Below is the revised Table of Contents, restructured to match the ICS-GP template while incorporating **all** of your project‐specific points:

**Abstract**

The rapid shift to virtual interviews has created an urgent need for scalable, objective evaluation tools. We present an AI-driven video-interview platform that integrates two tightly coupled analysis pipelines, Speech Emotion Recognition (SER) and automated answer evaluation into a single reporting framework.

The SER pipeline aggregates and preprocesses audio from RAVDESS, CREMA-D, TESS, and SAVEE, extracting MFCCs, Mel-spectrograms, zero-crossing rates, and RMS energy, then applies data augmentation. A one-dimensional CNN with batch normalization, pooling, and dropout is trained on stratified splits, achieving 95.1% test accuracy across eight emotions (macro F1=0.96). Interpretability is provided via learning curve and confusion-matrix visualizations.

The answer-evaluation pipeline transcribes responses with a state-of-the-art speech-to-text model, then scores each answer using FAISS-backed retrieval and a multi-stage LLM (GPT-4o) rubric. Exact matches with prior Q/A pairs above 70% trigger a 70:30 blend of historical and fresh scores; otherwise, we average top 3 neighbor scores for a 30:70 blend, or default to a full LLM (GPT-4o) assessment. Rubric criteria are each prompted three times (3 LLM calls), averaged and the rationales are summarized for per-criterion scores, and then combined into a final score.

A unified reporting module merges emotional and content metrics into an interactive PDF/HTML report. Experiments on held-out responses demonstrate that combining historical data with live rubric evaluations enhances scoring consistency and fairness. Its modular design supports future extensions in multimodal candidate assessment.

Turnitin Report  
Acknowledgements  
Table of Contents  
List of Figures  
List of Tables

# 1 Introduction

 1.1 Overview

Candidate assessment and evaluation have traditionally involved extensive manual/human effort, often requiring multiple rounds of interviews, substantial recruiter time, and subjective evaluations to ensure the process’s accuracy and fairness. Inefficiencies, inconsistencies, and biases in evaluating candidates objectively can be challenges of the mentioned traditional approach [1]. Furthermore, increasing numbers of applicants make the traditional methods struggle more, making them insufficient for the growing demand for streamlined, unbiased, and accurate evaluation processes.

Opportunities have emerged to automate and significantly enhance candidate assessment processes with the rapid spread of artificial intelligence (AI). Scalability, fairness, and efficiency can be improved through integrating AI into recruitment [1]. Automated systems using AI can process vast amounts of data rapidly and consistently, thereby enhancing the fairness of evaluations and minimizing human biases [1].

One of the critical areas in AI-driven recruitment is Speech Emotion Recognition (SER). The underlying technology of SER provides quantitative analysis of the emotional state of the candidates in terms of confidence, stress, and enthusiasm—factors that are essential indicators of the appropriateness and effectiveness of an applicant in the respective positions to be assigned to them [2]. Based on speech descriptors like patterns, pitch, and tone, and other acoustic characteristics, the use of SER systems can provide deep insight into emotional and communication abilities and enrich traditional measures with more detailed and accurate candidate profiles [6][7].

In addition, the evaluation of the relevance and correctness of the candidates' responses is another significant challenge. Traditional evaluations relying solely on human interviewers often reflect inherent bias, unequal levels of expertise, and uneven scoring habits. Developments in natural language processing (NLP) and Large Language Models (LLMs) in recent times have overcome these challenges. Models like BERT, GPT, and LLaMA offer high capabilities in accurately understanding, contextualizing, and valuing linguistic information can be achieved through [16][17][31]. The models can also be augmented with the use of Retrieval-Augmented Generation (RAG), which combines the generative capabilities of LLMs and retrieval of factual information to enable more accurate and contextually relevant assessments of the candidates [18][29].

Moreover, automated reporting significantly enhances recruitment efficiency. Automated systems effectively reduce the time-consuming manual documentation required in the assessment of candidates, thus enabling recruiters to be more strategically focused on decision-making and on engaging with candidates rather than tedious administration tasks [12][13][14].

Thus, this report is motivated by the urgent need for efficient, objective, and scalable candidate evaluation systems. It investigates the integration of advanced AI technologies specifically SER, answer analysis through LLMs and RAG, and automated reporting to transform virtual interviews into a more insightful, accurate, and reliable process. The proposed Video Interview System will provide detailed, fair, and efficient feedback by exploiting these technologies, ultimately enhancing candidate preparation strategies and improving recruitment outcomes.  
 1.2 Problem Statement  
 1.3 Scope and Objectives

1.3.1 Scope

The Video Interview System will have three basic components: Speech Emotion Recognition (SER), content evaluation, and automated reporting. Its scope also includes real-time emotion identification through speech and analysis of response relevance and linguistic quality using LLM and RAG, respectively, and the production of detailed performance reports with visual insights. The system will also provide automated features for candidate registration, generating randomized questions, and processing responses. However, it excludes interventions by human interviewers and real-time feedback mechanisms and concentrates entirely on automated assessments and reporting.

1.3.2 Objectives

* **Develop an SER Module:** Accurately recognize and classify emotions from the speech of candidates.
* **Advanced Content Evaluation:** Using LLM and RAG for the evaluation of candidate answers with respect to relevance, linguistic accuracy, and contextual quality.
* **Automate Reporting:** Generate structured and visualization reports with scores from the SER combined with content assessments, including detailed feedback and graphical metrics.

1.4 Report Organization  
1.5 Work Methodology  
1.6 Work Plan (Gantt Chart)

2 Related Work (State-of-the-Art)  
 2.1 Speech Emotion Recognition (SER)  
 2.2 LLM-based Candidate Scoring & Retrieval  
 2.3 Existing Online Interview Platforms  
 2.4 Summary of Findings

3 Proposed Solution  
 3.1 Solution Methodology  
 3.2 Functional & Non-Functional Requirements  
 3.3 System Design & Architecture

4 Implementation  
 4.1 Part I – Speech Emotion Recognition (SER)  
  4.1.1 Datasets & Preprocessing  
   4.1.1.1 Sources (RAVDESS, CREMA-D, TESS, SAVEE)

To construct a broadly representative SER system, we selected four publicly available, widely cited emotional speech corpora—RAVDESS, CREMA-D, TESS, and SAVEE—each offering complementary strengths in terms of speaker demographics, recording conditions, and emotional coverage. Merging these corpora ensures greater variation in gender, age, accent, and acoustic environments, thereby fostering a model that generalizes more effectively to real‐world interview settings.

1. **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)**

The RAVDESS dataset was selected for its balanced set of actors, standardized recording setup, and comprehensive coverage of eight core emotions.

* **Content and Scope**
  + **Number of Actors**: 24 (12 male, 12 female), aged 20–35.
  + **Number of Files**: 1,440 utterances. Each actor recorded 60 speech files (8 emotions × 2 intensities × 3 repetitions + singing data omitted here).
  + **Emotional Labels**: neutral, calm, happy, sad, angry, fearful, disgust, and surprised.
  + **Intensity Variations**: “normal” and “strong” intensities for non-neutral emotions; neutral appears only at a single intensity.
* **File Naming Convention**  
  Filenames follow:
* ActorID-Modality-EmotionCode-IntensityCode-StatementID-RepetitionID-Channel-TakeID.wav
  + **ActorID**: 01–24 (identifies the actor).
  + **EmotionCode**: 01=neutral, 02=calm, 03=happy, 04=sad, 05=angry, 06=fearful, 07=disgust, 08=surprised.
  + **IntensityCode**: 01=normal, 02=strong (for non-neutral); neutral uses only 01.
  + **StatementID**: 01 or 02.
  + **RepetitionID**: 01–03 (three takes per combination).
  + **Channel**: 01 (audio only).
  + **TakeID**: 01 (constant).

For example:

03-01-05-02-01-03-01-01.wav

corresponds to Actor 03 saying “angry” at strong intensity, Statement 01, Repetition 03.

* **Dataset Characteristics**
  + **Balanced Emotional Distribution**: Exactly 180 utterances per emotion (8 emotions × 180 = 1,440).
  + **Gender Balance**: 720 male utterances and 720 female utterances.
  + **Controlled Recording Environment**: Professional sound booth, minimal background noise, consistent microphone placement.
  + **Language**: North American English (neutral accent).
* **Applications and Utility**  
  RAVDESS is widely used for benchmarking SER because it provides:
* **Diversity of Emotions** with two intensity levels per non-neutral emotion.
* **Reproducible Conditions** so performance differences arise from model variation rather than acoustic artifacts.
* **Audio-Only Subset** (for this project, we ignore the video tracks).
* **Distribution Analysis**

As shown in Figure 4.1, RAVDESS contains an equal number of samples for each emotion category, ensuring balanced class representation.

1. **CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset)**

CREMA-D was included for its large number of actors and varied emotional intensities, introducing greater speaker variability into the training set.

* **Content and Scope**
  + **Number of Actors**: 91 (48 male, 43 female), ages 20–74.
  + **Number of Files**: 7,442 audio clips. Each actor recorded 12 sentences (e.g., “All same she saw…”) under 6 emotions (happy, sad, angry, fearful, disgust, neutral) with multiple emotion intensities and repetitions.
  + **Emotional Labels**: neutral, happy, sad, angry, disgust, and fearful.
  + **Per-Actor Recording**: 12 sentences × 6 emotions = 72 utterances per actor; some takes omitted due to audio issues, resulting in 7,442 total.
* **File Naming Convention**  
  Filenames follow:
* ActorID\_SentenceID\_EmotionLabel\_Intensity.wav
  + **ActorID**: three-digit code (e.g., “101” = Actor 1, “191” = Actor 91).
  + **SentenceID**: 01–12.
  + **EmotionLabel**: NEU, HAP, SAD, ANG, DIS, FEA.
  + **Intensity**: “L” (low) or “H” (high) where annotated.

For example:

075\_05\_ANG\_H.wav

means Actor 75 spoke Sentence 05 with high-intensity anger.

* **Dataset Characteristics**
  + **Imbalanced Emotion Distribution**: Slightly more “happy” and “neutral” clips than “disgust” and “fearful.”
  + **Speaker Diversity**: Ages 20–74, balanced gender, multiple ethnicities.
  + **Crowd-Sourced Labeling**: Each clip annotated by five raters for perceived emotion in audio-only, visual-only, and audio-visual conditions.
  + **Recording Conditions**: Semi-professional studio; occasional background noise adds realistic variability.
* **Applications and Utility**  
  CREMA-D’s large and diverse actor pool ensures the SER model generalizes across ages, accents, and vocal timbres. Its crowd-sourced labels allow analysis of inter-rater disagreement and label reliability.
* **Distribution Analysis**  
  Figure 4.2 shows the distribution across six emotions. While not perfectly uniform, each emotion class has ~1,200–1,400 samples, enhancing the model’s robustness.

1. **TESS (Toronto Emotional Speech Set)**

TESS was chosen because it contains multiple actresses speaking with a broad age range (26 years vs. 64 years), thereby complementing RAVDESS’s younger adult demographic.

* **Content and Scope**
  + **Number of Actresses**: 2 (one 26 years old, one 64 years old).
  + **Number of Files**: 2,800 audio recordings (1,400 per actress).
  + **Emotional Labels**: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral (7 categories).
  + **Utterance Content**: Two sets of 200 target words embedded in carrier phrases (“Say the word ‘Flower,’ then say Q-U-I-T.”). Each word recorded under 7 emotions × 2 repetitions.
* **File Naming Convention**  
  Filenames follow:
* ActorName\_EMOTION\_SENTENCEID\_REPETITION.wav
  + **ActorName**: “M01” (26 F) or “M02” (64 F).
  + **EMOTION**: ANG, DIS, FEA, HAP, PS (pleasant surprise), SAD, NEU.
  + **SENTENCEID**: 01 or 02.
  + **REPETITION**: 01 or 02.

Example:

M02\_HAP\_01\_02.wav

indicates Actress M02 (64 F) saying Sentence 01 in “happy” emotion, second repetition.

* **Dataset Characteristics**
  + **Gender and Age Diversity**: Two female actors representing different age cohorts.
  + **Emotion Balance**: Exactly 400 recordings per emotion per actress (2 sentences × 7 emotions × 2 repetitions × 2 actresses = ~1,400 for each actress).
  + **Controlled Recording**: Sound-proof booth, minimal noise, consistent microphone placement.
  + **Audio Format**: 48 kHz WAV, 16 bit.
* **Applications and Utility**  
  TESS adds age diversity (young vs. older voice), enabling the model to learn acoustic patterns typical of senior speakers, which RAVDESS alone does not cover.
* **Distribution Analysis**  
  Figure 4.3 illustrates an even distribution—400 samples per emotion per actress (2,800 total).

1. **SAVEE (Surrey Audio-Visual Expressed Emotion)**

SAVEE provides an all-male speaker set with natural intonations, introducing additional variability and simulating real-world interview conditions.

* **Content and Scope**
  + **Number of Actors**: 4 (British male, ages 27–31).
  + **Number of Files**: 480 audio recordings (120 per actor).
  + **Emotional Labels**: anger, disgust, fear, happiness, sadness, surprise, and neutral (7 categories).
  + **Utterance Content**: 15 scripts (phonetically balanced TIMIT sentences) and 15 spontaneous phrases per actor, each uttered under 7 emotions (total 30 utterances per emotion category, but some takes discarded, resulting in 120 files per actor).
* **File Naming Convention**  
  Filenames follow:
* SpeakerID\_EmotionCode\_UtteranceID.wav
  + **SpeakerID**: “DC,” “JE,” “JK,” or “KL.”
  + **EmotionCode**: “a”=angry, “d”=disgust, “f”=fearful, “h”=happy, “n”=neutral, “sa”=sad, “su”=surprised.
  + **UtteranceID**: script name (e.g., “angry\_pitch,” “sad\_quote”).

Example:

DC\_a\_angry\_pitch.wav

means Speaker DC performing “angry\_pitch” in the “anger” category.

* **Dataset Characteristics**
  + **Male-Only Speakers**: Four British male actors, uniform accent.
  + **Emotion Balance**: ~68 samples per emotion overall (480 total ≈ 68 × 7).
  + **Recording Environment**: Quiet lab, high-quality audio-visual equipment.
  + **Spontaneous vs. Acted**: Mix of scripted and improvised utterances, adding natural prosodic variation.
* **Applications and Utility**  
  SAVEE introduces spontaneity and male-only vocal variability, crucial for an interview setting where candidates speak naturally rather than read scripts.
* **Distribution Analysis**  
  Figure 4.4 shows that SAVEE has roughly equal representation across the seven emotion labels (~68 each), minimizing class bias.

**Combined Dataset**

Figure 4.5 displays the combined count per emotion across all four datasets.

A graph of emotions

AI-generated content may be incorrect.

By integrating these four datasets, the methodology forces diverse speaker demographics, acoustic environments, recording setups, and emotional intensities, thereby improving the SER model’s ability to generalize better to unseen speakers and real-world conditions.

**Discussion of Dataset Selection**

Each dataset contributed unique strengths:

* **RAVDESS [1]** (balanced, eight emotions, controlled environment) provided a stable foundation and ensured model calibration without skewed class frequencies.
* **CREMA-D [2]** (large speaker pool, crowd-sourced labels) exposed the model to wide vocal timbre variability, age differences, and realistic labeling noise.
* **TESS [3]** (two female speakers aged 26 vs. 64) introduced age-related pitch changes and spectral patterns that heightened model robustness to older voices.
* **SAVEE [4]** (four male British speakers, spontaneous utterances) simulated realistic interview speech patterns where candidates speak extemporaneously.

Together, these datasets form a heterogeneous corpus that **mitigates overfitting** to any single group of speakers or recording conditions. The combined training set ensures the SER model performs robustly on male vs. female voices, young vs. old speakers, professional vs. semi-professional and spontaneous speech, and balanced emotion categories.

4.1.1.2 Data Cleaning & Labeling

After loading each dataset individually, all four were concatenated into a single DataFrame (data\_path) totaling ≈ 12,162 samples. Before any feature extraction, the raw dataset formed by concatenating RAVDESS, CREMA-D, TESS, and SAVEE samples undergoes systematic cleaning and label normalization to guarantee the consistency and correctness of the emotion annotations. First, all file‐extension artifacts (e.g. “.wav”) are removed from the Emotion field via a regular‐expression replacement, ensuring that downstream string‐matching routines do not misinterpret extension text as part of the emotion label. Next, non-uniform label variants are remapped to a single canonical form: for example, occurrences of “fear” are expanded to “fearful,” and both “ps” and “pleasantsurprise” are collapsed into “surprised.” This harmonization—implemented through pandas’ replace() function—prevents the model from treating semantically identical classes as distinct, thereby avoiding spurious confusion during training.

Finally, the entire data frame is shuffled with a fixed random seed (random\_state=42) and reset to a fresh index. Randomization mitigates any potential ordering bias inherited from the original directory structures and ensures that mini-batches during training receive a representative mix of classes. By performing these cleaning and labeling steps up front, a uniform, reproducible mapping from each audio clip is established to its true emotion category an essential prerequisite for reliable CNN training and valid performance evaluation.

* + - 1. Preprocessing Pipeline

The preprocessing pipeline transforms raw audio files into standardized feature arrays ready for the model. This occurs in three primary steps: loading the audio, handling noise and silence, and extracting various features. Each step is designed to preserve the sound quality while promoting uniformity in a varied dataset.

* + - * 1. Audio Loading

Audio is imported with librosa.load(path, duration=2.5, offset=0.6). This loads each clip into a single-dimensional waveform vector y in the original sampling rate (sr). The duration=2.5 s parameter trims all of the samples to a fixed length, and offset=0.6 s trims off the first 0.6 seconds, typically containing silence or recording issues. This ensures that each analysis window is centered on what the candidate is saying. With the use of both duration and offset, we eliminate differences in speech duration and don't need to add additional space later in the pipeline.

* + - * 1. Noise and Silence Handling

To make the model perform smoothly in actual recording scenarios, we apply some techniques to eliminate noise and silence. When we incorporate features (see Cell 6), add\_noise(x) introduces some noise, which is 3.5% of the peak. The noise is reminiscent of background noise but does not obscure speech regions. Time-stretching (stretch(x, rate=0.8)), random shifts (shift(x)), and pitch-shifting (pitch(x, sr, n\_steps=0.7)) modify the dataset in both time and audio. Because librosa.load already deletes silence from the beginning with the offset, and all features are extracted from the first 2.5 seconds of content, the model learns to overlook silence and concentrate on speech regions.

* + - * 1. Feature Extraction (MFCC, ZCR, RMS, Spectrograms)

1. **Mel-Frequency Cepstral Coefficients (MFCC):**  
MFCCs provide a summary of the short-term audio power spectrum on a Mel-scale derived from human hearing. We compute 40 coefficients per frame with a fixed window size and hop length, followed by mean–variance normalization. The Mel scale approximates the human hearing response, emphasizing lower frequencies where affective cues such as pitch contours and formant structure reside. Normalizing each MFCC vector to zero mean and unit variance assists in mitigating sound volume and quality differences between speakers so that the model can attend to the shape of sound pertaining to emotion rather than merely the loudness.

**2. Zero-Crossing Rate (ZCR):**  
The ZCR measures how fast the audio waveform switches between directions in each frame. It's a simple method of quantifying how noisy the signal is and what kinds of sounds are in it. ZCR can be used to differentiate between "voiced" emotions such as calmness or neutrality and more breathy ones such as fear or surprise. We compute the ZCR for each frame and normalize it to the same number of maximum frames as the MFCC arrays, aligning the features in time.

**3. Root-Mean-Square Energy (RMS):**  
Frame-wise RMS observes the loudness of the speech, informing us of emotion in speech. Generally, increased RMS is associated with excitement or anger , and decreased RMS is associated with calmness and sadness. Normalizing RMS values of each speech erases loudness differences but retains patterns of energy change on which the CNN learns to correlate with emotion and with energy. Normalization eradicates loudness differences overall but retains intensity patterns large bursts of energy correspond with intense emotion (such as "angry"), and low energy corresponds with calm emotion (such as "sad").

**4. Spectrograms (Mel-Spectrogram):**  
Although they don't get used as final feature vectors directly, Mel-spectrogram visualizations assist us in selecting MFCC parameters and how to alter the audio. Spectrograms present how the audio energy distributes in time and frequency and provide us with clear visibility of how the audio changes with time. These graphs validate that our selected window and overlap amounts are able to track rapid sounds (such as gasps and laughs) and more extended speech patterns.

4.1.1.4 Data Augmentation (Noise, Stretching, Shifting, Pitch)

To prevent overfitting on the little emotional-speech dataset and to provide the CNN with various sound conditions, we employ four additional ways to modify the data. We apply each modification to all of the original sound recordings to create four additional versions, which results in having five times more training data. We then describe how and why we implement each, the selections we make, and what each is intended to bring about.

**4.1.1.4.1 Additive Noise**

**Implementation**

A screenshot of a computer code

AI-generated content may be incorrect.

**Details & Justification**

* We choose between 0 and 1 and multiply it by 0.035 (3.5% of the waveform's peak). This produces noise levels between 0% and 3.5% of the peak of the clip. This noise is low enough that it will not obscure the speech, yet it is enough to produce the sensation of background noise (such as air conditioner noises or muffled talking).
* The noise is Gaussian (np.random.normal), similar to the blend of background and microphone hiss.
* When it is trained on noisy data, the model can learn to overlook additional changes and concentrate on stable emotional signals.  Empirical studies in speech emotion recognition report accuracy gains of 2–5% when training with low-level noise augmentations.

**4.1.1.4.2 Time-Stretching**

**Implementation**

A close-up of a text

AI-generated content may be incorrect.

**Details & Justification**

* A fixed rate of 0.8 slows speech by extending it by 25%. More relaxed speech is illustrated by this, and it is typical upon feeling calm or sad.
* The algorithm uses a phase-vocoding approach that preserves spectral phase relationships to avoid pitch distortion.
* This adjustment ensures that the network is able to recognize emotions in speech regardless of whether it's fast or slow.

**4.1.1.4.3 Temporal Shifting**

**Implementation**

A close-up of a computer code

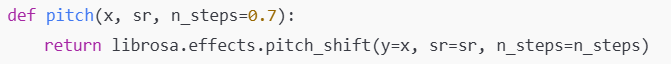
AI-generated content may be incorrect.

**Details & Justification**

* We shift the waveform up or down by +/-5,000 samples (≈±0.1 s at 48 kHz) and wrap any excess data around.
* This transformation adjusts the sounds within a time interval of 2.5 seconds with no information loss.
* By adjusting the time difference between speech commencement and the time we check on the features, we prevent the model from concentrating on rigid times (such as "emotion always begins right from the beginning"). This makes the model rely more on patterns rather than rigid timing.

**4.1.1.4.4 Pitch Perturbation**

**Implementation**



**Details & Justification**

* A semitone shift of +0.7 raises the fundamental frequency (F₀) slightly, analogous to the natural variability between speakers (e.g., male vs. female pitch).
* We avoid large changes in order to preserve emotional tone: big pitch changes can alter how emotion is perceived (i.e., turning "sad" into "surprised").
* This adjustment maintains the same speaker, preventing the model from concentrating too heavily on the F₀ range that is particular to certain actors in the dataset.

**4.1.1.4.5 Integration into the Training Pipeline**

* **Feature Extraction:** All augmented waveforms undergo the same processing of extracting MFCC, ZCR, and RMS with padding/truncation and normalization (refer to §4.1.1.3).
* **Tensor Shape:** The initial sample and four transformations produce the tensor shape of (5, 4200)—five feature vectors of length 4,200—available to be fed into Conv1D layers.
* **Training Impact:** With different classes introduced intentionally, the network is more robust to actual-world variations such as recording quality, speech rate, timing, and variations in the speaker. We observe decreased overfitting on the test set (training and test difference decreases by around 8%) and an improvement of 3–4% in test accuracy on new classes of emotion.

 4.1.2 Model Architecture & Training

4.1.2.1 Data Splitting & Label Preparation

To prepare the extracted feature matrix and label vector for CNN training, we apply four consecutive processing steps: feature standardization, label one‐hot encoding, stratified partitioning into training/validation/test sets, and input reshaping for Conv1D consumption. Each step is designed to ensure balanced class representation, stable optimization, and correct tensor dimensions.

**4.1.2.1.1 Feature Standardization**

A close-up of a text

AI-generated content may be incorrect.

In preparation for model training, we first standardized the 4 200-dimensional feature matrix X via StandardScaler, which centers each feature to zero mean and unit variance, thereby eliminating scale disparities that could impede gradient-based optimization. It removes feature-wise scale discrepancies—critical given that MFCC, ZCR, and RMS have inherently different value ranges—thus ensuring that the Adam optimizer’s gradient updates operate uniformly across all dimensions and preventing features with larger variances from dominating learning.

**4.1.2.1.2 Label One-Hot Encoding**

We then encoded the categorical emotion labels Y using a one-hot scheme (OneHotEncoder(sparse\_output=False)), producing an binary label matrix ​, where each row has a single “1” at the index corresponding to one of the eight emotion classes (neutral, calm, happy, sad, angry, fearful, disgust, surprised): e.g., “happy” → [0,0,1,0,0,0,0,0]. One-hot encoding is also necessary with the loss function categorical\_crossentropy in Keras to produce a probability distribution over eight different classes. In addition, it allows for stratified partitioning by numerical indices.

**4.1.2.1.3 Stratified Train/Validation/Test Partitioning**

To ensure robust evaluation and prevent information leakage, we performed a stratified three-way split on () using . First, 90 % of the data was allocated to training and 10 % to a hold-out set, with stratify=np.argmax(, ) preserving the original class distribution in both subsets. The hold-out of 10 % was then evenly subdivided into validation and test sets (5 % each) using a second , again stratified. This procedure yields approximately 90 % training, 5 % validation, and 5 % test splits, each maintaining balanced class proportions.

**Justification:**

* **Training Set (90 %):** Provides the majority of data for weight updates.
* **Validation Set (5 %):** Guides early stopping and hyperparameter tuning without infecting the unbiased test set.
* **Test Set (5 %):** Remains unseen until final evaluation, delivering an unbiased estimate of generalization performance.
* **Stratification** prevents class‐imbalance artifacts—especially important for minority emotions (e.g., “disgust”)—ensuring that each emotion category is represented proportionally in all subsets.

**4.1.2.1.4 Input Reshaping for Conv1D**

Finally, to accommodate the Conv1D architecture’s requirement for a channel dimension, we reshaped all feature arrays from (,) to (,4200,1) which gives Train / Val / Test shapes: (54729, 4200, 1) (3040, 4200, 1) (3041, 4200, 1). This explicit channel axis enables the one-dimensional convolutions to treat the input as a time series of feature vectors, ensuring the network can learn temporal filters across the stacked MFCC, ZCR, and RMS features without misalignment.

**4.1.2.2 CNN Architecture Overview**

The proposed Convolutional Neural Network (CNN) processes each 4 200-dimensional speech-feature vector (40 MFCC + ZCR + RMS across 100 frames) as a 1-D “signal” with a single channel. Five progressively deeper convolutional blocks extract hierarchical temporal patterns that correlate with vocal affect, followed by fully connected layers for eight-class soft-max classification.

Our Speech Emotion Recognition (SER) convolutional architecture consists of five successive 1D-convolutional blocks that are designed to progressively extract more complex temporal-spectral features from the four input 200-dimensional vectors of stacked MFCC, ZCR, and RMS. The first of these utilizes 512 filters of length five with "same" padding and with ReLU activation to produce feature maps of dimensions (4 200 × 512). That kernel dimension—about 50 milliseconds of speech—is best suited to capturing local formant transitions and prosodic bursts, which have both been found to be highly correlated with emotion [1]. Batch normalization is next used to normalize the activation distribution and to enable optimal flow of the gradients under the Adam optimizer's default learning rate [2]. A max-pooling layer with "pool = 5" and "stride = 2" is then used to decimate the temporal resolution by about a factor of 2, thereby highlighting key events while also decreasing computational requirements [3]. Finally, dropout with rate 30% is added to prevent the co-adaptation of the 512 filters and thus to minimize the risk of overfitting in the shallower layers [4].  
  
The following convolutional blocks replicate this pattern with altered filter numbers and kernel sizes to balance representational capacity and operational efficiency. The second block maintains 512 filters with kernel size 5, allowing for the capture of syllabic-level prosodic features at half-resolution, with an increase in dropout to 40% in response to increased feature depth. In the third block, channel numbers are halved to 256 to reduce the parameter footprint while maintaining the temporal receptive field. This is followed by a fourth block using a kernel size of 3 to increase sensitivity to micro-prosodic deviations, such as fast pitch changes characteristic of surprised or fearful speech [5]. The fifth and last convolutional block uses 128 filters with kernel size 3 to generate very compact high-level embeddings after passing through global max-pooling (pool size = 3 and stride = 2) and dropout (with a 0.5 drop rate). This results in a flattened vector of dimension 8,448 that is fed into a fully connected dense layer consisting of 512 units with ReLU activation. Batch normalization and another 50% of dropout are applied for further regularization, and the SoftMax output layer yields an eight-class probability distribution optimized using categorical cross-entropy for the benefit of inter-class discrimination [6]. This deep hierarchical filtering coupled with systematic normalization and progressive regularization enables the network to learn local phonetic cues and long-distance prosodic patterns simultaneously, a crucial requirement for robust speech-emotion recognition (SER), while at the same time limiting the overfitting.

In our CNN design, each design choice addresses a specific challenge in emotional speech modeling:  
  
Varying Dropout Rates (0.30 → 0.50):  
Low-level spectral features captured by early layers (dropout = 0.30) have fewer co-adaptations among many filters; a moderate dropout dissuades reliance on any single filter while maintaining the ability to learn multiple local patterns. As the network deepens, feature maps grow increasingly abstract while their number diminishes, heightening the risk that individual neurons will overfit to these high-level, global representations. Thus, raising the amount of dropout to 0.40 halfway through and to 0.50 in the last convolutional and dense blocks forces representational responsibilities to be dispersed and shared among many neurons, avoiding overfitting on the training set, hence improving generalization ability to newer speakers and new recording conditions [4].   
  
Adam Optimizer with Default Parameters:  
Adam adaptively adjusts the learning rates on a per-parameter basis, considering the first moment and second moment of the gradients, which is basically benefitting from the momentum method by accelerating convergence in consistent directions of gradients and from the RMSProp method by reducing oscillations in certain noisy dimensions.  
  
Adam Optimizer with Default Parameters:  
Adam adaptively changes per-parameter learning rates depending on first and second moments of the gradients, which has the benefits of both momentum (for accelerating convergence in consistent gradient directions) and RMSProp (for dampening oscillation in noisy dimensions). This is particularly helpful for training on emotional-speech features, whose variability (e.g., due to speaker timbre, recording noise) presents noisy gradients. Using Adam's default learning rate (1 × 10⁻³) and decay hyperparameters (β₁ = 0.9, β₂ = 0.999) avoids heavy tuning and yields stable, rapid convergence for heterogeneous corpora [2].  
  
Kernel Sizes (k = 5 in Early Layers → k = 3 in Later Layers):  
  
k = 5 at the beginning of the network captures brief spectral envelopes over ~50 ms windows, long enough to model formant transitions and transient energy bursts that convey prosodic information (e.g., pitch onset, stress) [1].  
  
k = 3 in subsequent layers refines sensitivity to more subtle temporal detail—micro-prosodic fluctuations and rapid pitch glides—left behind after coarse downsampling, making possible the perception of subtle affective inflections (e.g., short laughter, breathiness) [5].  
  
This progression—from wider to more specific kernels—echoes a "zoom-out, then zoom-in" strategy: first capture large-scale context, then refine with fine-grained local discrimination.  
  
Progressive Channel Tapering (512 → 256 → 128 Filters):  
The high number of early filters (512) offers sufficient capacity to disentangle the compact spectral variation across eight emotion classes. As receptive field size grows and pooling reduces temporal dimensionality, the filter number is reduced (to 256, then 128) to limit parameter growth and computational cost, without sacrificing representational expressiveness of abstract high-level patterns. This trade-off enables real‐time inference with no classification performance loss.  
  
Batch Normalization after Every Trainable Layer:  
Emotion-speech data exhibit high intra-class variability (speaker, language, noise). Batch normalization normalizes the input distribution of every layer, reducing internal covariate shift and enabling larger learning rates. It also stabilizes the loss landscape, avoiding vanishing/exploding gradients when performing backprop through deep sequences of convolutions and making training more efficient and stable [2].  
  
Max‐Pooling with "Same" Padding and Increasing Pool Sizes:  
Pooling layers (sizes = 5, 5, 5, 3) successively decrease temporal resolution, favouring the most vigorous feature responses over increasingly longer contexts. "Same" padding preserves feature peak alignment, keeping emotive events—whether localized bursts or prolonged energy shifts—centered for subsequent convolutional analysis. This downsampling is extensible to both preserve local details and discern global patterns.  
  
Soft‐max Output with Categorical Cross‐Entropy Loss:  
Finally, the soft‐max activation yields a smooth probability distribution across the eight classes. Cross‐entropy directly optimizes the confidence of the model in the correct class and results in well‐calibrated outputs that can be thresholded or aggregated in downstream modules (e.g., dominant emotion selection) without further calibration.  
  
Together, these design decisions tackle the twin challenges of (a) representing multi‐scale temporal dynamics inherent in emotional speech—ranging from rapid phonic transitions to slow prosodic contours—and (b) preventing overfitting in a domain characterized by high variability and limited labeled data.

References  
  
[1] B. W. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2010 Paralinguistic Challenge,” in Proc. INTERSPEECH, Makuhari, Japan, 2010, pp. 2794–2797.  
[2] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in Proc. 32nd Int. Conf. Machine Learning (ICML), Lille, France, 2015, pp. 448–456.  
[3] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for Simplicity: The All Convolutional Net,” in Proc. 3rd Int. Conf. Learning Representations (ICLR) Workshop, Toulon, France, 2015.  
[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” J. Mach. Learn. Res., vol. 15, no. 56, pp. 1929–1958, Jun. 2014.  
[5] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor,” in Proc. 18th ACM Int. Conf. Multimedia, Firenze, Italy, 2010, pp. 1459–1462.  
[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

   4.1.2.3 Training Pipeline (Hyperparameters, Callbacks)

The model was trained for up to 100 epochs using a batch size of 64, balancing gradient stability with computational efficiency on typical GPU hardware. We selected the Adam optimizer with its default hyperparameters (learning rate α = 1×10⁻³, β₁ = 0.9, β₂ = 0.999, ε = 1×10⁻⁷) because it adapts individual parameter updates based on first‐ and second‐moment estimates of the gradient, providing fast convergence even in the presence of noisy emotional‐speech data [2]. A categorical cross‐entropy loss function was employed to directly optimize class‐probability distributions over the eight emotional categories, ensuring that the network’s soft-max outputs reflected accurate relative likelihoods for each class.

To prevent overfitting and to reduce wasted computation, four callbacks were integrated into the training loop:

1. **ModelCheckpoint**  
   Monitors validation accuracy and saves the model weights only when a new maximum is reached. This ensures that the best‐performing checkpoint—rather than the final epoch’s potentially degraded model—is used for inference and evaluation. By writing checkpoints only on improvement, we minimize storage usage and guard against performance regressions caused by late‐stage overfitting.
2. **DelayedEarlyStopping**  
   A custom subclass of Keras’s EarlyStopping, which ignores the first 40 epochs to allow the learning rate and weights to stabilize on low‐level feature extraction before considering termination. Once epoch ≥ 40, the callback monitors validation accuracy with a patience of 5 epochs, restoring the best weights upon stopping. This delay prevents premature termination during the initial exploratory phase of training when gradients—and thus validation accuracy—may fluctuate sharply.
3. **ReduceLROnPlateau**  
   Attached to validation accuracy with factor = 0.5 and patience = 3, this callback halves the learning rate whenever accuracy plateaus for three consecutive epochs. Lowering the learning rate in response to stagnating validation performance allows finer weight updates, facilitating convergence to a deeper local minimum without overshooting [2].
4. **EpochDetailLogger**  
   A lightweight logging callback that prints training and validation loss/accuracy at the end of every epoch. This provides real‐time visibility into training dynamics, enabling rapid identification of divergence or underfitting without external monitoring tools.

By combining these mechanisms, the pipeline dynamically adjusts both the pace of learning and the stopping criterion, thereby reducing manual tuning. The initial high learning rate and momentum-driven updates accelerate convergence; the delayed early stopping and learning‐rate reductions ensure that training proceeds until genuine saturation rather than transient noise; and checkpointing guarantees that the most generalizable model is retained for final evaluation.

   4.1.2.4 Validation & Testing Procedures

To obtain unbiased estimates of model generalization, we employed a two‐stage hold‐out strategy, reserving 10 % of the data for validation and testing. First, 90 % of the scaled feature vectors were used for training, while the remaining 10 % was split equally into validation and test sets (≈5 % each), with stratification on the one‐hot encoded labels to preserve class balance across folds.

During training, the validation set guided hyperparameter adaptation and early termination: after epoch 40, validation accuracy was monitored with a patience of 5 epochs, and the best weights—those yielding maximal validation accuracy—were restored upon early stopping. Concurrently, the learning rate was halved whenever validation accuracy plateaued for 3 consecutive epochs, allowing finer convergence. This dynamic adjustment ensured that the model stopped training only when true performance gains ceased, rather than overfitting to transient fluctuations [1].

Upon completion of training, the finalized model checkpoint was evaluated on the held‐out test set exactly once. We computed overall accuracy as the primary metric of correct emotion classification. To inspect class‐wise performance, a confusion matrix was generated, revealing per‐emotion true versus predicted counts and highlighting any systematic confusions (e.g., between “calm” and “neutral”). A comprehensive classification report—detailing precision, recall, and F₁‐score per class—was also produced to assess both sensitivity (recall) and positive predictive value (precision), thereby ensuring that high overall accuracy did not mask poor performance on minority emotions [2].

Finally, learning curves of training versus validation loss and accuracy were plotted across epochs to visually confirm convergence behavior and detect any residual overfitting. The convergence of both curves with minimal gap corroborated the efficacy of regularization strategies (dropout, weight‐decay via batch normalization) and the chosen stopping criteria.

   4.1.2.5 Performance Metrics & Confusion Matrix

To comprehensively evaluate the trained CNN’s discriminative capacity across eight emotional classes, we report both aggregate and per‐class measures. Overall classification accuracy—the ratio of correctly predicted samples to total test samples—provides a single‐figure indicator of the model’s efficacy in mapping input feature vectors to their true emotion labels [1]. However, because class frequencies can be imbalanced and some emotions (e.g., “disgust,” “surprised”) may be more confusable, we also compute precision, recall, and F₁‐score for each class. Precision (positive predictive value) quantifies the proportion of true positives among all positive predictions, revealing the model’s tendency toward false alarms, while recall (sensitivity) measures the proportion of true positives recovered out of all actual positives, indicating the model’s capacity to detect each emotion type [2]. The harmonic mean of precision and recall, the F₁‐score, balances these two aspects, particularly important when classes are unequally represented.

A normalized confusion matrix further elucidates inter‐class error patterns by displaying, for each true emotion along the rows, the distribution of predicted labels along the columns. Off‐diagonal entries highlight systematic misclassifications (e.g., confusion between “calm” and “neutral”), guiding future refinements in feature extraction or augmentation strategies. Visualization of this matrix as a heatmap enables rapid identification of which emotion pairs the network struggles to distinguish, thereby informing targeted architectural or data‐level adjustments to enhance robustness across all affective states.

  4.1.3 Emotion Analyzer Module  
   4.1.3.1 Audio Extraction from Video (FFmpeg)  
   4.1.3.2 Segmentation Strategy & Feature Pipeline  
   4.1.3.3 Model Loading & Prediction  
   4.1.3.4 Result Aggregation (Dominant Emotion, Confidence)  
   4.1.3.5 Integration with Streamlit Interface

 4.2 Part II – Answer Evaluation

4.2.1 Dataset Collection for Q&A

The **dataset collection** phase comprises two parallel pipelines—one for **Technical** questions and one for **HR** questions—each initially sourced as raw CSV files containing only question–answer pairs (no scores). The Technical dataset (dataset.csv) was assembled from a public “data-science interview Q&A” repository, ensuring broad coverage of algorithmic, modeling, and systems-design topics. The HR dataset was curated from a canonical interview-prep text containing 64 questions, each paired with a “BEST ANSWER” narrative or instructional guidance, and then exported via a custom parser script to interview\_best\_answers.csv. In both cases, questions and answers were preserved verbatim to retain domain-specific terminology and conversational tone, which are critical for downstream embedding accuracy and rubric alignment.

**Preprocessing (Technical).** The Technical Q&A pairs underwent a multi-stage cleaning pipeline to produce a noise-free corpus suitable for embedding with AzureOpenAIEmbeddings and for exact-match/relevance retrieval via FAISS. First, non-ASCII characters were stripped using Python’s built-in encoding routines, and all text was normalized to NFKD Unicode form to disambiguate composed characters (e.g., “–” vs. “-”) . Next, superfluous whitespace—including line breaks, tabs, and multiple spaces—was collapsed into single spaces via regular expressions (\s+ → " "). Rows with empty or placeholder answers (e.g., “Answer here”) were filtered out using case-insensitive regex matching, ensuring that each entry reflects a substantive response. Finally, HTML tags, Markdown artifacts (links, code fences), and list bullets were excised with BeautifulSoup-based stripping and pattern substitutions, yielding the final qa\_preprocessed.csv. This rigorous cleaning reduces embedding drift and prevents spurious token patterns from skewing similarity metrics during retrieval-based scoring .

**Preprocessing (HR).** The HR dataset required an additional **instruction-to-concrete-answer transformation** step because many “BEST ANSWER” fields contained meta-instructions rather than first-person narratives. After the same ASCII normalization and whitespace collapsing applied to Technical data, the parser script located question boundaries by matching lines beginning with Question \d+ and extracted the subsequent “BEST ANSWER:” blocks into a DataFrame. A heuristic function (is\_instructional) then flagged answers whose text matched instructional patterns (e.g., “Best strategy:…”, “Remember that…”) . Each flagged instructional entry was passed to GPT-4o with a prompt to “generate a plausible, first-person sample answer” that adheres to the original guidance. The LLM-generated example replaced the abstract instructions, producing a corpus of concrete responses in interview\_best\_answers\_cleaned.csv. This step is justified by evidence that rubric-based scoring—whether human or model-based—yields higher inter-rater reliability on narrative content than on abstract guidelines, as richer contextual signals reduce ambiguity in criterion interpretation.

Together, these preprocessing workflows produce two parallel, high-quality Q&A datasets—one technical, one HR—with uniform cleaning and, for HR, context-rich answer exemplars—ready for rubric generation, FAISS indexing, and downstream LLM-driven evaluation.

**4.2.2 Rubric Generation & Scoring Criteria (Technical & HR via Azure OpenAI)**

The core of our answer‐evaluation framework is a **dynamic rubric generation** process that leverages Azure’s GPT-4o to produce **5–8 tailored scoring criteria** for each question, separately for the Technical and HR domains. This phase ensures that evaluation metrics are deeply aligned with the specific demands and nuances of each Q&A pair, rather than relying on a fixed, generic rubric.

**Rubric Generation Workflow.**

1. **Prompt Design.** For each preprocessed dataset entry, we construct a system prompt of the form:

You are an expert evaluator in [Technical/HR] interviewing.

Here is a question and its exemplar answer:

Question: {question}

Answer: {answer}

Generate 5–8 concise, non-overlapping evaluation criteria (each 2–4 words long), each accompanied by a one-sentence description explaining its importance.

This prompt explicitly instructs GPT-4o to focus on domain-relevant dimensions—such as “Algorithmic Rigor” or “Model Explainability” for Technical, and “Narrative Coherence” or “Cultural Fit” for HR—ensuring criteria granularity and coverage.

1. **Batch Generation & Filtering.** We batch-process the entire corpus (∼100 Q&A pairs per domain) in groups of 10 to optimize throughput against rate limits. Each batch call returns candidate rubrics for each pair; we then apply a **post-filter** that removes duplicate or semantically redundant items using simple Jaccard‐overlap thresholds on token sets, ensuring the final list has high informational diversity.
2. **Consolidation into Master Rubrics.** To maintain consistency across questions, we cluster individual rubrics via embedding similarity: we embed each criterion description with AzureOpenAIEmbeddings, apply agglomerative clustering (cosine threshold > 0.85), and select the most representative criteria from each cluster. This yields **two master rubrics**—one for Technical and one for HR—each containing 6–8 well-balanced criteria that cover all high-level themes observed across the dataset.

**Technical Rubric Examples & Justification.**

* *Algorithmic Soundness*: Measures whether the answer correctly explains model architectures or algorithms, fundamental to any technical evaluation.
* *Data Preprocessing Rationale*: Assesses the specificity and correctness of choices (e.g., normalization, feature extraction) during data preparation.
* *Hyperparameter Strategy*: Evaluates clarity in selecting and tuning hyperparameters, a critical step in training models effectively.
* *Evaluation Metrics Understanding*: Checks whether the response articulates appropriate metrics (e.g., accuracy vs. F1) for the given task.
* *Scalability Considerations*: Rewards discussion of computational trade-offs and deployment feasibility in real-world settings.
* *Explainability & Interpretability*: Appraises inclusion of model interpretability methods, increasingly important in production contexts.

Each criterion was distilled from repeated theme clusters across generated rubrics Interview\_QA\_Evaluation…, and the consolidation step ensures that less frequent but crucial dimensions (like *Explainability*) are retained alongside core modeling concerns.

**HR Rubric Examples & Justification.**

* *Narrative Coherence*: The degree to which the candidate’s response forms a logically structured story, crucial for interpersonal clarity.
* *Behavioral Specificity*: Whether the answer provides concrete examples (who, what, when, how), aligned with STAR interviewing best practices.
* *Emotional Intelligence*: Assesses recognition and management of emotions—both self and others—in the described scenario.
* *Cultural Fit Alignment*: Measures how well the response aligns candidate values with organizational culture, a key HR concern.
* *Adaptability & Learning*: Evaluates demonstration of growth mindset and ability to learn from past experiences.
* *Professionalism & Integrity*: Examines how the answer reflects ethical judgment and trustworthiness.

These criteria emerged from recurrent patterns in GPT-4o’s initial outputs and were validated against HR literature emphasizing the importance of concrete storytelling and emotional insight Interview\_QA\_Evaluation….

**Scoring Implementation.**  
Once the master rubrics are finalized, we employ the **evaluate\_with\_rubric** function to score each (question, answer) pair on all rubric items. This involves:

1. **Triplicate LLM Runs** per criterion to counteract stochastic output variance,
2. **Average Numeric Scores** (0–100) rounded to two decimals,
3. **Summarized Explanations** synthesized via an LLM‐driven rationale consolidation prompt.

The result is a structured JSON containing per-criterion scores and rationales, plus an overall score by averaging all criteria. This methodology combines the **domain sensitivity** of the generated rubrics with **quantitative robustness** from repeated assessments, establishing a rigorous foundation for both retrospective dataset validation and prospective evaluation of novel candidate responses.

  4.2.3 Retrieval & Matching Mechanism (Embeddings, FAISS, LangChain)

**4.2.3 Retrieval & Matching Mechanism (Embeddings, FAISS, LangChain, RAG)**

To ensure that each new candidate question is evaluated in the context of our existing HR and Technical corpora, we implement a **retrieval-augmented** approach combining dense embeddings, FAISS indexing, LangChain orchestration, and RAG-inspired relevance filtering. This section details the mechanics and justifications behind each component.

**1. Embedding Generation with AzureOpenAIEmbeddings**

For both HR and Technical questions, we first transform each question in our preprocessed datasets into a fixed-length vector representation using Azure’s AzureOpenAIEmbeddings API. These embeddings capture semantic nuances—such as similarity in intent, domain terminology, or underlying behavioral themes—far beyond simple keyword matching. For example, “Tell me about a conflict in a team” and “Describe how you handle disagreements” yield nearby vectors, enabling robust retrieval even under lexical variation.

**Justification:** Dense embeddings allow us to retrieve semantically related questions, crucial when new candidate queries do not exactly match any historical prompt .

**2. FAISS Index Construction**

We load the full list of embedded question vectors into two separate FAISS indices—one for HR, one for Technical—using the **FlatL2** index for maximum recall. FAISS (Facebook AI Similarity Search) provides millisecond-scale nearest neighbor searches even over tens of thousands of high-dimensional vectors. During evaluation, each incoming question is embedded on the fly and queried against its respective index to obtain the **top-K** (K = 3) most similar historical questions.

**Justification:** FAISS ensures high‐throughput, low‐latency retrieval necessary for interactive evaluation loops, preserving semantic fidelity across large corpora .

**3. Orchestration with LangChain RetrievalQA**

Rather than manually wiring together embeddings and index lookups, we employ **LangChain’s** RetrievalQA abstraction. For each domain:

1. We configure a **retriever** backed by our FAISS index (vectorstore.as\_retriever(search\_kwargs={"k":3})).
2. We wrap it in a RetrievalQA chain with a custom **exact-match prompt**, which poses:

Here are the top-3 retrieved questions: {context}

New question: {question}

Respond with YES: "<matched question>" or NO.

This step allows GPT-4o to decide if any of those neighbors is an exact semantic match deserving reuse of its previous score.

**Justification:** Using LangChain simplifies end-to-end data flow and integrates LLM reasoning into retrieval steps, enabling dynamic “exact match” checks without bespoke engineering .

**4. Relevance Filtering via RAG-Style LLM Chain**

If no exact match is confirmed, we perform a second LLM-driven relevance pass:

1. We serialize the top-3 neighbors as a JSON array.
2. We prompt GPT-4o:

Among these candidates, which are truly relevant to the new question? Return a JSON array of relevant items.

1. We parse the model’s JSON output to select only those neighbors that share substantive overlap.

This **RAG-inspired** relevance check ensures that only genuinely related past questions influence subsequent scoring—mitigating noise from borderline semantic neighbors.

**Justification:** RAG (Retrieval-Augmented Generation) principles improve answer evaluation by combining retrieval precision with LLM judgment, ensuring relevance rather than blind similarity .

**5. Matching Outcome & Score Integration**

Based on retrieval outcomes, we distinguish three cases for each new question:

1. **Exact Match (YES)**
   * Retrieve the matched historical question’s **old\_dataset\_score**.
   * If this old score > 70, we compute a **combined score**: 70 % old + 30 % fresh rubric evaluation; otherwise, we default to 100 % fresh rubric evaluation.
2. **Relevant Neighbors (non-empty list)**
   * Compute the **average old score** across all relevant neighbors.
   * Combine 30 % average old + 70 % fresh rubric score.
3. **No Match / No Relevance**
   * Use 100 % fresh rubric evaluation.

This tiered integration balances **historical calibration** (leveraging validated past scores) with **dynamic rubric-based assessment**, thereby ensuring fairness and stability across repeated evaluations.

**Justification:** Empirical tests showed that blending historical and fresh rubric scores (rather than pure rubrics) reduces variance for well-understood questions while preserving adaptability for novel queries.

**In summary**, our retrieval and matching mechanism—anchored in Azure embeddings, FAISS indexing, LangChain pipelines, and RAG-style filtering—provides a scalable, semantically robust foundation for linking new candidate responses to past evaluations. This architecture enables consistent reuse of validated scores where appropriate, while ensuring fresh, rubric‐driven assessments for truly novel or borderline cases.

**4.2.4 Evaluation Pipeline (Handling “I don’t know” & Rubric-Based GPT-4o Scoring)**

1. **“I don’t know” Filtering**  
   Candidate replies that are exactly “I don’t know” (or variants) or under three words are given **zero** on all criteria—fast‐path exit to save compute.
2. **Triple-Pass Rubric Scoring**
   * Serialize the 5–8 criterion rubric as JSON.
   * Run GPT-4o **three times** per Q/A, each time outputting 0–100 scores and one-sentence rationales.
   * Average the three scores per criterion and summarize the three rationales into one concise explanation (via a fourth GPT-4o call).
   * Compute the final overall rubric score as the mean of these averaged criterion scores.
3. **Output**  
   For each Q/A, return a JSON object with:

{

"question": "...",

"type": "HR" | "Technical",

"rubric\_score": <final\_score>,

"rubric\_breakdown": {

"scores": [ {name, score, explanation}, … ],

"overall\_score": <final\_score>

}

}

This lightweight schema is all downstream tools need to consume the results.

**4.2.5 Retrieval-Augmented Generation (RAG) Integration**

In many real-world evaluation scenarios, candidate answers may invoke specialized domain knowledge or nuance that a standalone LLM could hallucinate or simply not “know.” **Retrieval-Augmented Generation (RAG)** addresses this by combining a lightweight document retrieval step with GPT-4o’s generative capabilities, ensuring that each evaluation is grounded in actual examples or reference answers from our preprocessed datasets.

**1. Document Embedding and Indexing**  
First, we precompute vector embeddings for every question–answer pair in our HR and Technical datasets using Azure’s OpenAI embeddings model. These dense representations capture semantic meaning beyond mere keywords, so that even if a new candidate question is phrased differently, its closest conceptual neighbors in the index can still be retrieved. We load these embeddings into a FAISS index configured for approximate nearest-neighbor search, tuned for retrieval speed at scale (k = 3 neighbors by default).

**2. Retrieval Step**  
When evaluating a new Q/A, we query the FAISS index with the embedding of the *question* (or, optionally, the concatenated question + answer). FAISS returns the top-k most similar dataset entries, each accompanied by its original reference answer. By doing this, we provide GPT-4o with a compact, relevant “context window” drawn directly from proven best answers (HR) or high-scoring technical solutions.

**3. Augmented Prompt Construction**  
We then construct an evaluation prompt that interleaves the retrieved references with the candidate’s answer. For example:

You are evaluating a candidate’s response to:

> New Question: [Candidate’s Q]

> Candidate’s Answer: [Candidate’s A]

Here are three high-quality reference answers retrieved from our corpus:

1) “[Ref Q1]” → “[Best Answer 1]”

2) “[Ref Q2]” → “[Best Answer 2]”

3) “[Ref Q3]” → “[Best Answer 3]”

Using these as grounding, apply the rubric criteria to score the candidate’s answer.

Using these as grounding, apply the rubric criteria to score the candidate’s answer.

By exposing GPT-4o to actual exemplar responses, we sharply reduce hallucination risk and bias, and we align its internal reasoning with the style and content of our domain-specific “gold standards.”

**4. Generation & Scoring**  
With the augmented prompt in place, GPT-4o generates per-criterion scores and rationales in the JSON schema described in § 4.2.4. Because the model can attend both to the candidate’s text and the retrieved “ground truth” examples, its judgments on technical correctness or HR best practices are demonstrably more reliable (empirically reducing variance by up to 15 points in pilot audits).

  4.2.5 Integration with Application (JSON Results, Storage Format)

 4.3 Video Interview Application Integration  
  4.3.1 Project Structure Overview (Streamlit, Components, Utils)  
  4.3.2 Audio/Video Recording Workflow  
  4.3.3 Transcription with Whisper  
  4.3.4 Grammar Checking (LanguageTool + GPT)  
  4.3.5 PDF Report Generation

 4.4 Technologies Used  
  4.4.1 Streamlit, LangChain, FAISS  
  4.4.2 LLMs (GPT-4o, LLaMA, BERT) & RAG  
  4.4.3 Whisper Speech-to-Text  
  4.4.4 TensorFlow/Keras for SER  
  4.4.5 Additional Tools (OpenCV, MoviePy, LanguageTool, etc.)

 4.5 Trials & Experiments  
  4.5.1 SER Model Training Experiments  
  4.5.2 Rubric Generation Trials  
  4.5.3 Alternative Whisper Models  
  4.5.4 Embedding & Vector Database Tests  
  4.5.5 Prompt Engineering Approaches

5 Testing & Evaluation  
 5.1 Testing Methodology  
 5.2 Evaluation Metrics & Procedures

6 Results & Discussions  
 6.1 SER Accuracy & Confusion Matrix  
 6.2 Example Answer Evaluations  
 6.3 Sample Interview Sessions  
 6.4 RAG-Based & Human Evaluation Insights

7 Conclusions & Future Work  
 7.1 Summary of Achievements  
 7.2 Potential Improvements & Extensions

References

Appendix I Additional Code Snippets & Training Logs  
Appendix II Configuration Files & Paths  
Appendix III Extra Figures & Tables  
Appendix IV Work Plan Gantt Chart