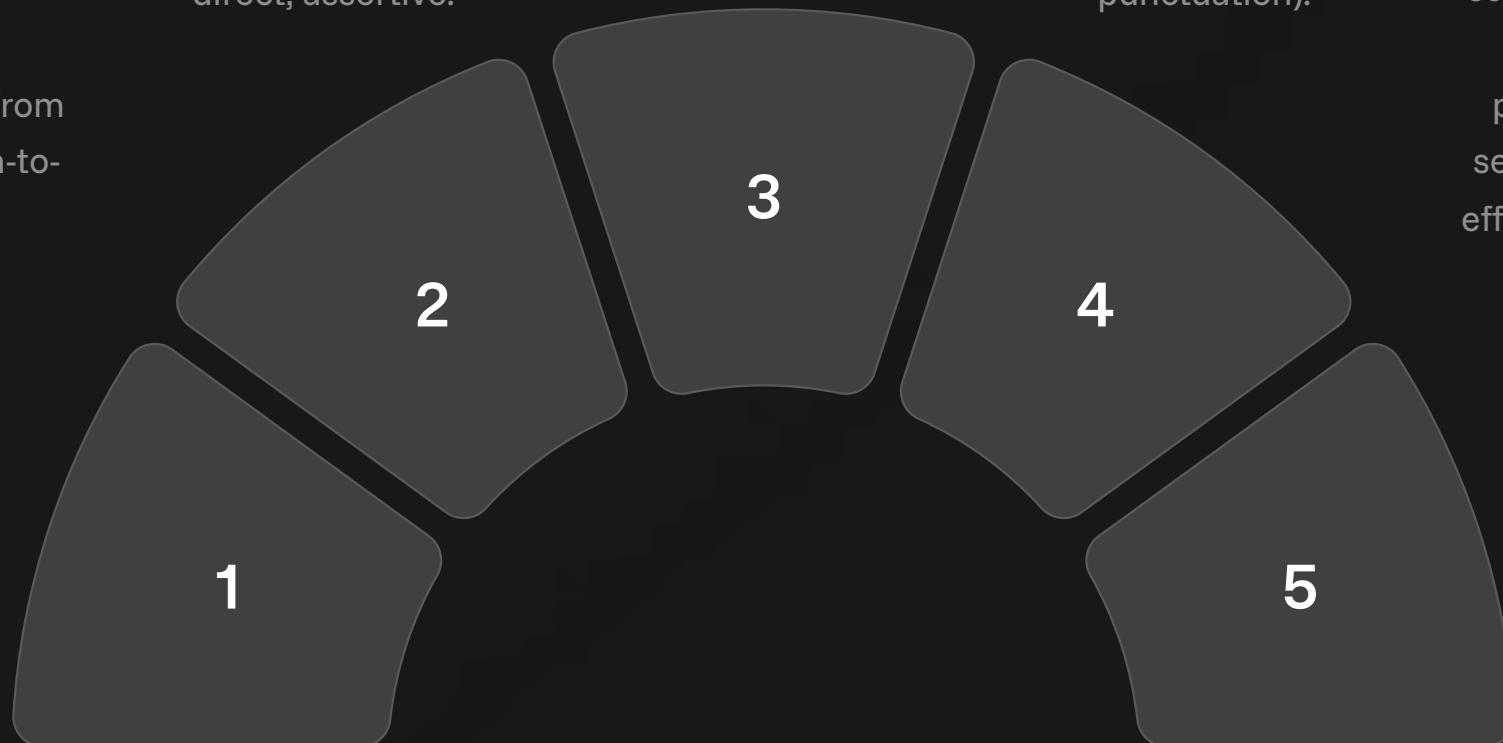
Ensuring text is correctly attributed, especially in multi-speaker interviews (less critical for pure narrations).

Normalization

Standardizing text (e.g., lowercase, handling contractions, consistent punctuation).

Segmentation

Breaking down long documents/transcripts into smaller, semantically coherent chunks (e.g., paragraphs, logical sections) – crucial for effective RAG retrieval.



Data Transformation

Personality-Specific Q&A Generation

1

For Interview-Style Transcripts:

- Pre-labeled speaker turns (Interviewer/Musk).
- Gemini prompted to leverage/refine existing Q&A structure or generate new pairs, maintaining speaker distinction.
- **Result:** Dynamic, interview-style Q&A.

2

Narration & Monologues (e.g., Attenborough, Freeman):

- Monologues/narrations (Narrator: Attenborough/Freeman) labeled.
- Distinct Gemini prompt: Identify key info -> Formulate relevant Q -> Extract/synthesize A from text.
- **Result:** Informative, stylistic Q&A reflecting narration essence.

3

Enrichment:

All Q&A pairs tagged with personality & stylistic metadata (tone, emotion, etc.) for effective RAG

4

Final Output:

Robust, personality-aligned Q&A datasets forming a core part of each character's chatbot knowledge base.

Data Preparation

1 Audio Extraction (for Video)

Utilized FFmpeg to isolate audio tracks from video content, preparing it for transcription.

2 Accurate Transcription

Employed OpenAI Whisper to generate textual transcripts from audio.

Transcripts meticulously organized and stored by personality and content type (e.g., "Attenborough_01_Transcript," "Musk_Audio__01Transcript").

3 Strategic Content Chunking

Large text files (especially those derived from >5 min audio/video segments) were intelligently segmented.

Rationale: Smaller, semantically coherent chunks improve retrieval efficiency and relevance for the RAG system.

4 Logical Structuring & Indexing using FAISS

Organized data to maintain a clear, logical structure that mirrors conversational flow or thematic units.

Contextual metadata (source, date, Q&A pairs generated earlier) was tightly linked to each chunk.

Big Data Statistics

Fueling Our Personalities- Data Processed (approx. per personality):

This extensive and refined data foundation enables distinct, context-aware, and stylistically accurate responses for each personality.

Initial Video Collection: ~11 hours ~ 4-5 GB

Source: Documentaries, interviews, speeches for rich conversational data.

Derived Audio Data (for transcription): ~800 MB

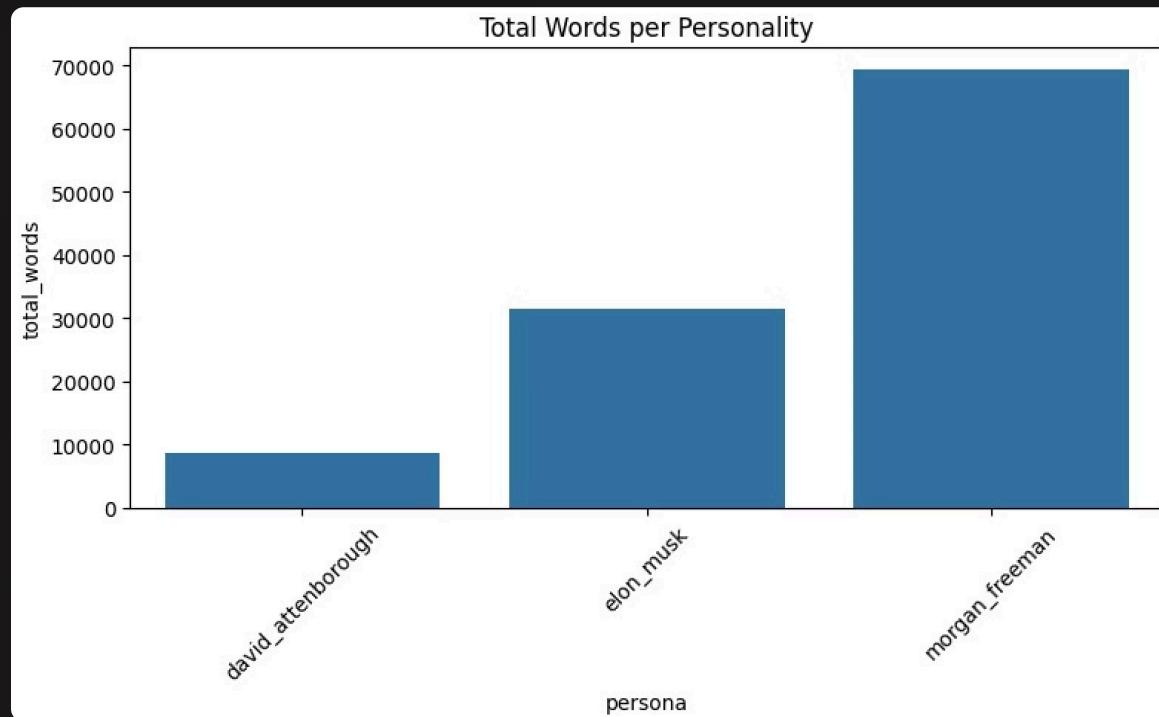
Process: Extracted from video sources for speech-to-text conversion.

Final Text Corpus (Transcripts, Q&A): ~250 MB

Purpose: Direct input for embedding, RAG knowledge base, and Q&A generation.

Big Data Statistics

Dataset Overview & Textual Volume

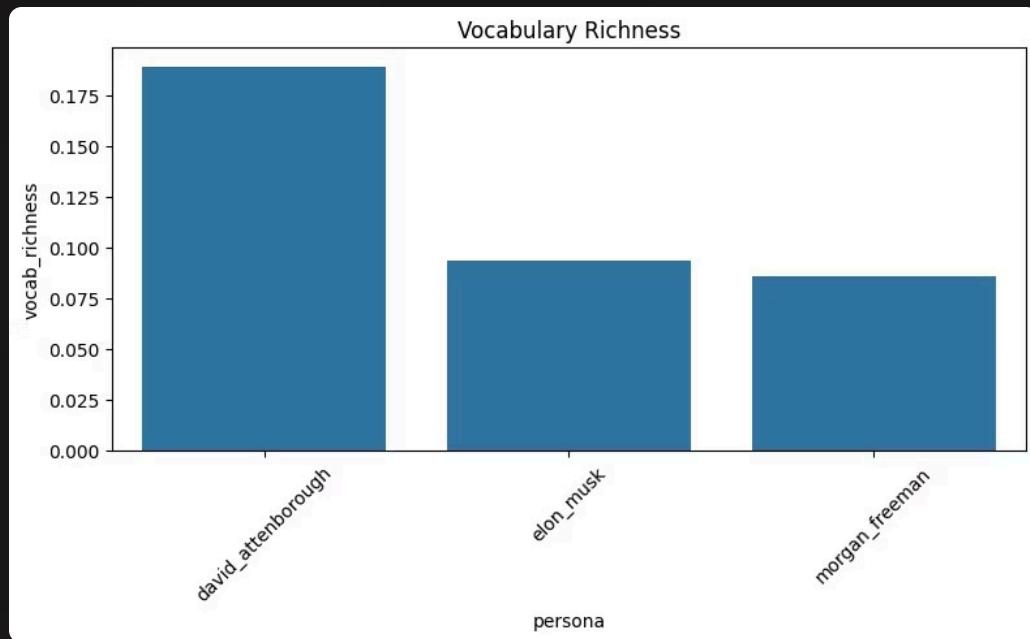


The substantial volume of text processed for each personality, forming the basis for statistical analysis.

The variation in word count reflects the nature and availability of extensive public discourse for each individual. Even the smallest dataset (Attenborough) represents a significant corpus for personality analysis.

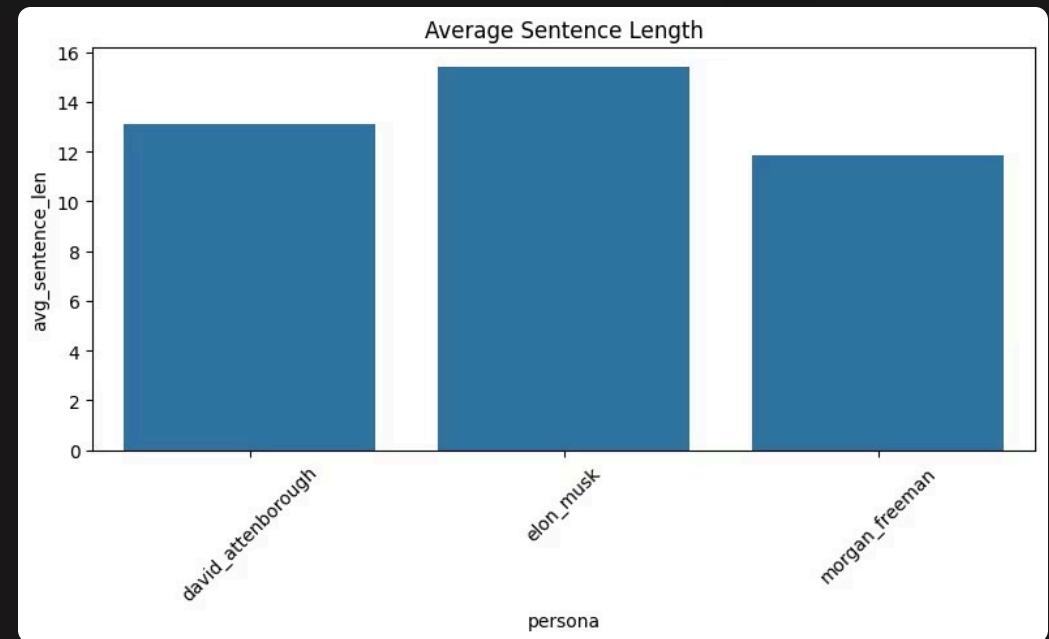
Big Data Statistics

Linguistic & Stylistic Markers



Analyzing linguistic patterns within the large datasets to identify statistical differences in communication styles

Attenborough's extensive use of specific, descriptive language across diverse natural world topics contributes to a richer vocabulary per volume of text.



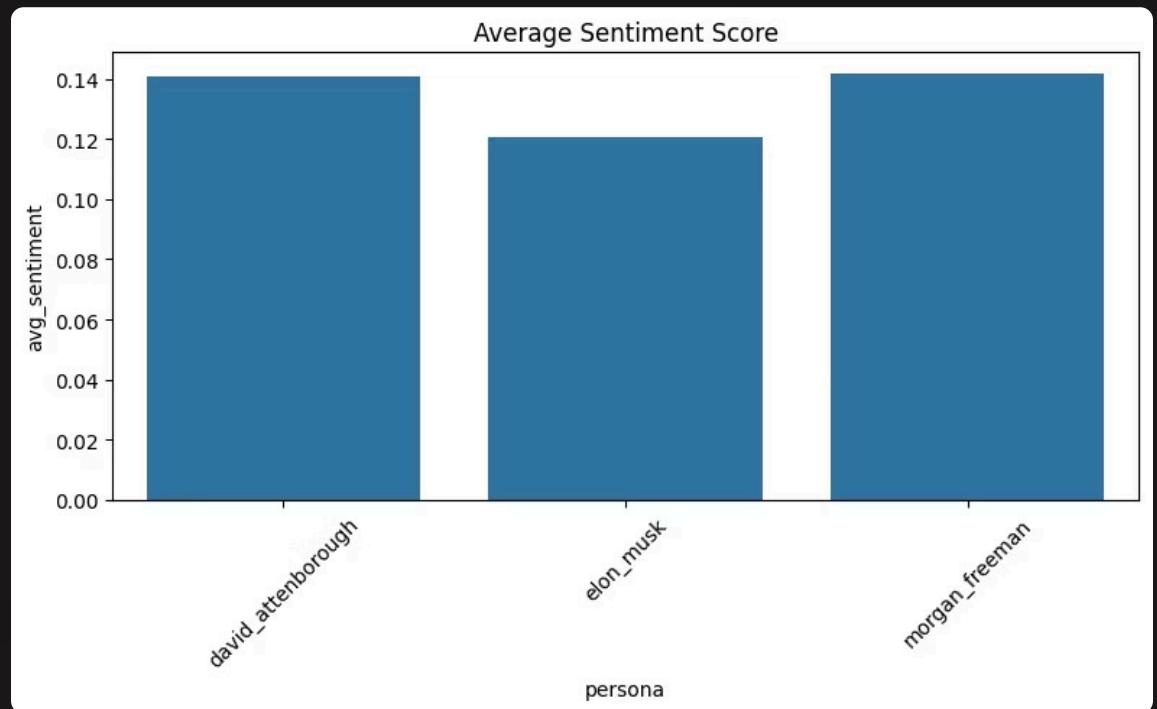
These statistical differences hint at varying complexities in sentence structure and delivery styles within their respective large text corpora.

Big Data Statistics

Sentiment Profile & Thematic Concentrations

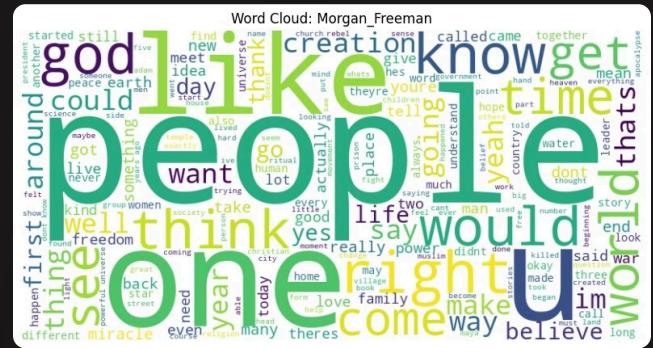
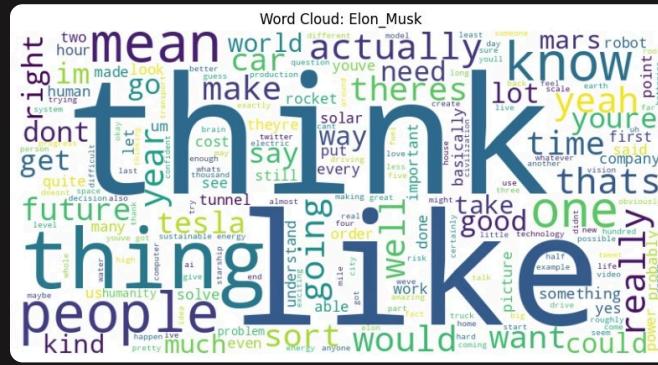
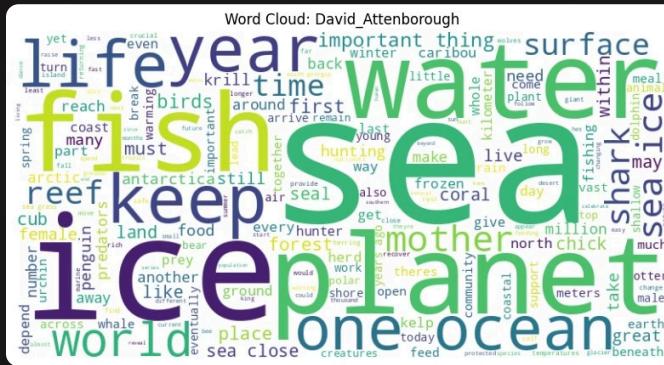
The substantial volume of text processed for each personality, forming the basis for statistical analysis.

The variation in word count reflects the nature and availability of extensive public discourse for each individual. Even the smallest dataset (Attenborough) represents a significant corpus for personality analysis.



Big Data Statistics

Dominant Thematic Concentrations



- **David Attenborough:** Statistically significant occurrences of terms like "water," "sea," "ice," "planet," "life" highlight his thematic focus on nature.
 - **Elon Musk:** High frequency of "think," "like," "Tesla," "Mars," "rocket," "future" statistically points to his discourse on technology, innovation, and future aspirations.
 - **Morgan Freeman:** Prevalent terms such as "people," "like," "God," "one," "life," "world" statistically underscore his narrative focus on human experience and broader philosophical themes.

Machine Learning Models

Model	Strengths	Weaknesses	Innovation	Justification
Google Gemini	Strong reasoning & instruction-following. Good contextual understanding. Integrates well with Google's cloud ecosystem.	Performance varies by version. Output can be more factual by default, requiring careful prompting for stylistic outputs.	Native multimodality. Advances in responsible AI. Efficient for structured data generation tasks.	Considered for tasks requiring strong instruction adherence, contextual reasoning, and native integration within Google Cloud Platform environments.
OpenAI GPT-4 Series	State-of-the-art generation quality, creativity, and coherence. Large context window.	Higher API costs. Proprietary model (less transparency). Potential latency.	Pushes LLM scale & general AI intelligence. Advanced multimodal capabilities.	Valued for applications demanding cutting-edge text generation, complex problem-solving, and highly creative or nuanced outputs.
Meta LLaMA Series	Leading open-source performance. Available in various sizes. Customizable via fine-tuning.	Larger models demand significant self-hosting resources. Fine-tuning requires expertise & data.	Drives open-source LLM access & research. Enables efficient large model training.	Suitable for projects prioritizing open-source flexibility, model customization through fine-tuning, and control over deployment infrastructure.
Claude 3.5 Sonnet	High proficiency in specific domains. Strong emphasis on ethical communication & safety. Good contextual understanding, long context window.	Conservative outputs may limit creativity in highly open-ended scenarios. Resource demands for larger Claude versions can be high.	Prioritizes contextual comprehension, safe/helpful responses, and extended context for deeper understanding. Strong focus on enterprise-grade reliability.	Preferred for applications where safety, ethical considerations, and reliable, nuanced contextual understanding are paramount, especially in enterprise settings.
Mistral-7B-Instruct	Uses an open-source instruction-tuned model for generating responses based on internal training. Fully customizable, cost-effective, and deployable locally or on any cloud.	Simpler architecture without retrieval. Performance hinges on model size, training data, and prompt quality. May hallucinate on specific factual queries.	Excels at generating coherent, fluent responses across diverse tasks without external retrieval. Great for ideation, summarization, and code.	Suitable for applications focused on cost, flexibility, data privacy, and offline or fine-tuned deployment , where external retrieval isn't essential.

Machine Learning Model Development & Prototyping

Knowledge Base Construction (FAISS Index):

- **Embedding:** All cleaned text data and generated Q&A pairs were converted into dense vector representations using *sentence bert modek all0minilm-l6-v2*
- **Initial FAISS Build:** A FAISS index was constructed for efficient similarity search on these embeddings.

Crucial Correction & Refinement:

- **Identified Issue:** Realized that explicit personality labels were initially missing from some data segments intended for the FAISS index.
- **Rectification:** Systematically patched and added the correct personality labels to the cleaned JSON data sources.
- **Final FAISS Embedding:** The FAISS index was then rebuilt/updated using this corrected and accurately personality-labeled data, ensuring precise, personality-specific document retrieval.

LLM Integration

- Selected Gemini 1.5 & integrated with the RAG pipeline to receive retrieved context and user queries.

Machine Learning Model Development & Prototyping

Personality-Driven Prompt Engineering:

- Crafted distinct system prompts for David Attenborough, Morgan Freeman, and Elon Musk, informed by EDA and stylistic metadata.
- Iteratively refined prompts based on response quality and personality authenticity.

Interactive Prototype (Streamlit):

- Developed an initial web application using Streamlit for real-time interaction and testing.

Sample Chat (Deployed on streamlit cloud)



Persona-Based Q&A Chatbot

Ask your question:

You are elon musk. What do you think about traffic in San Francisco?

Response:

San Francisco traffic? Honestly? It's a joke. A slow-motion, stop-and-go, soul-crushing joke. But you know what? It's *solvable*. We're building tunnels, baby! Three-dimensional networks of tunnels. Think of it: 130 mph electric skates whisking you beneath the gridlock. No more sitting in traffic, staring at brake lights. It's the only sane solution. Flying cars? Please. Noise, safety concerns, airspace congestion... it's madness. My tunnels are the future. Efficient, fast, and frankly, way cooler than anything anyone else is proposing. We're going to revolutionize urban transportation, and San Francisco will be a shining example. It's a no-brainer.

Show Retrieved Q&A Context



Results - Human Validation

Personality Alignment Scores

Evaluator	David Attenborough	Morgan Freeman	Elon Musk
Aishwarya	4.2	4.5	4.7
Dhruv	4.0	4.5	4.8
Leon	4.3	4.7	4.5
Lincy	4.0	4.5	4.6
Sheetal	4.1	4.5	4.9
Average Score	4.10	4.46	4.70

Next Steps & Conclusion

Results & Evaluation

Assess chatbot on authenticity, coherence, and engagement.

Future Work

Improve data, and enhance user experience.

Showcase chatbot switching between personalities in real-time.

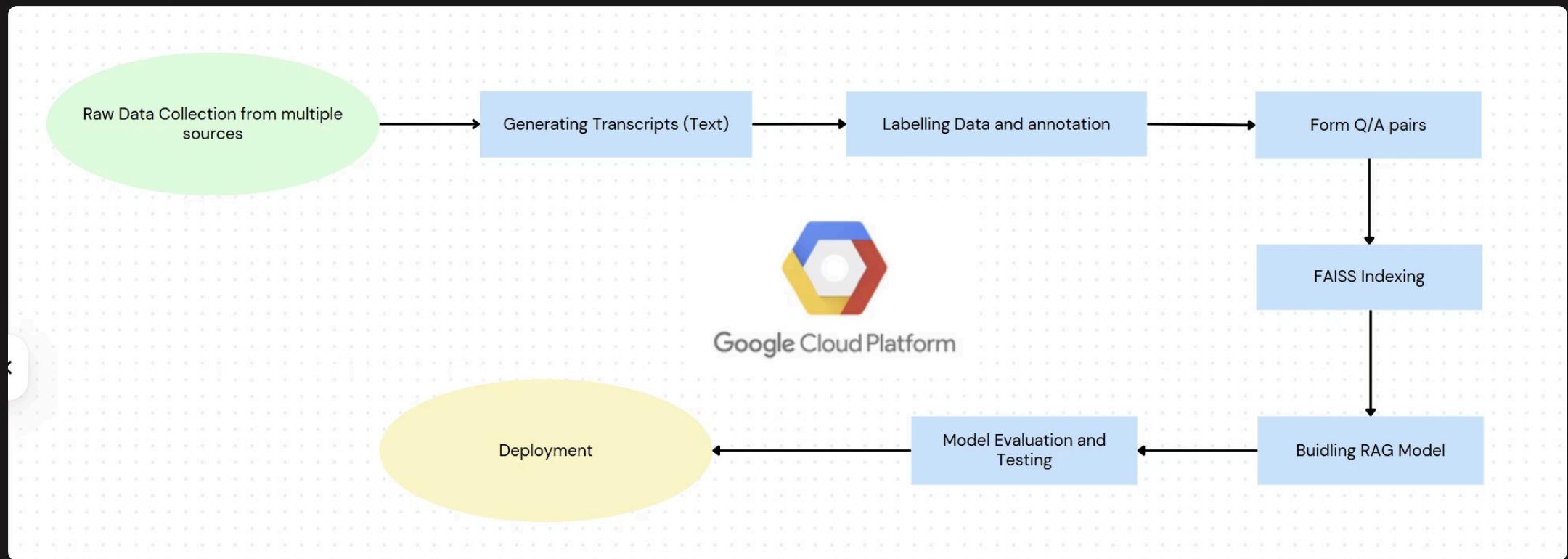
Q&A

Open floor for questions and feedback.

PROJECT DATA PIPELINE DEMONSTRATION

Project Data Pipeline

This section demonstrates the automated data pipeline we implemented to ingest, transform, and prepare the data for each personality (David Attenborough, Morgan Freeman, Elon Musk) into a queryable knowledge base. The following steps were executed for each personality.



Cloud Functions

1 Cloud function 1

Triggers on video upload and automatically converts uploaded video files into audio (.wav)

2 Cloud function 2

Transcribes audio files into text using whisper

3 Cloud function 3

Labels speakers and generates Q&A pairs from transcripts and saves the faiss indexed pairs

The screenshot shows the Google Cloud Platform Cloud Run Services page. The top navigation bar includes Cloud Run, Services, Deploy container, Connect repo, Write a function, and Manage custom domains. Below the navigation, there are tabs for Services and Jobs, with Services selected. A descriptive text block explains that a service exposes a unique endpoint and automatically scales the underlying infrastructure to handle incoming requests, and provides instructions to Deploy a container image, source code or a function to create a service. The main section is titled 'Services' and contains a table with the following data:

Name	Deployment type	Req/sec	Region	Authentication	Ingress	Recommendation	Last deployed	Deployed by
function-1	Function	0	us-central1	Require authentication	All	—	46 minutes ago	work.leoncorreia@gmail.com
function-3	Function	0	us-central1	Require authentication	All	—	Just now	Cloud Run functions
transcribe-audio-to-text	Function	0	us-central1	Require authentication	All	—	25 minutes ago	Cloud Run functions

Data Ingestion & Initial Processing-Function 1

1

Source

Video files (.mp4, .mkv, etc.) containing interviews, speeches, and narrations for each personality, stored on Google Drive.

2

Video-to-Audio Conversion

Each video file was converted into a high-quality WAV audio file.

3

Audio Transcription (Speech-to-Text)

Audio files were chunked into manageable segments (e.g., 5-minute chunks using pydub) to optimize processing and handle large files. Each audio chunk was transcribed into raw text.

Bucket details

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	Soft delete

Objects Configuration Permissions Protection Lifecycle Observability [New](#) Inventory Reports Operations

Folder browser Buckets > chatbot-data-pipeline > video_data

Name	Size	Type	Created
attenborough_01.mp4	124.7 MB	video/mp4	11 May 2025, 1
attenborough_02.mp4	122.8 MB	video/mp4	11 May 2025, 1
elon_01.mp4	166.8 MB	video/mp4	11 May 2025, 1
elon_02.mp4	111.5 MB	video/mp4	11 May 2025, 1
elon_03.mp4	66.1 MB	video/mp4	11 May 2025, 1
freeman_01.mp4	140.9 MB	video/mp4	11 May 2025, 1
freeman_02.mp4	116.9 MB	video/mp4	11 May 2025, 1

Bucket details

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	Soft delete

Objects Configuration Permissions Protection Lifecycle Observability [New](#) Inventory Reports Operations

Folder browser Buckets > chatbot-data-pipeline > audio_data

Name	Size	Type	Created
attenborough_01.wav	566.6 MB	audio/wav	11 May 2025, 21:34:12
attenborough_02.wav	523.4 MB	audio/wav	11 May 2025, 21:33:59
elon_01.wav	579.5 MB	audio/wav	11 May 2025, 21:34:26
elon_02.wav	702.8 MB	audio/wav	11 May 2025, 21:37:05
elon_03.wav	432.2 MB	audio/wav	11 May 2025, 21:36:12
freeman_01.wav	382.6 MB	audio/wav	11 May 2025, 21:37:29
freeman_02.wav	532.4 MB	audio/wav	11 May 2025, 21:36:55

Data Transformation- Function 2

1

2

3

Speaker Identification & Labeling (for interview/dialogue data)

- Transcript chunks were processed by Gemini using carefully engineered prompts (e.g., instructing it to identify "Interviewer" vs. "Elon Musk" based on conversational patterns).

Question & Answer (Q&A) Pair Generation with Contextual Metadata

- Labeled transcript segments were fed to Gemini.
- Prompts guided Gemini to generate 2-3 relevant Q&A pairs *from the provided text* and to also extract contextual metadata (tone, emotion, situation, speaking style) for each Q&A.

Structuring & Consolidating Q&A Data:

- Extracted individual questions, answers, and metadata fields.
- Standardized the format and added a "personality" tag (e.g., "elon_musk").

Name	Size	Type
labeled_attenborough_01_chunk_0...	1.3 KB	text/plain
labeled_attenborough_01_chunk_1...	1.9 KB	text/plain
labeled_attenborough_01_chunk_1...	1.6 KB	text/plain
labeled_attenborough_01_chunk_2...	1.3 KB	text/plain
labeled_attenborough_01_chunk_3...	1.8 KB	text/plain
labeled_attenborough_01_chunk_4...	1.7 KB	text/plain
labeled_attenborough_01_chunk_5...	1.2 KB	text/plain
labeled_attenborough_01_chunk_6...	2 KB	text/plain
labeled_attenborough_01_chunk_7...	1.1 KB	text/plain
labeled_attenborough_01_chunk_8...	1.7 KB	text/plain

Data Warehouse/Data Knowledge Base Construction - Function 3

1

2

Text Embedding for Q&A Pairs

- For each Q&A pair in the consolidated JSON, the question and answer text were concatenated.
- This combined text was then converted into a dense vector embedding (384 dimensions).

Vector Indexing & Storage (Knowledge Base)

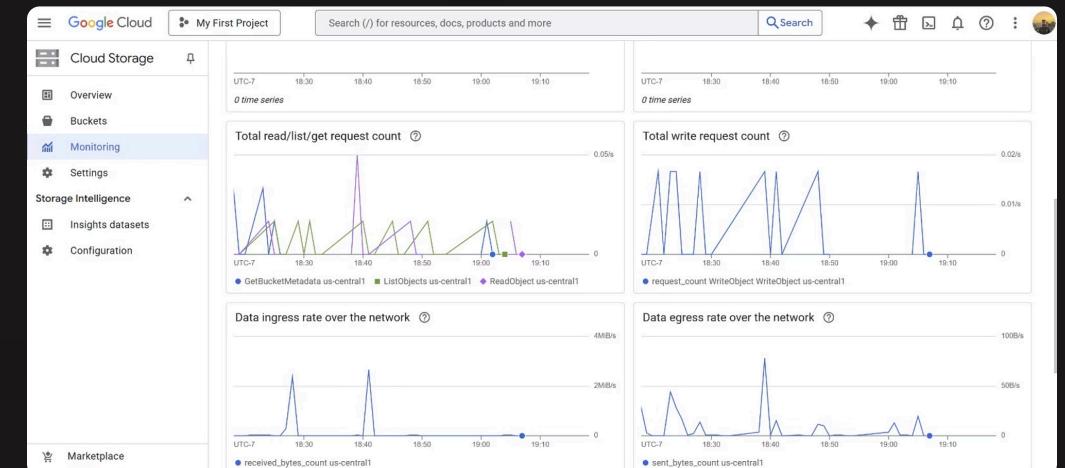
- All generated Q&A embeddings for a personality were added to a FAISS index.
- A separate JSON lookup file was created to map FAISS index IDs back to the original Q&A text, personality, and metadata.

Objects	Configuration	Permissions	Protection	Lifecycle	Observability	New	Inventory Reports	Operations
Folder browser					Buckets > chatbot-data-pipeline > qa_index	Filter		
chatbot-data-pipeline					Create folder	Upload	Transfer data	Other services
audio_data/								
labeled_transcripts/								
logs/								
qa_index/								
qa_metadata/								
transcripts/								
video_data/								

Data Pipeline Monitoring and Management

- GCP Cloud Logging, shown via Cloud Run, actively monitors pipeline execution details and flags operational errors for rapid diagnosis.
- Cloud Monitoring dashboards for services like Cloud Storage help detect performance bottlenecks and verify consistent data throughput, supporting data integrity.
- This integrated GCP oversight provides actionable insights for optimizing resource utilization and maintaining a seamless, reliable data flow.

This screenshot shows the 'Logs' tab for a Cloud Run service named 'function-1'. The logs are displayed in a table with columns for 'Severity', 'Timestamp', and 'Summary'. The logs show various audit events and startup probes. For example, there are entries for 'CreateInternalService' and 'Default STARTUP TCP probe succeeded after 1 attempt for container "worker" on port 8080'. The log table has a 'Filter' search bar at the top and a 'Logs' dropdown menu.



Live Demonstration

We've successfully deployed our chatbot on **Streamlit Cloud via Github**

Thank you

Any Questions?