# PRVIA: AI-Based Framework for Automated Evaluation of Pre-Recorded Video Interviews


Assistant Prof. Mona Abdelazim
Information System. Ain Shams University
Cairo, Egypt
monaabdelazim@cis.asu.edu.com

TA. Aya Nasser
Scientific Computing . Ain Shams University
Cairo, Egypt
aya.naser@cis.asu.edu.com

Ammar Mohamed Hassan
Artificial Intelligence. Ain Shams University
Cairo, Egypt
ammarhassan7018@gmail.com

Mohamed Ashraf Mahran
Artificial Intelligence. Ain Shams University
Cairo, Egypt
mohamedmahran6325@gmail.com

Mohamed Samy Mohamed
Artificial Intelligence. Ain Shams University
Cairo, Egypt
mohamedsamyy02@gmail.com

Nadine Haitham Ali
Artificial Intelligence. Ain Shams University
Cairo, Egypt
nadeenhaitham2003@gmail.com

Yomna Mohamed Bassam
Artificial Intelligence. Ain Shams University
Cairo, Egypt
yomnamuhammedbassam.67@gmail.com

Youssef Tamer Mahmoud
Artificial Intelligence. Ain Shams University
cCairo, Egypt
yousseftamer891@gmail.com



***Abstract*—Pre-recorded video interviews have emerged as a vital tool for modern, remote-friendly recruitment, yet manual evaluation remains costly, inconsistent, and prone to bias. We introduce PRVIA (Pre-Recorded Video Interview Analysis), the first end-to-end, AI-driven framework that automates the assessment of asynchronous interviews by extracting and integrating verbal and nonverbal cues. Our system adopted automatic speech recognition, large language models for semantic relevance, facial expression analysis, and personality traits estimation to produce detailed, multi-dimensional scores for English proficiency, response coherence, affective engagement, and personality indicators. Deployed via a web-based platform, PRVIA processes each interview in under 2 minutes, delivering standardized reports that exhibit high correlation with human evaluators while substantially reducing time-to-hire and mitigating common evaluative biases. Extensive experiments on real-world interview datasets demonstrate PRVIA's ability to streamline large-scale hiring workflows, enhance fairness, and maintain decision quality. This work fills a critical gap in the literature by offering a scalable, fully integrated solution for automated pre-recorded interview evaluation.**



**Keywords—AI recruitment, video interview analysis, multimodal evaluation, expression recognition, speech-to-text, personality detection, semantic similarity, automated hiring.**


## I. Introduction

The modern job market is undergoing a fundamental transformation as organizations increasingly adopt virtual hiring practices. With the rise of remote work and distributed teams, traditional face-to-face interviews are being supplemented—and often replaced—by online and pre-recorded formats. Among these, pre-recorded video interviews have emerged as an efficient solution for initial candidate screening. They offer scheduling flexibility and enable recruiters to review responses at their convenience. However, despite their advantages, the manual evaluation of these video submissions remains a time-intensive and often inconsistent process, subject to human bias and fatigue.

Pre-recorded interviews are typically used during the early stages of hiring, where candidates respond to a predefined set of questions on video. While scalable in theory, the task of reviewing a large number of these recordings can lead to delays in the recruitment cycle and inconsistent decision-making. Moreover, technical disparities—such as poor lighting, low-resolution cameras, or weak audio—may unfairly influence the perception of candidate performance. These limitations are compounded by common cognitive biases in human evaluation, including stereotyping, recency effects, and first impressions, all of which can distort hiring outcomes.

To address these challenges, this work proposes PRVIA (Pre-Recorded Video Interview Analysis), an AI-powered web platform that automates the assessment of video interviews using a modular, multimodal approach. The system integrates deep learning techniques across audio processing, natural language processing (NLP), and computer vision. models to analyze both verbal and non-verbal cues, generating comprehensive candidate profiles based on language proficiency, personality traits, emotional expression, and the semantic relevance of responses. The goal is to streamline the evaluation process, reduce manual effort, and promote fairness through consistent, data-driven scoring.

The key contributions of this project are as follows:

1. The development of an end-to-end system for automated interview analysis using audio, visual, and textual data.

2. The application of deep learning techniques for emotion recognition, English proficiency estimation, and semantic similarity scoring;

3. A dual-interface platform supporting both HR reviewers and job applicants;

4. A significant reduction in time-to-hire through real-time processing, with average report generation completed in under 2 minutes per candidate.

By minimizing manual workload, mitigating evaluator bias, and enabling scalable candidate assessment, PRVIA offers a practical solution for modern recruitment workflows. Its integration into existing HR systems has the potential to enhance hiring efficiency while ensuring fairness and inclusivity.

## II. Related Work

The field of **automatic pronunciation assessment (APA)** has evolved significantly over the past two decades. Early systems like the one proposed by [1] **Witt and Young (2000)** introduced the **Goodness of Pronunciation (GOP)** framework, which relied on forced alignment using DNN-HMM-based ASR systems and handcrafted features. While foundational, these systems were limited in scalability and struggled with accent diversity.

Subsequent research, such as [2] **Zhang et al. (2008)**, addressed these limitations by incorporating language-specific modeling, particularly for tonal languages like Mandarin. However, the reliance on phoneme-level alignment remained a bottleneck.

More recently, the [3] **GOPT model by Gong et al. (2022)** Adopted Transformer-based self-attention networks in a multi-task learning setup. It extracted multi-aspect pronunciation features (e.g., accuracy, fluency, completeness) using traditional GOP features and regression heads. Despite richer feedback, it retained the computational overhead of alignment-based preprocessing.

A major shift came with [4] **alignment-free scoring techniques**. **Fu et al. (2022)** proposed a pronunciation system powered by **wav2vec2**, a self-supervised model trained on raw audio data. This eliminated the need for explicit alignment and improved robustness to diverse accents and spontaneous speech.

Building on this, [5] **Fu et al. (2024)** introduced a **multi-modal large language model (LLM)**–based framework combining **data2vec2** and a decoder-only LLM (**Qwen-7B**), achieving utterance-level predictions without intermediate feature engineering. This end-to-end, deep learning-based pipeline offered a scalable and more generalizable approach to pronunciation scoring, which serves as the basis for the audio module in PRVIA.

Personality and emotion inference from video has seen rapid development, particularly since the ChaLearn First Impressions Challenge (2017–2018). Güçlütürk et al. proposed a hybrid CNN-LSTM architecture, capturing both spatial and temporal features for personality trait prediction. Gabor et al. demonstrated that simpler temporal pooling strategies could achieve competitive results, especially with limited data. More recent approaches favor lightweight, spatiotemporal networks such as X3D, which efficiently model behavioral dynamics like eye gaze, smile duration, and head movement while maintaining high accuracy and reducing computational load—an ideal fit for real-time recruitment pipelines [22].

In parallel, facial expression-based emotion recognition has benefited from large-scale datasets like FER-2013 and AffectNet. DeepFace, originally designed for face verification, now includes an emotion classification module trained on these datasets. Its modular design and inference-only architecture make it suitable for scalable, -latency deployment, as required in our PRVIA system [23]. These tools support plug-and-play integration without fine-tuning, aligning with the system's design principles.

Multiple academic and commercial systems have explored AI-based candidate screening. HireVue, for instance, uses proprietary AI models to assess both verbal and non-verbal communication, integrating directly with Applicant Tracking Systems (ATS) to reduce time-to-hire. VidCruiter provides customizable video interview analysis platforms used in enterprise settings. From the academic side, Chen et al. [25] introduced a large-scale dataset and evaluation framework that analyzes candidates using multimodal cues. Lee et al. demonstrated a real-world hiring system combining visual, vocal, textual, and physiological signals to predict job fit, achieving 82% accuracy and 85.5% user satisfaction. Naim et al. and Suen et al. [26] validated the use of prosody and facial features in assessing engagement and communication skills.

## III. System Architecture

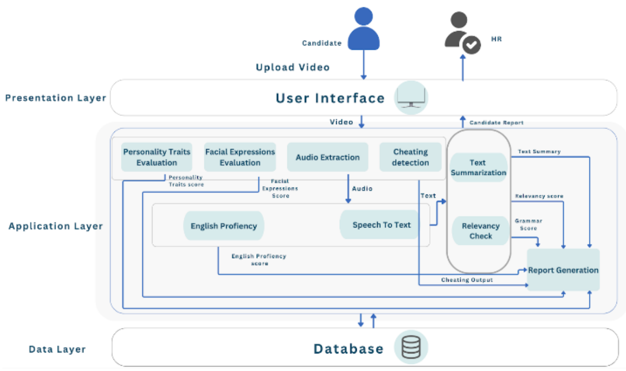


**Fig. 1** System Architecture

The system consists of four core analytical modules—Video Module, Audio Module, Text Module, and Report Generation Module—in addition to supporting components including the User Interface and Database. Each module includes specialized sub-modules responsible for processing distinct aspects of candidate behavior and communication.

### A. Core Modules Description

#### 1) Video Module

The Video Module initiates processing by accepting the uploaded interview video. It contains four sub-modules:

Personality Traits Evaluation: Predicts Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) using spatiotemporal deep learning models on extracted facial features.

Facial Expressions Evaluation: Analyzes frame-level emotional expressions to infer candidate affect and engagement using DeepFace and valence-based categorization.

Cheating Detection: Tracks eye movement to detect off-screen behavior such as reading answers, measuring gaze duration deviation from the camera.

Audio Extraction: Separates audio from the video stream for downstream processing in the Audio Module.

**2) Audio Module**

The Audio Module processes the extracted audio and performs two main tasks:

English Proficiency Evaluation: Assesses fluency, pronunciation, and clarity using self-supervised audio representations and transformer-based scoring models. Generates an overall English level classification (e.g., low, average, high).

Speech-to-Text Conversion: Converts spoken audio to text via automatic speech recognition (ASR), enabling further textual analysis.

**3) Text Module**

The Text Module receives the transcribed interview content and analyzes it through three sub-modules:

Text Summarization: Generates a concise summary highlighting key points from the candidate's spoken responses.

Relevancy Checking: Evaluates alignment between candidate responses and the intended interview questions.

**4) Report Generation Module**

This module integrates outputs from all analytical components and produces a structured report. The report includes:

Personality trait scores

Facial expression distribution

English proficiency score

Relevance and grammar assessment

Cheating alerts (if applicable)

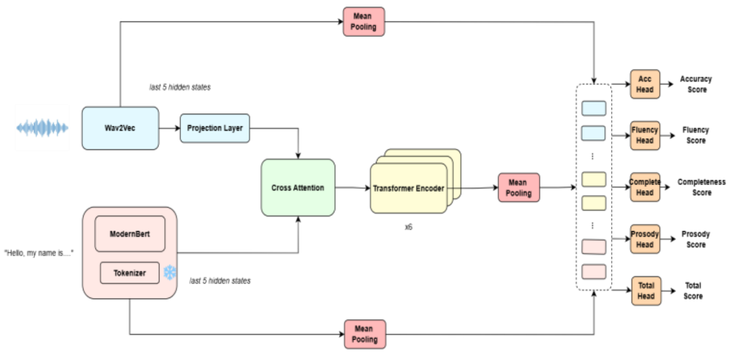


**Fig. 2** Illustration of the proposed architecture with sample utterance“Hello, my name is”, actual utterances used are longer.

## IV. Audio Module

**A. Speechocean 762 Dataset**

Speechocean 762[27] is a free open-source dataset designed for pronunciation assessment, consisting of a total of 5,000 English utterances collected from 250 non-native speakers. One major advantage of speechocean762 is that it provides rich label information. Specifically, for each utterance, it provides five utterance-level aspect scores: accuracy, fluency, completeness, prosody, and total score (ranging from 0-10). For each word, it provides three word-level aspect scores: accuracy, stress, and total score (ranging from 0-10). It also provides an accuracy score for each phoneme (ranging from 0-2). Each score is annotated by five experts. Thus, it provides a total of 8 labels for different granularities and pronunciation quality aspects. We restricted our analysis to the utterance-level annotations available in the Speechocean762 dataset, specifically the scores for accuracy, fluency, prosody, and completeness at the sentence level. This decision was driven by the nature of our task: in job interview scenarios, the primary concern is overall spoken communication quality, not fine-grained phonetic feedback. The training set consists of 2,500 utterances, 15,849 words, and 47076 phones; the test set consists of 2,500 utterances, 15,967 words, and 47,369 phones.

**B. Model Architecture**

An overview of our architecture is shown in Figure 2. For the pronunciation assessment task, the transcription is known.

**1.Speech Encoder**

Wav2Vec2-Large-Xlsr-53-English was employed as the audio encoder due to its strong performance in phonetic representation and robustness across accents. 16 kHz raw audio is transformed into a sequence of hidden representations, each with a dimensionality of 1,024. The last five hidden states are concatenated, as they capture essential

**Table 1.** Comparison of utterance-level pronunciation assessment performance across multiple scoring dimensions for PRVIA and baseline models.

| *Type* | *Model* | *PCC Acc* ↓ | *PCC Flu* ↑ | *PCC Pro* ↑ | *PCC Comp* ↓ | *PCC Total* ↓ | *AVG PCC* |
|---|---|---|---|---|---|---|---|
| *Align Based* | *GOPT* [3] | *0.714* | *0.753* | *0.760* | ***0.155*** | ***0.742*** | - |
| | *Multi- task* [6] | *0.694* | *0.730* | - | - | - | - |
| | *MultiPA* [7] | *0.705* | *0.772* | - | - | - | - |
| | *HiPAMA [8]* | ***0.730*** | *0.762* | - | - | - | - |
| *Align Free* | *SSL* [9] | - | *0.780* | - | - | - | - |
| | *Multi-modal LLM-based* [5] | *0.713* | *0.777* | - | - | - | - |
| | *Proposed (PRVIA)* | *0.693* | ***0.790*** | ***0.770*** | *0.060* | *0.717* | *0.608* |

.temporal and phonetic patterns, contributing to more accurate pronunciation assessment

## 2. Text Encoding with ModernBERT

ModernBERT was adopted as the text encoder due to its strong generalization capabilities and rich semantic representations. It processes tokenized textual inputs and generates contextual embeddings with a dimensionality of 768. To preserve both semantic and syntactic information, the final five hidden states are concatenated. During training, the parameters of ModernBERT were frozen to retain its pre-trained knowledge and reduce computational overhead, allowing the model to focus optimization on the downstream components.

## 3. Cross-Attention Mechanism

To effectively integrate audio and textual modalities, a cross-attention mechanism is employed to align speech and language features. Prior to fusion, audio embeddings are linearly projected from 1,024 to 768 dimensions to match the dimensionality of the text embeddings. The cross-attention layer, configured with 32 attention heads and a dropout rate of 0.2, enables the text representations to attend to the projected audio features. This bidirectional interaction enhances the model's capacity to capture dependencies between linguistic content and acoustic patterns, facilitating more accurate multimodal pronunciation evaluation.

## 4. Transformer Encoder

A stack of six Transformer encoder layers is applied to the cross-attended features to refine the fused audio-text representations. Each layer employs Flash Attention for an efficient multi-head self-attention mechanism with 32 heads, a position-wise feed-forward network with 2,048 hidden units, and layer normalization. The encoder captures complex interdependencies across modalities and enhances the contextual representation, outputting a sequence of embeddings with a dimensionality of 768.

## 5. Multimodal Fusion and Regression Scoring

Mean pooling is applied to the outputs of the Transformer encoder, along with the final five hidden states from both wav2vec2-large-xlsr-53-english and ModernBERT, to generate fixed-length feature vectors. These vectors are concatenated into a unified multimodal representation, which serves as input to five dedicated multi-layer perceptron (MLP) regression heads. Each head predicts one of the scoring dimensions—accuracy, fluency, completeness, prosody, or total score—using a three-layer architecture with GELU activation and a 0.2 dropout rate. The outputs are scaled to produce final scores in the range of 0 to 10.

## 6. Experiments

The model is optimized using the AdamW optimizer with an initial learning rate of 0.001, cosine learning rate decay, and a weight decay of 0.0001. During training, only the cross-attention layer, Transformer encoder, selected layers from the Wav2Vec2 model, and the multi-layer perceptron (MLP) regression heads are updated, while ModernBERT remains frozen to preserve its pre-trained language representations and reduce computational overhead. The loss function combines Mean Squared Error (MSE) for score prediction with a correlation-aware regularization term designed to maximize the Pearson correlation coefficient (PCC) between predicted and ground-truth scores. Since the Speechocean762 dataset contains imbalanced labels skewed toward higher scores, PCC is adopted as the primary evaluation metric to better reflect the model's ranking ability across the score range. Training is performed over 50 epochs with a batch size of 16, completing in approximately 4 hours using GPU acceleration

### 7. Results and Evaluation

In this subsection, we evaluate the performance of the proposed multi-modal pronunciation scoring system, which adopts an align-free, end-to-end architecture. For a fair comparison, we benchmarked our model against two categories of baseline systems: (1) align-based methods that rely on traditional DNN-HMM ASR pipelines trained on the LibriSpeech dataset, and (2) align-free approaches based on self-supervised learning (SSL) features. We excluded align-based systems enhanced with SSL features [6], as they introduce additional complexity and reduce practicality in real-world deployments. As shown in Table 1, our model achieves competitive performance across key pronunciation metrics. In particular, PRVIA attains a fluency PCC of 0.790—surpassing all align-based baselines—and achieves solid performance in accuracy (0.693) and prosody (0.770), while maintaining completeness and total score consistency. Compared to align-based models, which require more intricate preprocessing and alignment steps, PRVIA provides a simpler and more scalable solution for pronunciation assessment. Additionally, relative to existing align-free methods, our model closes the performance gap in accuracy scoring, an area where previous methods underperformed. Notably, these results are achieved using a lightweight architecture trained on limited GPU resources, highlighting the efficiency and practical deployability of the proposed framework.

## V. Video Module

This module extracts behavioral insights from visual content, providing personality profiling, emotional state assessment, and attention-based integrity monitoring. It comprises three main components: **Personality Traits** Prediction, **Emotion Recognition & Engagement Analysis**, and **Gaze-Based Cheating Detection.**

### A. Personality Traits Prediction

#### 1. ChaLearn First Impressions V2 Dataset

Our personality analysis is based on the **ChaLearn First Impressions V2 dataset** [22], a benchmark for apparent personality recognition containing **10,000 short video clips** (~15 seconds) of individuals speaking directly to the camera. Each clip is annotated with continuous values for the **Big Five personality traits**: **Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness**, based on the aggregated judgments of multiple annotators. These scores capture apparent personality—i.e., how individuals are perceived—rather than self-reported traits. The dataset includes multimodal inputs such as video, audio, and transcripts, making it suitable for a range of behavioral modeling tasks.

We also refer to the large-scale interview corpus by **Chen et al.** [25], which comprises **1,891 annotated videos** (totaling over 63 hours), designed for trait evaluation and hiring recommendation. Their multimodal framework integrates facial expressions, vocal prosody, and textual content to emulate real-world virtual interviews.

While both datasets support multimodal learning, our system focuses exclusively on the **visual modality**—specifically facial and motion-based cues—Despite the availability of multimodal inputs, we adopt a visual-only approach, as recent findings demonstrate that models such as X3D-S can approach state-of-the-art performance using facial and motion cues alone [22]. This choice enhances deployability by reducing complexity and computational cost

#### 2. Model Architecture

We adopt a lightweight, visual-only pipeline tailored to predict continuous personality traits. The architecture includes the following components:

#### i. Face Detection and Frame Extraction

Videos are decoded using the `av` library, from which 60 evenly spaced frames are extracted. Each frame undergoes face detection and alignment using MTCNN, followed by resizing to 160×160 pixels. This ensures temporal diversity and input consistency across samples.

#### ii. Spatiotemporal Feature Encoding (X3D-S)

To encode facial expressions and motion cues, we use a pre-trained **X3D-S** model, originally trained on **Kinetics-400**. As shown by Li et al. [22], deep spatiotemporal models effectively capture short-term social signals. The X3D-S model produces a **400-dimensional feature vector** through global average pooling over time.

#### iii. Multi-Head Regression for Trait Prediction

The extracted feature vector is passed to **five parallel MLP regressors**, each predicting one personality trait. Each MLP follows a 400 → 128 → 64 → 1 architecture, with **GELU activations** and a dropout rate of 0.2. Outputs are normalized to the range **[0, 1]** to match the dataset's ground truth format.

#### iv. Training Strategy

The dataset is divided into five equal parts (2,000 videos each). Progressive training is applied: each chunk is trained for **50 epochs** with checkpoint continuation. We use the **AdamW optimizer** with an initial learning rate of **0.001**, cosine scheduler, **weight decay of 0.0001**, and batch size **16**. Training completes in approximately **5 hours** on a single NVIDIA GPU for each Chunk.

#### v. Experimental Results and Evaluation

The model achieves a **mean accuracy of 90.52%**, surpassing the strongest published visual-only baseline

This highlights its ability to capture expressive personality cues from facial dynamics alone.

TABLE 2
MEAN ACCURACY PER PERSONALITY TRAIT

| Trait | Mean Accuracy (%) |
|---|---|
| Extraversion | 91.3 |
| Agreeableness | 89.7 |
| Conscientiousness | 90.8 |
| Neuroticism | 89.9 |
| Openness | 91.7 |
| **Overall Mean Accuracy** | **90.52** |

**vi. Trait-to-Text Mapping**

To enhance interpretability, trait scores are mapped to human-readable personality descriptors using defined thresholds. For example, a high extraversion score results in a label such as "friendly", while lower values correspond to "reserved". This semantic mapping enables easier integration into candidate evaluation reports.

**B. Emotion Analysis**

**1. CAER Dataset**

To evaluate the emotion recognition module, we used a subset of **350 annotated videos** from the **CAER (Context-Aware Emotion Recognition) dataset** [24]. CAER consists of short video clips extracted from television scenes and labeled with seven universal emotions: **Angry, Disgust, Fear, Happy, Sad, Surprise**, and **Neutral**. Despite its entertainment origin, CAER captures **naturalistic facial behaviors**, spontaneous expressions, and varied contexts—qualities that align closely with real-world virtual interviews. Its variability in **scene context**, **occlusion**, and **emotional intensity** makes it a strong proxy for asynchronous candidate assessments. For consistency and computational efficiency, each video was sampled at **5 frames per second (FPS)** using OpenCV (`cv2`), ensuring uniform temporal representation across all clips.

**2. Model Architecture**

Our emotion recognition system employs a **pre-trained, inference-only pipeline**, prioritizing portability and computational efficiency. The architecture comprises the following components:

**i. Frame Sampling and Preprocessing**

Input videos are processed using cv2.VideoCapture. To reduce temporal redundancy while preserving emotional dynamics, we sample five evenly spaced frames per second. Each frame is resized to the required input resolution for the downstream emotion classifier.

**ii. Face Detection and Emotion Classification**

Face detection is performed on each sampled frame using **MediaPipe Face Detection**, a high-speed, CPU-efficient solution that handles real-time applications. When a face is detected, it is cropped and passed to the **DeepFace Emotion Classifier**, a pre-trained CNN trained on **FER-2013** and optionally fine-tuned on **AffectNet**. The model outputs probability distributions over the seven emotion classes for each valid frame.

**iii. Emotion Aggregation**

Emotion predictions from all valid frames are aggregated using mean probability scores. The emotion with the highest mean value is selected as the dominant emotion for the video. This approach reduces the impact of transient or noisy frame-level predictions and improves temporal consistency.

**iv. Engagement Assessment**

To assess candidate engagement, the module groups predicted emotions into two main valence categories: positive (including happy, surprise, and neutral) and negative (including angry, sad, fear, and disgust). The relative distribution of these emotional categories across frames is analyzed to infer high-level engagement states. If positive emotions dominate, the candidate is labeled as Confident; a roughly balanced distribution results in a Neutral label; and if negative emotions prevail, the candidate is considered Nervous. This valence-based summarization provides interpretable and human-readable insight into the candidate's affective demeanor during the interview.

**3. Techniques and Implementation Details**

**i. Frame Sampling with OpenCV (cv2)**: Efficient frame extraction and decoding ensures low-latency video handling.

**ii. Face Detection with MediaPipe**: Lightweight, real-time detection that runs efficiently on CPU and supports edge deployment.

**iii. Emotion Classification with DeepFace**: Utilizes a CNN trained on FER-2013, with optional AffectNet augmentation for broader generalization.

**iv. Fault Tolerance**: Frames without valid face detections are skipped, making the pipeline robust to occlusions, motion blur, or low lighting.

**v. Valence-Based Mapping**: Aggregated emotion distributions are simplified into broad emotional tendencies, improving interpretability for non-technical stakeholders.

**vi. Technologies**

**DeepFace Emotion Engine**: A pre-trained CNN-based emotion recognition tool requiring no fine-tuning, enabling scalable deployment across users and hardware.

**MediaPipe Detection Pipeline**: Provides real-time, accurate face detection optimized for high-FPS scenarios with minimal hardware dependency.

**Valence-Based Engagement Mapping**: Translates emotion probabilities into semantically meaningful engagement labels (e.g., Confident, Nervous).

**Inference-Only Design**: The pipeline runs entirely using pre-trained components, making it highly portable, fast, and resource-efficient.

**Vii. Experimental Results and Evaluation**

The system was evaluated using the 350-video subset of the CAER dataset. Three main evaluation criteria were considered:

TABLE 3
VIDEO MODULE EVALUATION METRICS

| Metric | Score |
|---|---|
| Frame-Level Face Detection Rate | >90% |
| Dominant Emotion Agreement (w/ Ground Truth) | >85% |

Our design aligns with recent advances in **context-aware emotion modeling** [24] and **vision-language-based inference** [23], where emotion understanding benefits from both facial expressions and high-level contextual interpretation. While our current approach is visual-only for efficiency, future extensions may integrate transformer-based multimodal pipelines as proposed in [23] and [26] for deeper behavioral insight.

**C. Gaze-Based Cheating Detection**

This system aims to detect potential cheating behavior by analyzing gaze direction inferred from facial landmarks using the MediaPipe Face Mesh model. The underlying hypothesis is that a subject may be engaged in cheating if they consistently look away from the screen for a sustained period.

**1. Model Architecture**

**i. Frame Processing and Facial Landmark Detection**

Video frames are captured using cv2.VideoCapture and processed with the MediaPipe Face Mesh model. The model detects and localizes 478 facial landmarks, including refined iris points when refine_landmarks=True is enabled. Each frame is analyzed independently to identify key eye and iris regions.

**ii. Iris Localization and Gaze Direction Estimation**

From the set of detected landmarks, specific indices are used to extract the horizontal eye boundaries (left eye: 33, 133; right eye: 362, 263) and iris centers (468 for left eye, 473 for right eye). Gaze direction is estimated using a normalized geometric equation: If the result falls within the empirically defined threshold range $0.35<x<0.65$, the gaze is considered centered.

$$\text{relative\_position} = \frac{\text{iris}_x - \text{eye\_left}_x}{\text{eye\_right}_x - \text{eye\_left}_x}$$

**iii. Temporal Gaze Validation**

To reduce noise from frame-level fluctuations, a temporal logic mechanism is applied.

Gaze must remain centered for ≥ 0.5 seconds to be confirmed. If gaze remains off-center for ≥ 3.0 seconds, it is flagged as potential cheating behavior.

**iv. Decision Logic**

A cumulative counter tracks the duration of non-centered gaze. Once this exceeds the 3-second threshold, the system raises a cheating alert. This approach ensures that only sustained gaze deviations are flagged, reducing false positives due to natural eye movement.

**2. Techniques and Implementation Details Facial Landmark Extraction with MediaPipe**

**i. MediaPipe Face Mesh** provides a real-time, CPU-efficient facial landmark model capable of detecting 478 landmarks per face, including iris points for precise gaze tracking.

**ii. Geometric Gaze Estimation**: Uses eye and iris landmarks to calculate normalized iris position without additional learning, enabling fast, model-free direction inference.

**iii. Temporal Logic Filtering**: Applies time-based rules to distinguish between transient and sustained gaze deviations, improving the system's reliability.

**iv. Threshold Calibration**: The $0.35<x<0.65$ threshold range was empirically determined to balance natural eye movement and cheating detection accuracy.

**v. Robustness to Missing Data**: Frames without valid face or iris detection are skipped, allowing the system to remain robust under occlusions

**3. Technologies**

**MediaPipe Face Mesh:** a real-time facial landmark estimation framework developed by Google. It predicts 478 3D facial landmarks from RGB input and operates efficiently on CPUs. When refine_landmarks=True is enabled, the model adds five additional high-precision iris landmarks per eye (landmarks 468–473), enabling accurate and subpixel-level gaze estimation. This makes it particularly suitable for attention tracking in resource-constrained environments.

**BlazeFace Detector:** the initial face detection component within the MediaPipe pipeline. It is a lightweight, single-shot face detector optimized for mobile and real-time performance, capable of exceeding 200 FPS on modern CPUs. Its robustness to varying face poses and lighting conditions allows it to consistently detect facial regions even in challenging video inputs.

**OpenCV (cv2):** employed for video decoding, frame extraction, color space conversion, and real-time rendering of results. It handles cv2.VideoCapture for reading input video streams, frame-wise resizing, and output generation using cv2.VideoWriter. OpenCV also provides functions for drawing eye landmarks, bounding boxes, and overlaying gaze information on frames.

**NumPy:** used for numerical operations such as gaze ratio computation and time-based accumulation of gaze duration per frame. Its efficient array operations simplify frame-level data tracking for iris position and gaze classification.

**Custom Temporal Filtering Logic:** A custom, lightweight temporal gating mechanism is implemented to improve decision reliability. This logic tracks the duration of both centered and uncentered gaze states and enforces time-based thresholds to confirm attention shifts, minimizing false positives due to transient distractions.

**Inference-Only Pipeline:** The entire system operates using pre-trained and inference-only components. No training or fine-tuning is required at runtime, allowing for quick deployment, minimal computational overhead, and broad hardware compatibility.

## VI. Text Module

### A. Personality Traits from Text

Personality traits can be inferred from linguistic patterns in both spoken and written language. Prior research has demonstrated strong correlations between word usage and the Big Five personality dimensions [14]. For instance, extroverted individuals often use more positive emotion words, while those with high neuroticism scores tend to frequently employ first-person pronouns.

In this module, natural language processing (NLP) techniques and pre-trained language models are employed to analyze interviewees' textual responses and predict personality traits based on these linguistic indicators. Furthermore, the system integrates visual cues—such as facial expressions, gaze direction, and body posture—captured during the interview. By fusing textual and visual modalities, the system aims to provide a more comprehensive and accurate assessment of the interviewee's personality.

#### 1. Dataset

The Essays Dataset was used for both training and evaluation. This publicly available resource consists of 2,468 stream-of-consciousness essays written by psychology students. Each essay is annotated with multiple binary labels corresponding to the Big Five personality traits: Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU). As the dataset contains only textual data, it has been widely adopted in both personality psychology and natural language processing (NLP) research [15]. The distribution of samples across the five traits is summarized in Table 4.

TABLE 4
PERSONALITY TRAIT DISTRIBUTION IN THE ESSAYS DATASET

| Trait | Sample Count | Percentage |
|---|---|---|
| Extraversion (EXT) | 1,276 | 51.70% |
| Openness (OPN) | 1,271 | 49.90% |
| Neuroticism (NEU) | 1,233 | 53.10% |
| Agreeableness (AGR) | 1,310 | 50.79% |
| Conscientiousness (CON) | 1,253 | 51.52% |

#### 2. Feature Extraction

We employed large pre-trained language models, including BERT [16] and ALBERT [17], to extract meaningful features from textual responses. To identify the most effective configuration, we conducted a series of experiments exploring various model settings and preprocessing strategies.

Initially, we analyzed how features extracted from different layers of the models affect performance. Prior studies [18] suggest that lower layers tend to capture syntactic (grammatical) features, while higher layers focus on semantic (meaning-related) information. Based on this, we tested several approaches to feature representation, including using the [CLS] token embedding and averaging the embeddings of all tokens.

In addition, we experimented with multiple text preprocessing methods—such as alternative tokenization schemes and cleaning strategies—to assess their impact on feature quality. Given the 512-token input limit of transformer models, we also evaluated three truncation techniques: retaining the beginning of the text, retaining the end, and combining both (e.g., first 256 and last 256 tokens).

These experiments allowed us to determine the optimal configuration for our task and provided insights into best practices for applying large language models to similar NLP problems.

#### 3. Experiment Configuration

Our models rely on embeddings generated by pre-trained language models as input. To determine the best setup for personality trait prediction, we tested various combinations of embedding models—such as BERT, ALBERT, and GloVe—with both classical and neural classifiers.

For each of the Big Five personality traits, we trained a separate binary classifier to predict whether an individual demonstrates that trait. Our primary architecture was a multi-layer perceptron (MLP) consisting of three hidden layers, using ReLU activations, batch normalization, and dropout to reduce overfitting. The model was optimized using the Adam algorithm [19] with categorical cross-entropy as the loss function.

To ensure robustness and generalization, we applied 10-fold cross-validation, early stopping, and learning rate scheduling (reduction on plateau). Among all tested configurations, the combination of BERT-Large embeddings and the MLP architecture produced the highest accuracy across most traits.

We also evaluated traditional machine learning models such as Support Vector Machines (SVMs), Random Forests, and Logistic Regression. However, these approaches generally performed worse than neural networks when using contextual embeddings, which is consistent with previous findings [20].

**4. Experimental Results**

The classification performance of different text-based models for predicting the Big Five personality traits—Openness (OPN), Conscientiousness (CON), Agreeableness (AGR), Neuroticism (NEU), and Extraversion (EXT)—is presented in Table 5. The models were evaluated using 10-fold cross-validation on the Essays Dataset, with classification accuracy reported as the primary metric.

TABLE 5
PERSONALITY TRAIT PREDICTION ACCURACY USING TEXT MODELS

| *Model* | OPN | CON | AGR | NEU | EXT |
|---|---|---|---|---|---|
| *BERT-Base + MLP* | 0.650 | 0.640 | 0.630 | 0.618 | 0.615 |
| *BERT-Large + MLP* | **0.680** | **0.652** | **0.634** | **0.615** | **0.628** |
| *ALBERT-Base + MLP* | 0.660 | 0.619 | 0.600 | 0.581 | 0.603 |
| *BERT-Large + SVM* | 0.623 | 0.549 | 0.534 | 0.581 | 0.555 |

Among the evaluated configurations, the BERT-Large + MLP model consistently achieved the highest accuracy across all five traits. This result underscores the benefit of deeper transformer architectures paired with neural classifiers. The ALBERT-Base + MLP model, while more efficient, generally exhibited lower performance, likely due to its parameter-reduction strategy.

Traditional machine learning classifiers, such as Support Vector Machines (SVMs) combined with BERT-Large embeddings, underperformed significantly compared to neural approaches. This is consistent with prior findings [20] that contextual embeddings are most effective when paired with deep neural architectures.

**B. Checking the Answer Relevance**

To ensure fair evaluation in automated video interviews, it is critical to verify that candidate responses align with the intent and context of the questions. This relevance checking step affects the accuracy of scoring and hiring decisions. However, challenges arise due to transcription errors, subjective interpretations, and contextual complexity [21].

**1. Text Similarity Techniques**

**i. Cosine Similarity**

```
[CLS] Question [SEP] Answer [SEP]
```

The first method involved calculating cosine similarity between the question and answer embeddings. We used models like BERT, GloVe, and all-MiniLM-L6-v2 to convert both texts into vectors after basic preprocessing (tokenization, lowercasing, and stopword removal).

A similarity score between 0.7 and 0.9 was typically used as the threshold to classify an answer as relevant. While this method is fast and captures some overlap (e.g., shared terms like "project" or "team"), it often gives false positives, especially when irrelevant answers contain common words. This makes it unreliable for nuanced or context-dependent relevance judgments.

**ii. Semantic + Keyword-Based Similarity**

To improve accuracy, we used a hybrid method that considers both the semantic meaning and keyword alignment.
- First, a model answer is created—either manually or generated using GPT-4 or BERT.
- Both the candidate's answer and the model answer are embedded using a transformer (e.g., all-MiniLM-L6-v2), and cosine similarity is computed between them.
- In parallel, we apply TF-IDF to extract and weigh important keywords from both texts.

The final relevance score is computed as a weighted sum.

$$\text{Score} = 0.7 \cdot \text{Semantic Similarity} + 0.3 \cdot \text{TF-IDF Similarity} \quad (1)$$

This method better captures context while still considering keyword overlap.

**Limitations include**:

- The accuracy depends on the quality of the model answer.

- Repetitive use of keywords in shallow answers can still mislead the system.

- Choosing a clear threshold remains difficult due to overlapping score ranges between relevant and irrelevant responses.

**iii. Cross-Encoder**

A cross-encoder model (e.g., ms-marco-MiniLM-L-12-v2) was used to jointly process question–answer pairs for relevance scoring. The input is formatted as:

and passed through a transformer to output a score between 0 and 1. This approach captures deep contextual relationships between the question and answer, offering improved semantic understanding compared to separate embedding methods.

However, performance was limited due to the lack of fine-tuning on hiring-specific data and the presence of transcription errors. The model also misclassified nuanced answers with unusual phrasing and was computationally expensive, making it impractical for large-scale use.

#### iv. LLM-Based Prompting

The final and currently adopted method uses large language models (LLMs) such as Gemini to evaluate the relevance of candidate answers through prompt engineering. This approach leverages the LLM's ability to understand context and semantics without requiring model fine-tuning or complex infrastructure.

Several prompt formats were tested, including those enriched with job role context. However, the most effective version was a concise, instruction-based prompt asking the model to rate the relevance of a question using the candidate's answer as context.

This method proved to be both flexible and effective. Unlike traditional models that require embeddings or supervised training, the LLM understands the question–answer relationship through instructions and produces an interpretable score. It avoids the overhead of training, adapts well to different question styles, and handles subtle nuances in candidate responses.

#### 2. Conclusion

Each method tested for relevance checking in video interview analysis revealed specific limitations, emphasizing the complexity of evaluating candidate responses in hiring scenarios.

**Cosine similarity**, while fast, relied too heavily on surface-level word overlap and failed to reliably distinguish relevant from irrelevant answers.

**Semantic similarity combined with TF-IDF** improved contextual understanding but required high-quality model answers and was prone to false positives from keyword repetition.

**Cross-encoders** offered deeper contextual analysis but underperformed due to a lack of fine-tuning data, transcription noise, and high computational cost.

Prompt engineering with large language models (LLMs)—despite challenges such as sensitivity to phrasing and potential bias—provided the best balance of accuracy, flexibility, and ease of deployment when carefully designed and validated.

Based on these evaluations, the LLM-based prompting approach was selected as the final method due to its strong performance and adaptability across diverse interview contexts.

### C. Text Summarization

Reviewing lengthy interview recordings is often time-consuming for Human Resources (HR) professionals, especially when handling a large number of candidates. To address this challenge, our project focuses on applying text summarization techniques to automatically generate concise summaries of candidate responses extracted from video interviews. The objective is to condense each candidate's spoken content into a comprehensive textual summary that highlights the most important points, such as their qualifications, experience, goals, and personal attributes. This approach allows HR personnel to efficiently access key information without the need to watch entire interview recordings, significantly improving the candidate evaluation process and reducing manual effort. To enable this, two essential steps are required: (1) extracting audio from the video interviews, and (2) transcribing the extracted audio into text for further summarization.

#### 1. Dataset

During the development of the summarization component, we conducted an extensive search for publicly available datasets that contained interview-style speech transcripts paired with reference summaries. However, we found that such datasets were either unavailable or not well-suited to our specific use case, which focuses on candidate-style self-presentations typically encountered in job interviews.

Consequently, we created a custom dataset tailored to the unique requirements of our project. The dataset comprises **800 entries**, each simulating a candidate discussing themselves, their professional background, work experiences, or personal aspirations—closely mirroring the type of responses provided during real job interviews.

**Table.6.** Sample from our custom dataset

| Transcripts | Reference Summaries |
|---|---|
| Hi, I'm Avery Campbell. I've been working as a project manager for 7 years. Recently, I managed cross-functional teams in software rollout. Outside of work, I enjoy volunteering and exploring new technologies. | Avery Campbell, a project manager with 7 years of experience, recently managed cross-functional teams in software rollout. |
| During my time at worked as a UX designer at InnovateX where I led the development of a data pipeline migration. This involved working closely with stakeholders, managing timelines, and ensuring high quality deliverables. The project significantly improved team efficiency and customer satisfaction. | A UX designer at InnovateX led a data pipeline migration, improving efficiency and client outcomes. |
| I'm deeply motivated by the opportunity to solve meaningful problems. I find purpose in contributing to projects that make life easier for people. For example, working on a user-friendly mobile health app made me realize the potential of technology in improving lives. That sense of impact keeps me going. | They are driven by impact and personal growth, often taking the initiative to learn, adapt, and contribute meaningfully to projects. |
| I'm Robin Stewart. I've been working as a Designer, jewellery for around 6 years now. In my last role at Vega, Henry and Craig, I led several successful projects and collaborated with cross-functional teams. I really enjoy problem-solving and creating efficient workflows. One challenge I faced was transitioning to a remote setup, but I quickly adapted by organizing regular check-ins and streamlining communication tools. I'm currently looking for opportunities where I can grow and contribute meaningfully. | Robin Stewart, a Designer, jewellery with 6 years of experience, discussed their work at Vega, Henry and Craig, highlighting project leadership, adaptability to remote work, and interest in growth opportunities. |

To generate this dataset efficiently and with a high degree of realism, we utilized **OpenAI's GPT-4 model**, carefully prompting it to produce diverse, natural-sounding candidate responses along with corresponding human-like summaries. Special attention was given to maintaining linguistic variety, topic diversity, and a realistic conversational tone throughout the dataset to ensure it provided a meaningful basis for evaluating the summarization models.

Each dataset entry consists of:

**Transcripts:** A GPT-4-generated passage simulating a spoken-style candidate response.

**Reference Summary:** A concise summary capturing the key information from the transcript.This custom dataset enabled us to perform a more targeted and context-appropriate evaluation of text summarization methods within the domain of video interview analysis.

**2. Speech-to-Text Transcription Using Whisper**

Following audio extraction, the next critical step is converting the candidate's speech into text. This transcription process was performed using Whisper, an open-source automatic speech recognition (ASR) model developed by OpenAI.

The transcription pipeline begins with language detection, a built-in feature of Whisper. This step is essential to ensure that only English-language audio segments are processed further. If Whisper detects that the spoken language is not

In English, the transcription step is automatically skipped. This filtering mechanism is crucial to maintaining consistency across the dataset and to prevent the generation of low-quality or inaccurate summaries resulting from language mismatches.

For audio segments identified as English, Whisper proceeds to the transcription phase. In this project, we employed the Whisper Medium model, which offers an effective balance between transcription accuracy and computational efficiency. Our selection was guided by the evaluation presented in the original Whisper research paper, where the Medium model demonstrated consistently low Word Error Rates (WER) across a wide range of benchmark datasets. Although larger models,

such as Whisper Large, can achieve even lower ERs, their significantly higher computational requirements made them less practical given the constraints of this project.

The output of this stage is a text transcript of the candidate's spoken responses, provided the detected language is English.

These transcripts serve as the input for the subsequent text summarization step.

## 3. Summarization Models

To identify the most effective summarization approach for condensing candidate transcripts, we evaluated several state-of-the-art pre-trained models using the custom dataset we developed. This comparative evaluation aimed to determine which model produces the most informative, fluent, and concise summaries aligned with the specific requirements of job interview contexts.

**Models Evaluation:**

### i. BART-Large-CNN

One of the primary models used in this study is **BART-Large**, a powerful transformer-based architecture pre-trained on large-scale English corpora and fine-tuned on the CNN/DailyMail summarization dataset. BART (Bidirectional and Auto-Regressive Transformers) combines a BERT-like encoder and a GPT-like decoder, making it

particularly effective for **abstractive summarization**. This structure enables BART to both understand input context and generate coherent, paraphrased summaries.

### ii. FLAN-T5 Large

We also evaluated **FLAN-T5 Large**, an instruction-tuned variant of Google's T5 (Text-to-Text Transfer Transformer) model. T5 frames all natural language processing tasks as text-to-text problems, where both inputs and outputs are treated as textual sequences. FLAN-T5 distinguishes itself through fine-tuning on a broad range of instruction-based datasets, enhancing its ability to follow detailed prompts, even in zero-shot or few-shot settings. The **FLAN-T5 Large** model was selected for its balance between performance and computational efficiency. Summarization was guided using a carefully designed prompt that simulates the expectations of a professional recruiter:

**Prompt Used for Summarization**

> You are an AI assistant specializing in summarizing job interview responses. Your task is to generate a clear, well-structured, and natural-sounding summary of the candidate's answer while keeping it concise and professional.
>
> - Ensure the summary is written in a natural, flowing paragraph, not bullet points.
>
> - Maintain the key ideas from the response while eliminating unnecessary details.
>
> - Keep the tone formal, coherent, and human-like, as if written by a professional recruiter.
>
> Here is the candidate's response: `"transcript"`

Now, provide a well-written summary in paragraph form that retains the most important information from the response."

### iii. Gemini 2.0 Flash

Additionally, we evaluated **Gemini 2.0 Flash**, a cutting-edge large language model developed by Google DeepMind. Gemini was prompted using the same structured instruction applied to the other models, ensuring a consistent evaluation setup. Despite being a general-purpose model, Gemini demonstrated strong performance in text summarization when guided by task-specific prompts, effectively capturing key information and generating coherent summaries.

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**

ROUGE measures the quality of generated summaries by comparing them to reference summaries based on word overlap and sequence similarity.

**ROUGE-1:** Measures unigram (single-word) overlap.

Recall: Overlapping unigrams ÷ Total unigrams in the reference summary.

Precision: Overlapping unigrams ÷ Total unigrams in the generated summary.

F1 Score: 2 × (Precision × Recall) ÷ (Precision + Recall)

**ROUGE-2:** Measures bigram (two-word sequence) overlap.

Recall: Overlapping bigrams ÷ Total bigrams in the reference summary.

Precision: Overlapping bigrams ÷ Total bigrams in the generated summary.

F1 Score: 2 × (Precision × Recall) ÷ (Precision + Recall)

**ROUGE-L:** Measures the longest common subsequence (LCS) between the generated and reference summaries, capturing fluency and word order.

Recall: LCS length ÷ Length of reference summary.

Precision: LCS length ÷ Length of generated summary.

F1 Score: 2 × (Precision × Recall) ÷ (Precision + Recall)

**ROUGE-Lsum:** An extension of ROUGE-L designed for multi-sentence summaries.

Recall: LCS length ÷ Total words in the reference summary.

Precision: LCS length ÷ Total words in the generated summary.

F1 Score: 2 × (Precision × Recall) ÷ (Precision + Recall)

**BLEU (Bilingual Evaluation Understudy)**

BLEU is traditionally used in machine translation but is also applicable for summarization. It focuses on **precision**, penalizing models that overuse paraphrasing or generate overly brief outputs.

**BLEU Score Formula:**

BLEU = Brevity Penalty (BP) × exp ($\Sigma_n$ $w_n$ log($p_n$)) Where:

**BP**: A penalty applied when the generated summary is shorter than the reference.

**$w_n$**: Weight assigned to each n-gram level. In BLEU-4, each is weighted at 0.25.

**$p_n$**: Precision for each n-gram level (unigrams, bigrams, trigrams, 4-grams).

**N-gram Precisions:**

**$p_1$**: Unigram precision.

**$p_2$**: Bigram precision.

**$p_3$**: Trigram precision.

**$p_4$**: 4-gram precision.

In this study, BLEU was calculated using the **BLEU-4 configuration**, which provides a comprehensive assessment by considering overlaps at various n-gram levels. This approach balances sensitivity to both short and long phrase matches, offering a more nuanced evaluation.

**BERTScore**

Unlike ROUGE and BLEU, **BERTScore** evaluates semantic similarity using contextual embeddings from pre-trained transformer models, making it particularly sensitive to meaning rather than exact word overlap.

**BERTScore Precision:**

Sum of maximum similarity scores for each token in the generated summary to tokens in the reference summary ÷ Total tokens in the generated summary.

**BERTScore Recall:**

Sum of maximum similarity scores for each token in the reference summary to tokens in the generated summary ÷ Total tokens in the reference summary.

**BERTScore F1:**

2 × (Precision × Recall) ÷ (Precision + Recall)

BERTScore offers a complementary perspective by capturing semantic equivalence, even when surface-level wording differs between the generated and reference summaries.

Table 7: Summarization Models Performance Comparison

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Lsum | BLEU | BERT Score (F1) |
|---|---|---|---|---|---|---|
| BART-Large | **0.4930** | **0.2686** | **0.4345** | **0.4344** | **0.1738** | **0.9110** |
| FLAN-T5 Large | 0.3530 | 0.2164 | 0.3425 | 0.3423 | 0.1117 | 0.8996 |
| Gemini 2.0 Flash | 0.4530 | 0.2373 | 0.4073 | 0.4072 | 0.1047 | 0.9079 |

## 4. Model Performance and Analysis

During our evaluation, we observed that the **ROUGE and BLEU scores** across all tested models were relatively low. To understand this result, we reviewed relevant literature and confirmed that this is a **well-known limitation in Natural Language Generation (NLG) tasks**. Both ROUGE and BLEU are based on surface-level n-gram overlap and do not account for **semantic meaning**. As a result, these metrics tend to penalize summaries that use alternative phrasing, paraphrasing, or synonymous expressions, even when the generated summaries accurately convey the intended information.

This limitation is particularly significant in **summarization tasks**, where multiple valid summaries can express the same content using different words or structures. Therefore, low ROUGE and BLEU scores do not necessarily reflect poor summarization quality but rather highlight the lexical differences between the generated and reference summaries.

In contrast, **BERTScore** provides a more reliable evaluation in this context, as it measures semantic similarity using contextual embeddings derived from transformer-based language models. BERTScore is less sensitive to exact word matches and better captures the underlying meaning of the text.

**Model Performance Summary**

**i. BART-Large-CNN:**

Achieved the highest ROUGE-1 (0.4930), ROUGE-2 (0.2686), ROUGE-L (0.4345), and ROUGE-Lsum (0.4344) scores.

Recorded the highest BERTScore F1 (0.9110), indicating strong semantic alignment with reference summaries.

These results suggest that BART-Large-CNN generated summaries that were both lexically and semantically close to the reference summaries.

**ii. FLAN-T5 Large:**

Scored the lowest in ROUGE and BLEU metrics but achieved a competitive BERTScore F1 of 0.8996.

The lower n-gram-based scores may be attributed to the model's tendency to paraphrase, which is penalized by ROUGE and BLEU but is less problematic from a semantic perspective.

**iii. Gemini 2.0 Flash:**

Achieved moderate ROUGE scores, indicating reasonable lexical overlap with the references.

Recorded the lowest BLEU score (0.1047), potentially due to higher variability in word choice and sentence structure.

Maintained a strong BERTScore F1 of 0.9079, suggesting good semantic preservation comparable to BART-Large.

**iv. Human Evaluation**

It is critical to note that **human judgment remains the gold standard** when assessing the quality of text summarization. While automated metrics such as ROUGE, BLEU, and BERTScore offer valuable quantitative insights, they cannot fully capture **semantic coherence, informativeness, and readability** as perceived by human evaluators.

During our manual review of the summaries generated by BART-Large, FLAN-T5 Large, and Gemini, we found that **Gemini consistently produced superior summaries**. Its outputs were more informative, retained key details effectively, and exhibited a more natural and human-like writing style compared to the other models, despite BART-Large achieving marginally higher BERTScore F1 values.

The average BERTScore F1 results were as follows:

BART-Large: **0.9110**

Gemini: **0.9079**

Although the numerical difference was minimal, **the qualitative difference was significant**. Based on this observation, we concluded that **Gemini 2.0 Flash was the most reliable and effective model** for our summarization task. Accordingly, we selected Gemini as the preferred model for generating the final candidate summary reports in this project.

## Sample outputs for each model:

**Transcript 1:**

Hello, my name is Nabil Samaham. I am 24 years old. I am from Syria. I am a first-year master's student in robotics and artificial intelligence. I finished my bachelor's degree when I was in Syria, and then I moved to Russia to study for my master's degree.Actually, I have many skills that can help me get a job in particular companies. For example, I have no problem working under stress, I enjoy working in teams, and I always meet deadlines. I do not like to exceed deadlines, and I am always on time.However, I do not enjoy sales work very much because I don't think I'm good at convincing others to buy products from me. If I were to do sales, it would likely be for financial reasons—or if I invented a new product and needed investors to turn my prototype into a marketable product that benefits others.Over the past three years, my proudest achievement has been completing my bachelor's degree in mechatronics engineering, becoming an engineer, and getting the opportunity to move to Russia to pursue my master's degree.

**Gemini 2.0 Flash:**

Nabil is a 24-year-old first-year master's student in Robotics and Artificial Intelligence, originally from Syria. He completed his bachelor's degree in Mechatronics Engineering before relocating to Russia for his master's studies. He highlights his ability to work well under pressure, collaborate effectively in teams, and consistently meet deadlines. While he is not particularly drawn to sales roles due to a perceived lack of persuasiveness, he would consider it if it involved securing investment for a product he developed. Nabil considers completing his bachelor's degree and pursuing his master's in Russia as his most significant achievements over the past three years.

**BART-Large-cnn:**

Nabil Samaham is a first-year master's student in Robotics and Artificial Intelligence. He believes he possesses several valuable skills that can help him secure employment in specific companies. However, he is not fond of sales roles, as he feels that persuasion is not one of his strengths.

**Flan T5:**

Nabil Samaham is a first year master student in robotics and artificial intelligence. He is from Syria and he is studying his master degree in Russia

**Transcript 2:**

Hi, I'm Mario Meza. I've been working as a Field seismologist for around 5 years now. In my last role at Mccarthy LLC, I led several successful projects and collaborated with cross-functional teams. I really enjoy problem-solving and creating efficient workflows. One challenge I faced was transitioning to a remote setup, but I quickly adapted by organizing regular check-ins and streamlining communication tools. I'm currently looking for opportunities where I can grow and contribute meaningfully.

**Gemini 2.0 Flash:**

Mario Meza is a Field Seismologist with approximately five years of experience, most recently at Mccarthy LLC, where he led successful projects and fostered collaboration within cross-functional teams. He enjoys problem-solving and workflow optimization. He demonstrated adaptability by successfully transitioning to a remote work environment through proactive communication strategies. Mr. Meza is seeking an opportunity for professional growth and meaningful contribution.

**BART-Large-CNN :**

Mario Meza has been working as a Field seismologist for around 5 years. In his last role at Mccarthy LLC, he led several successful projects. He is currently looking for opportunities where he can grow and contribute meaningfully.

**FLAN T5:**

Mario Meza is a Field seismologist. He's looking for a new job

## VIII. Conclusion

This paper has presented PRVIA, a fully end-to-end, align-free framework that integrates deep learning across audio, text, and video modalities to automate pre-recorded interview assessment. The Video Module captures facial expressions, head gestures, and personality traits while detecting potential cheating behaviors; the Audio Module performs automatic speech recognition and pronunciation evaluation without phoneme-level alignment; and the Text Module generates concise summaries and checks semantic relevance,. A unified Report Generation Module then aggregates these multimodal insights into standardized, interactive candidate reports, enabling HR teams to conduct fairer, faster, and more consistent screening. By eliminating manual bottlenecks and reducing evaluative bias, PRVIA establishes a new baseline for scalable, practical interview analysis.

## VIII. Future Work

While PRVIA provides a strong foundation for automating the video interview evaluation process, there are several areas where the system can be further enhanced and expanded in future iterations:

**1. Model Improvements**

Future versions of the system can incorporate more advanced and fine-tuned models that are specifically trained on interview datasets. Additionally, exploring larger language models or fine-tuning existing models on domain-specific data could improve both the accuracy and naturalness of the generated summaries.

**3. Faster Model Inference**

Future improvements could aim to optimize model latency and response time by integrating faster, lightweight models or using model distillation techniques to enable quicker analysis, especially in high-traffic environments.

**4. Integration with Applicant Tracking Systems (ATS)**

A valuable next step is to integrate the PRVIA system with existing Applicant Tracking Systems (ATS) to provide a seamless experience for HR professionals. This would allow automatic syncing of candidate reports, video submissions, and summaries directly into the company's recruitment management platform.

**5. Development of an HR Virtual Agent**

Building a smart HR agent that can autonomously interact with candidates during the interview process is another promising direction. This agent could ask pre-defined or dynamic interview questions, follow up with candidates in real-time, and guide them through the video submission process without human supervision.

**6. Real-Time Video Analysis**

Future versions can aim to support real-time video analysis and live interview processing, providing instant feedback to both candidates and HR teams. This would significantly reduce waiting times and enable more interactive recruitment experiences.

**7. Chatbot Integration**

Developing an AI-powered chatbot to assist candidates throughout the application and interview process would enhance user experience. The chatbot could answer questions, provide interview instructions, and offer feedback in real-time, making the system more user-friendly and accessible.

## IX. References


[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, vol. 30, no. 2–3, pp. 95–108, 2000.

[2] F. Zhang et al., "Automatic mispronunciation detection for Mandarin," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 4385–4388.

[3] Y. Gong et al., "Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment," in ICASSP 2022, pp. 8232–8236.

[4] L. Peng et al., "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis," in Proc. Interspeech, 2021.

[5] K. Fu et al., "Pronunciation Assessment with Multi-modal Large Language Models," arXiv preprint arXiv:2407.09209, 2024.

[6] J. H. M. Wong et al., "Variations of multi-task learning for spoken language assessment," in Proc. Interspeech, 2022, pp. 4456–4460.

[7] Y.-W. Chen et al., "Multipa: a multi-task speech pronunciation assessment system for a closed and open response scenario," arXiv preprint arXiv:2308.12490, 2024.

[8] H. Do et al., "Hierarchical pronunciation assessment with multi-aspect attention," in ICASSP 2023, pp. 1–5.

[9] E. Kim et al., "Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning," in Proc. Interspeech, 2022, pp. 1411–1415.

[10] A. Q. Jiang et al., "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.

[11] Y. Labrak et al., "BioMistral: a collection of Open-Source pretrained large language models for medical domains," arXiv preprint arXiv:2402.10373, 2024.

[12] Gemini Team, "GemMa: Open models based on Gemini research and technology," arXiv preprint arXiv:2403.08295, 2024.

[13] Label Studio, "Fine-Tuning Llama 3: Enhancing Accuracy in Medical Q&A with LLMs," Jun. 2024. [Online]. Available: https://labelstud.io/blog/fine-tuning-llama-3-enhancing-accuracy-inmedical-q-and-a-with-llms/

[14] H. M. C. N. Wijerathne et al., "Smart recruitment tool with AI Technology," in Proc. ICCCNT, 2023, pp. 1–8.

[15] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," J. Pers. Soc. Psychol., vol. 77, no. 6, pp. 1296–1312, 1999.

[16] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[17] Z. Lan et al., "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," in Proc. ICLR, 2020.

[18] I. Tenney et al., "BERT Rediscovers the Classical NLP Pipeline," in Proc. ACL, 2019, pp. 4593–4601.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learn. Representations (ICLR), 2015.

[20] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[21] T. N. Nguyen et al., "Trust AI: A Human-Centered Perspective on Trustworthy Automated Hiring," in Proc. ACL, 2023, pp. 1333–1353.

[22] Y. Li et al., "Cr-net: A deep classification-regression network for multimodal apparent personality analysis," Int. J. Comput. Vision, vol. 128, pp. 2763–2780, 2020.

[23] A. Xenos et al., "VLLMs Provide Better Context for Emotion Understanding through Common Sense Reasoning," arXiv preprint arXiv:2404.07078, 2024.

[24] J. Lee et al., "Context-aware emotion recognition networks," in Proc. IEEE/CVF Int. Conf. on Computer Vision, 2019, pp. 10143–10152.

[25] L. Chen et al., "Automated video interview judgment on a large-sized corpus collected online," in Proc. ACII, 2017, pp. 504–509.

[26] I. Chakraborty et al., "AI and AI-Human Based Salesforce Hiring using Conversational Interview Videos," SSRN preprint 4137872, 2023.

[27] Junbo Zhang, Zhiwen Zhang, et al., "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in Interspeech, 2021.