# Agentic Video Interview Analyzer — Comprehensive Specification

This document explains the end-to-end flow, inputs/outputs, prompts, evaluation criteria, thresholds, weightages, decision logic, and expectations for the Agentic Video Interview Analyzer used for the Ambassador Program assessments.

Audience: Engineering, Product, and Leadership stakeholders. Use this to review design, tune thresholds/weights, and provide feedback.

# 1) High-Level Overview

- **Goal**: Automatically assess candidate video interviews across identity verification, technical video quality, speech transcription, content quality, behavioral signals, and aggregate to a pass/review/fail decision with transparent reasoning.
- **Core Components (Agents)**:
  - Identity Verification (OCR + name match + face match)
  - Video Quality Assessment (OpenCV-based metrics)
  - Speech-to-Text Transcription (Google Cloud Speech-to-Text)
  - Batched Content + Behavioral Evaluation (single Gemini call)
  - Decision Aggregation (weighting + thresholds + LLM rationale)
- **Optimization**: Two execution modes
  - Sequential LangGraph workflow (`app/agents/graph.py`)
  - Optimized semi-parallel workflow (`app/agents/graph_optimized.py`)

# 2) Execution Flow

## 2.1 Sequential Flow (default)

Order: Identity → Quality → Transcription → Batched Evaluation (Content + Behavioral) → Aggregation

```
# START → verify_identity → check_quality → transcribe_videos → batched_evaluation → aggregate_decision → END
```

Notes:

- All agents run regardless of identity success to collect full evidence (fraud detection and transparency).
- Batched evaluation consolidates multiple LLM calls into a single call.

## 2.2 Optimized Flow (30–45s typical)

Phases: Preparation → Semi-parallel processing → Batched evaluation → Aggregation → Cleanup

```
# Phase 2: Quality + Transcription in parallel, then Identity; merge results; run batched evaluation; aggregate decision
```

# 3) Inputs — What the System Expects

Required per assessment (see `InterviewState`):

- `user_id` (string)
- `username` (string; candidate's full name for name matching)
- `profile_pic_url` (GCS URL)
- `gov_id_url` (GCS URL)
- `video_urls` (list of 5 videos):
  - `video_0`: identity check video
  - `video_1`–`video_5`: interview question responses
- `interview_questions` (list of 5 dicts): each has `question_number`, `question`, `goal`, `criteria`

```
class InterviewState(TypedDict):
    # inputs, questions, and agent outputs schema
```

# 4) Agent Nodes — Details, Criteria, Thresholds

## 4.1 Identity Verification

- Components:
  - OCR on government ID via Google Vision to extract text and heuristic name (`extract_text_from_image`, `extract_name_from_text`).
  - Name similarity to provided `username` with multi-step matching.
    - Threshold: name similarity ≥ 50% considered match.
  - Face match using DeepFace (ArcFace) between `profile_pic` and a frame extracted from `video_0`.
    - Face verified if similarity ≥ 60% (lenient threshold logic).
  - Overall identity verified only if both name_match and face_verified are true.
- Confidence score: 50% name similarity + 50% face similarity.
- Failure does not block evaluation; it produces red flags and affects concerns in aggregation.

Key thresholds and logic:

- Name match threshold: 50% similarity.
- Face verified: average similarity ≥ 60% on `video_0` check.
- Combined identity confidence = 0.5 × name_similarity + 0.5 × avg_face_confidence.

```
# Name similarity ≥ 50%; face verification on video_0; overall verified requires both
```

Red flags captured (examples):

- Name mismatch, low similarity, face verification failure, extraction errors.

## 4.2 Video Quality Assessment

- Metrics (sampled frames via OpenCV): resolution, fps, duration, brightness, sharpness (Laplacian variance), face visibility ratio, multiple faces.
- Issues flagged only for: "Blurry/out of focus" and "Poor face visibility".
- Quality score (0–100) weighted:
  - Resolution 25%, FPS 15%, Brightness 20%, Sharpness 20%, Face visibility 20%.
- Pass criterion: overall quality score ≥ 60.

```
# Scoring weights; pass if overall_score ≥ 60
```

## 4.3 Speech-to-Text Transcription

- Google Cloud Speech-to-Text, FLAC 16kHz mono.
- Auto-selects LongRunningRecognize for audio ≥ 60s, otherwise synchronous recognize.
- Outputs per video: transcript, confidence, word_count, filler_words, duration.
- Aggregates: `avg_confidence` across videos, `total_words`.

Important details:

- Language code: `en-IN`.
- Filler words counted: ["um", "uh", "like", "you know", "basically", "actually"].

```
# Config, long-running, confidence aggregation, filler words
```

## 4.4 Batched Content + Behavioral Evaluation (Single LLM Call)

- Consolidates analysis for all 5 questions and behavioral signals in a single Gemini call.
- System prompt enforces JSON-only return.
- The batched prompt includes:

    - Identity summary (verified flag + confidence)
    - Full question text, goals, and criteria
    - All transcripts
    - Strict instructions and JSON schema for outputs

Per-question pass threshold:

- Score ≥ 70 ⇒ pass
- `questions_passed`/`questions_failed` computed accordingly

Behavioral analysis output includes: `behavioral_score`, confidence/engagement levels, stress indicators (0–10), authenticity, communication clarity, strengths, concerns, red_flags, overall impression.

```
# JSON schema + IMPORTANT: Score ≥ 70 = pass for each question
```

Prompts (structure summary):

- System: "You are an expert interview evaluator. Return ONLY valid JSON, no markdown."
- Human: Includes identity context, questions/goals/criteria, transcripts, and the full JSON output schema with guidance:
    - Be strict
    - Score ≥ 70 = Pass
    - Look for specific evidence
    - Red flags: self-centered motives, negative language, lack of empathy

### 4.5 Decision Aggregation

- Final score = 70% content overall score + 30% behavioral score.
- Identity and Quality are gatekeepers (0% weight). Their issues influence red flags/concerns but not numeric weighting.
- Decision thresholds:
    - PASS: final_score ≥ 70
    - REVIEW: 60 ≤ final_score < 70
    - FAIL: final_score < 60
- Identity failure does not auto-fail; it adds red flags and concerns and is visible to reviewers.
- LLM reasoning: A concise 3–4 sentence rationale is generated to explain decision.

```
# PASS ≥ 70, REVIEW 60–69, FAIL < 60; weights Content 70% + Behavioral 30%
```

## 5) Weightages and Thresholds (Quick Reference)

- Identity
    - Name match: ≥ 50% similarity to pass name check
    - Face match: ≥ 60% average similarity on video_0
    - Overall verified: name_match AND face_verified
    - Confidence: 50% name + 50% face
    - Weight in final score: 0% (gatekeeper; impacts concerns/red flags)
- Quality
    - Resolution 25%, FPS 15%, Brightness 20%, Sharpness 20%, Face visibility 20%
    - Pass: overall quality ≥ 60 (used for concerns/red_flags)
    - Weight in final score: 0%
- Transcription
    - Avg confidence reported; influences behavioral signals and concerns
    - Language: en-IN; LongRunningRecognize for ≥ 60s
- Content (per question in batched evaluation)
    - Pass threshold: score ≥ 70
    - Overall content score: average of 5 question scores

    - Weight in final score: 70%

- Behavioral

    - Outputs include behavioral_score (0–100), confidence/engagement, stress indicators 0–10, authenticity, clarity
    - Weight in final score: 30%

- Final Decision

    - PASS: final_score ≥ 70
    - REVIEW: 60–69
    - FAIL: < 60

# 6) Prompts — What We Ask the Model(s)

- Batched Evaluation System Prompt:

    - "You are an expert interview evaluator. Return ONLY valid JSON, no markdown."

- Batched Evaluation Human Prompt (structure):

    - Identity summary (verified + confidence)
    - Questions with goals and criteria
    - Candidate transcripts for Q1–Q5
    - Strict JSON schema for content_evaluation and behavioral_analysis
    - Guidance: be strict, score ≥ 70 passes, look for evidence, list red flags

- Aggregation Reasoning Prompt:

    - "You are a hiring decision expert. Based on the assessment data, provide: 1) reasoning; 2) strengths; 3) concerns; 4) final recommendation. Keep 3–4 sentences."

# 7) Outputs — What We Produce

`final_decision` object includes:

- `decision`: PASS | REVIEW | FAIL
- `final_score`: weighted (Content 70% + Behavioral 30%)
- `confidence_level`: heuristic based on distance from 75
- `component_scores`: identity, quality, content, behavioral, transcription(avg_confidence×100)
- `weighted_breakdown`: contributions (identity/quality/transcription = 0 weight)
- `reasoning`: 3–4 sentence LLM rationale
- `recommendation`: action string
- `strengths`, `concerns`, `red_flags`: lists compiled from all agents

```
# final_decision structure with weights and narrative explanations
```

# 8) Red Flags and Concerns

- Identity: name mismatch, face verification failure, low similarity, OCR errors, age/gender discrepancy flags
- Quality: poor face visibility (< 50%), blurry/out-of-focus
- Transcription: low confidence (< 70%), no speech detected
- Content: most/all questions failed, low overall content score, specific red-flag phrases
- Behavioral: very low behavioral_score (< 30) or indicators of stress/disengagement; inappropriate tone

# 9) Operational Notes

- Models

    - LLM: Gemini (gemini-2.0-flash-exp for batched eval and reasoning; gemini-2.5-flash-exp in content module)

  - Face: DeepFace with ArcFace backend
  - OCR: Google Cloud Vision
  - Transcription: Google Cloud Speech-to-Text (en-IN)

- Performance

  - Optimized run: ~30–45 seconds typical (semi-parallel)
  - Sequential run: 4–5 minutes historically

- Resilience

  - Every node has exception handling; on failures, defaults are set and errors appended to `state.errors`
  - Identity/Quality failures become red flags but do not block evaluation

# 10) What Feedback We’re Seeking

- Are the final decision thresholds appropriate? (PASS ≥ 70, REVIEW 60–69)
- Are weightages right for our hiring bar? (Content 70% / Behavioral 30%)
- Should identity/quality contribute a small numeric weight or remain gatekeepers only?
- Are per-question pass thresholds (≥ 70) and "strict but evidence-based" guidance correct?
- Should we tune name similarity (50%) or face similarity (60%) thresholds?
- Any additional red flags/concerns to auto-detect at aggregation time?
- Is the LLM rationale sufficient for reviewers, or should we expand to include evidence snippets?

# 11) Appendix — State Schema (abridged)

```
# InterviewState contains: inputs, questions, agent outputs, control flow, errors, metadata
```

# 12) Change/Tuning Log (to track decisions)

- Weights updated per CTO direction: Content 70%, Behavioral 30%; Identity and Quality are gatekeepers (0 weight).
- Identity name similarity relaxed to ≥ 50%; face similarity pass at ≥ 60% on video_0.
- Batched evaluation consolidates content + behavioral into one call to reduce latency and cost.

— End —