TITLE

1 Generative AI-Based Multimodal Interview Coach
2 **An Intelligent Multimodal Interview Coaching System Using LLMs**
3 **AI-Driven Multimodal Interview Evaluation and Feedback System**

Step-by-Step Implementation with Tools Used

Multimodal Resume & Interview Coach

Step 1: User Input Collection

What we do:

The system first collects inputs from the user.

Inputs:

- Resume in **PDF format**
- Mock interview recording (**audio or video**)

Why:

These inputs represent the same information a human interviewer would have.

Tools Used:

- Streamlit (for file upload interface)

Step 2: Resume Text Extraction (Text Modality)

What we do:

The uploaded resume PDF is read and converted into plain text.

Why:

AI models cannot directly understand PDFs; text extraction is required for analysis.

Tools Used:

- pdfplumber (PDF text extraction)
- Python text preprocessing (cleaning)

Output:

- Raw resume text

Step 3: Resume Information Understanding

What we do:

Important information is extracted from the resume such as:

- Technical skills
- Projects
- Experience

Why:

This information provides context to evaluate interview answers.

Tools Used:

- NLP techniques
- Large Language Model (GPT / LLaMA) with prompt-based extraction

Output:

- Structured resume features

Step 4: Audio Extraction from Interview Video (if video is provided)

What we do:

If the interview is uploaded as a video, the audio track is extracted.

Why:

Speech analysis requires clean audio input.

Tools Used:

- OpenCV / FFmpeg

Output:

- Audio file (.wav)

Step 5: Speech-to-Text Conversion (Audio Modality)

What we do:

The interview audio is converted into text.

Why:

Textual transcript is needed for content evaluation.

Tools Used:

- Whisper (speech-to-text model)

Output:

- Interview transcript
- Word-level timestamps

Step 6: Audio Feature Extraction (Confidence Analysis)

What we do:

The system analyzes speech characteristics such as:

- Speech energy
- Pitch variation
- Pause duration

Why:

These features indicate confidence, fluency, and hesitation.

Tools Used:

- Librosa (audio signal processing)
- NumPy

Output:

- Audio confidence score

Step 7: Video Frame Extraction (Video Modality)

What we do:

Key frames are extracted from the interview video.

Why:

Facial expressions and body language must be analyzed frame-by-frame.

Tools Used:

- OpenCV

Output:

- Video frames

Step 8: Facial Landmark Detection

What we do:

Faces are detected and facial landmarks are identified.

Why:

Landmarks are required to detect eye contact and head movement.

Tools Used:

- MediaPipe (Face Mesh)

Output:

- Facial landmark coordinates

Step 9: Emotion Detection from Video

What we do:

Facial expressions are analyzed to detect emotions such as:

- Calm
- Nervous
- Stressed

Why:

Emotions help evaluate the candidate's comfort level.

Tools Used:

- FER (Facial Emotion Recognition)
- Deep learning emotion models

Output:

- Emotion scores

Step 10: Video Confidence Estimation

What we do:

Eye contact, head movement, and emotions are combined to compute a video-based confidence score.

Why:

Non-verbal behavior is a key indicator in interviews.

Tools Used:

- Custom Python logic
- Weighted scoring

Output:

- Video confidence score

Step 11: Multimodal Feature Fusion

What we do:

Features from:

- Text (resume & transcript)
- Audio (speech confidence)
- Video (body language)

are combined into a single representation.

Why:

Individual modalities are incomplete; fusion provides holistic judgment.

Tools Used:

- Python (weighted fusion logic)

Output:

- Final confidence score

Step 12: LLM-Based Interview Evaluation

What we do:

All extracted information is sent to a Large Language Model for reasoning.

Why:

Interview evaluation requires reasoning, not just classification.

Tools Used:

- GPT / LLaMA (LLM)
- Prompt engineering

Tasks performed by LLM:

- Resume–answer alignment
- Technical depth evaluation
- Communication quality analysis
- Weakness identification

Step 13: Feedback Generation (Generative AI)

What we do:

The LLM generates structured feedback.

Why:

The goal is coaching, not just scoring.

Tools Used:

- Generative AI (LLM)

Output Includes:

- Strengths
- Weaknesses
- Improvement suggestions
- Sample improved answer

Step 14: Result Presentation

What we do:

Results are displayed to the user in an interactive interface.

Why:

Clear visualization improves usability.

Tools Used:

- Streamlit

Step 15: System Evaluation

What we do:

System outputs are compared with human evaluation.

Why:

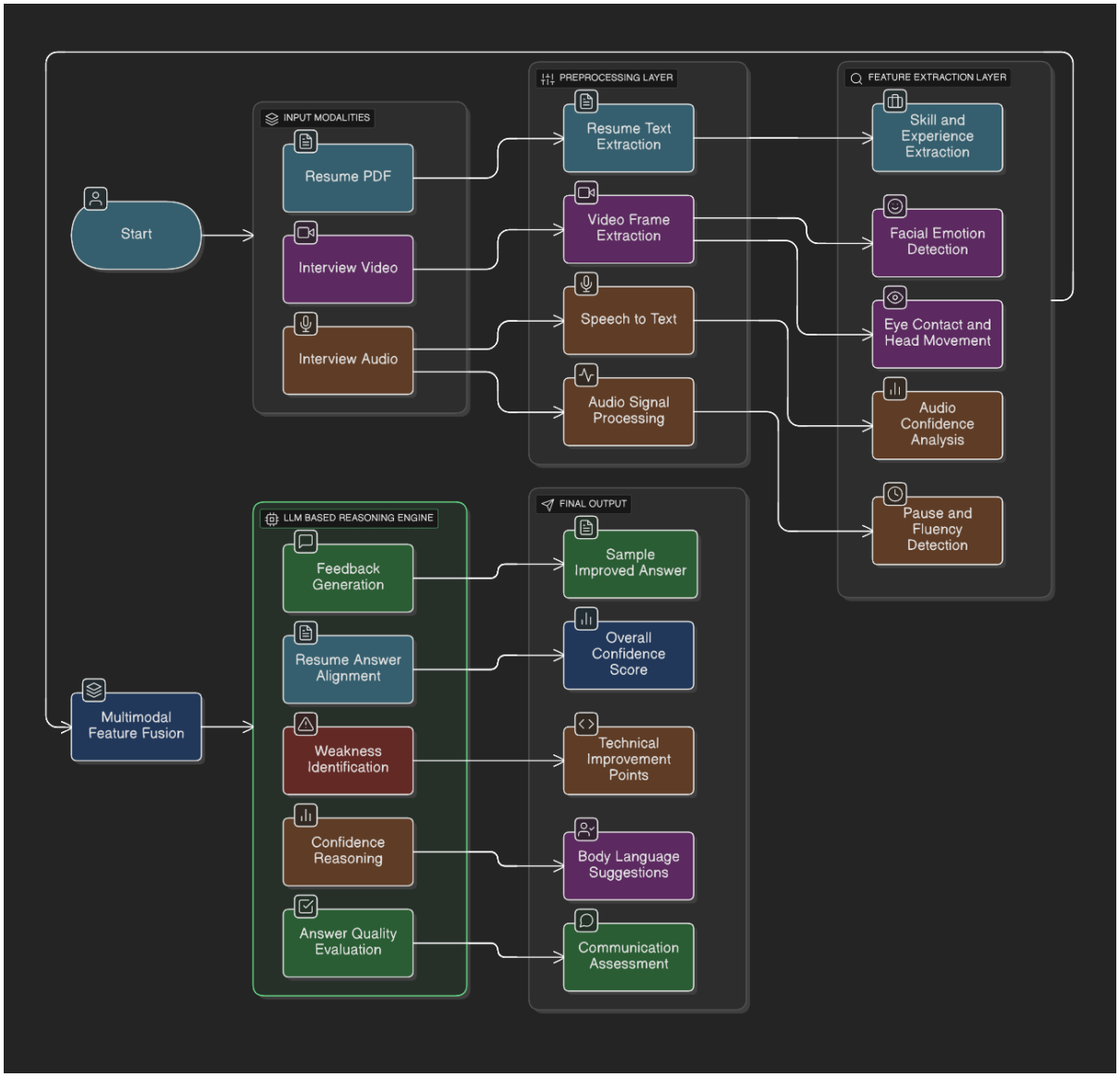
To validate effectiveness.

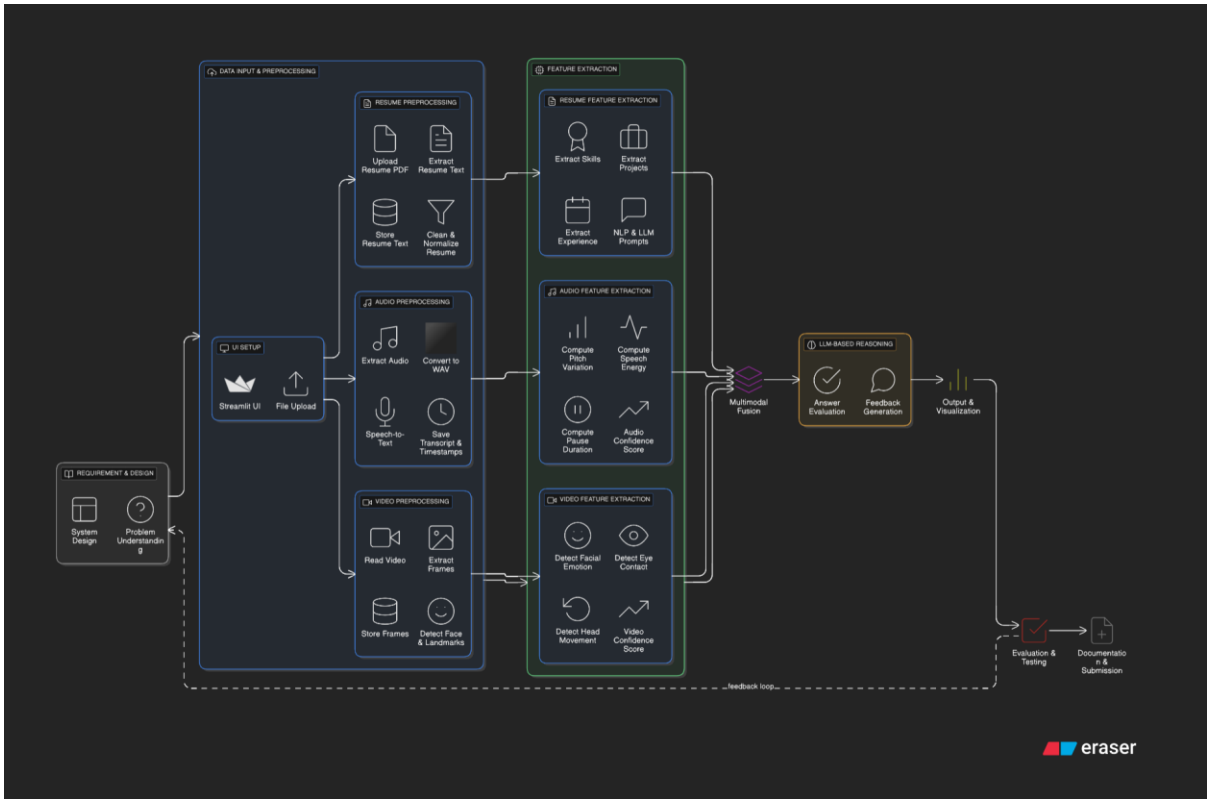
Metrics Used:

- Word Error Rate (STT)
- Emotion detection accuracy
- Human feedback correlation

One-Line Implementation Summary

The system processes resume text, interview speech, and video independently, extracts meaningful features from each, fuses them together, and uses a large language model to reason and generate personalized interview feedback.


Start
INPUT MODALITIES
Resume PDF
Interview Video
Interview Audio
PREPROCESSING LAYER
Resume Text Extraction
Video Frame Extraction
Speech to Text
Audio Signal Processing
FEATURE EXTRACTION LAYER
Skill and Experience Extraction
Facial Emotion Detection
Eye Contact and Head Movement
Audio Confidence Analysis
Pause and Fluency Detection
Multimodal Feature Fusion
LLM BASED REASONING ENGINE
Feedback Generation
Resume Answer Alignment
Weakness Identification
Confidence Reasoning
Answer Quality Evaluation
FINAL OUTPUT
Sample Improved Answer
Overall Confidence Score
Technical Improvement Points
Body Language Suggestions
Communication Assessment


REQUIREMENT & DESIGN
System Design
Problem Understanding
DATA INPUT & PREPROCESSING
UI SETUP
Streamlit UI
File Upload
RESUME PREPROCESSING
Upload Resume PDF
Extract Resume Text
Store Resume Text
Clean & Normalize Resume
AUDIO PREPROCESSING
Extract Audio
Convert to WAV
Speech-to-Text
Save Transcript & Timestamps
VIDEO PREPROCESSING
Read Video
Extract Frames
Store Frames
Detect Face & Landmarks
FEATURE EXTRACTION
RESUME FEATURE EXTRACTION
Extract Skills
Extract Projects
Extract Experience
NLP & LLM Prompts
AUDIO FEATURE EXTRACTION
Compute Pitch Variation
Compute Speech Energy
Compute Pause Duration
Audio Confidence Score
VIDEO FEATURE EXTRACTION
Detect Facial Emotion
Detect Eye Contact
Detect Head Movement
Video Confidence Score
Multimodal Fusion
LLM-BASED REASONING
Answer Evaluation
Feedback Generation
Output & Visualization
Evaluation & Testing
Documentation & Submission
feedback loop
eraser