

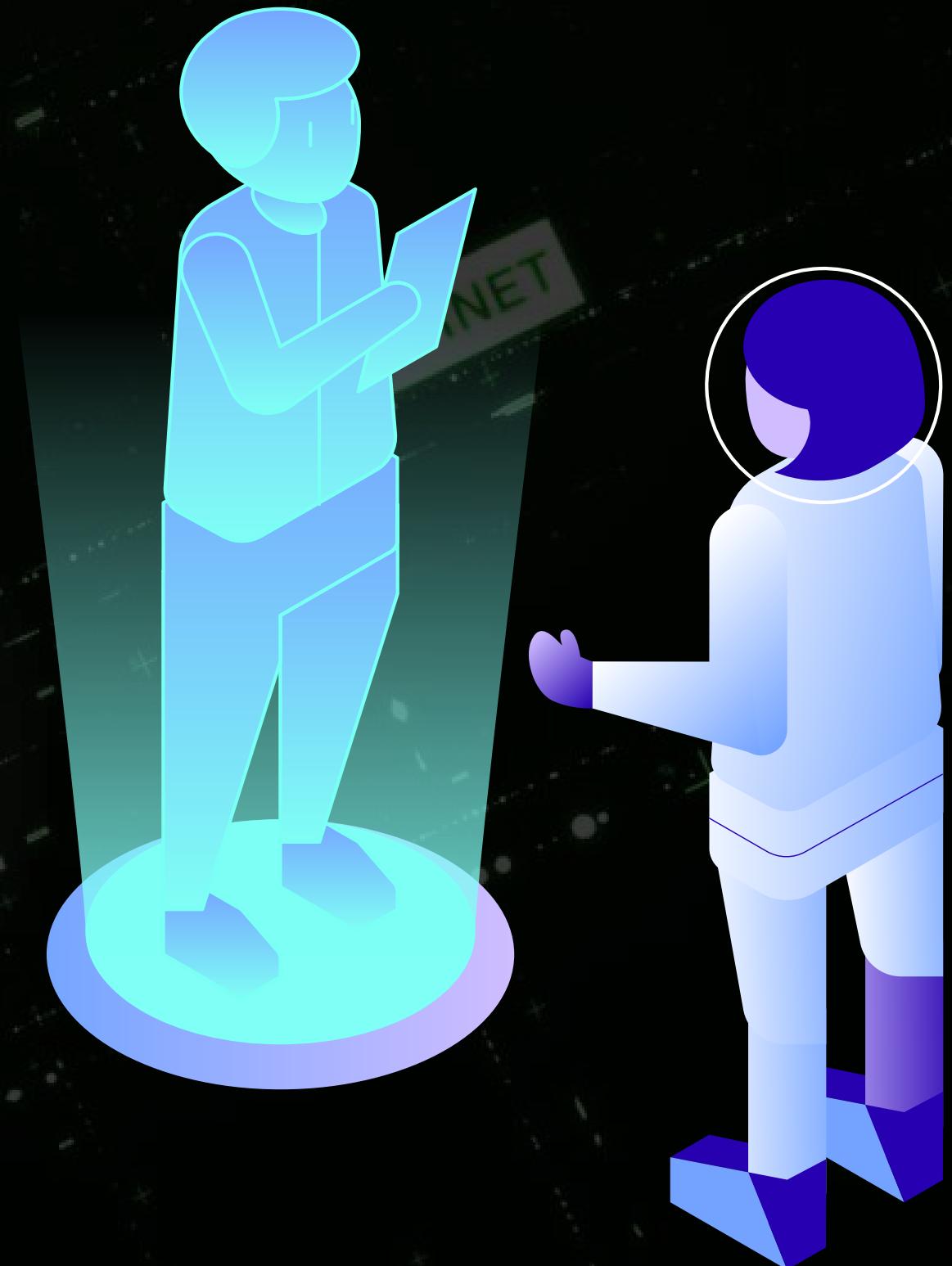
Faculty of Computers & Artificial Intelligence

Helwan University

AI Department

VIRTUAL AI INTERVIEWER

Supervisor : Dr. Soha Ahmed Ehssan



CONTENT

- 1. INTRODUCTION**
- 2. LITERATURE REVIEW**
- 3. SYSTEM ARCHITECTURE**
- 4. SYSTEM METHODOLOGY**
- 5. IMPLEMENTATION**
- 6. EVALUATION AND RESULTS**
- 7. CONCLUSION & FUTURE WORK**



1. INTRODUCTION

INTERNET

WHAT'S WRONG WITH TRADITIONAL INTERVIEWS?

01

We found over 60% of interviews show signs of bias , often leading to unfair hiring and increased candidate stress.

02

Limited access to realistic practice and feedback,

03

A 2022 study by SHRM and Glassdoor found that over 65% of recruiters struggle to fairly evaluate large applicant pools due to high volume and time constraints.

1. INTRODUCTION

Continued..

01

No stress anymore

02

AI scoring can match or exceed human **fairness**.

03

Provide candidates with detailed, **personalized** reports.

04

Automation reduces recruiter **workload** and improves outcomes.

WHAT ARE WE TRYING TO PROVE?

2. LITERATURE REVIEW

INTERNET

WHAT DOES THE LITERATURE SAY?

01

Most rely on **text/audio only**

02

limited feedback and **no real-time adaptivity**

03

They lack **multimodal emotion analysis**

A 2024 **IEEE Access** review on AI recruitment tools found that **76%** of systems use only text and voice input, with no video or visual emotion integration.

A 2024 **Springer AI & Ethics** study reported that **68%** of AI interview platforms provide **static question sets** and **delayed or generic feedback**

According to a 2024 **ACM Transactions on HCI** paper, only **12%** of evaluated platforms support multimodal emotion detection

HOW DOES OUR SYSTEM PUSH BEYOND EXISTING WORK?

01

MULTIMODAL INPUT

Voice, face, and
text fusion for
deeper analysis

03

FAIR SCORINGS

Judgment agent
(Mistral) with
explainable
feedback

02

SMART QUESTIONS

RAG + LLaMA for role-
based, adaptive
interviewing

04

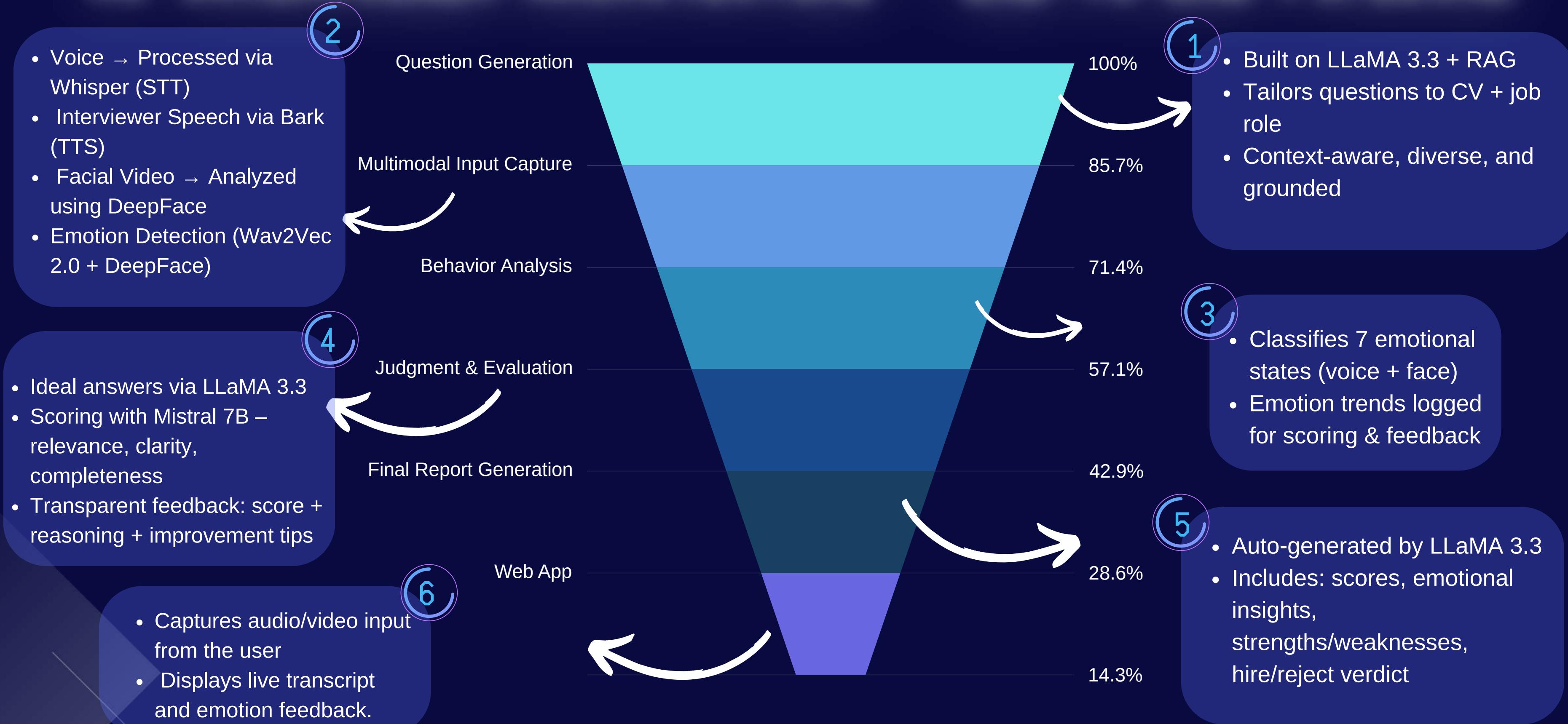
INSIGHTFUL REPORTING
Generates detailed
reports with scores,
emotions & a clear
hire/reject verdict

3. SYSTEM ARCHITCTURE

INTERNET

3. SYSTEM ARCHITECTURE

AI INTERVIEWER ARCHITECTURE - END-TO-END PIPELINE



4. SYSTEM METHODOLOGY

INTERNET

1. KEY AI MODELS & THEIR ROLES

LLAMA 3.3 (70B)

01

- Generates context-aware, domain-specific interview questions
- Also used to create high-quality reference answers for evaluation

BARK (TTS)

02

- Evaluates candidate answers based on relevance, clarity, and accuracy
- Produces scores + natural language feedback for fairness and explainability

WHISPER (BASE)

03

- Converts spoken answers into accurate transcripts in real time
- Handles varied accents, informal tone, and background noise

1. KEY AI MODELS & THEIR ROLES

WAV2VEC 2.0

- Detects emotions in voice (e.g., stress, confidence, calmness)
- Fine-tuned on labeled datasets to recognize real-time vocal cues

DEEPFACE

- Analyzes facial expressions from video to classify emotional states.
- Processes sampled frames for performance efficiency and emotion trends.

MISTRAL-7B-INSTRUCT

- Evaluates candidate answers based on relevance, clarity, and accuracy
- Produces scores + natural language feedback for fairness and explainability

2. PREPROCESSING

AUDIO

- Resampled to standard frequencies (16kHz for Whisper, 22kHz for wav2vec)
- MFCCs (Mel-frequency cepstral coefficients) extracted for emotion features.

VIDEO

DeepFace runs on every 15th frame from webcam video for efficient facial emotion detection.

TEXT & RETRIEVAL

Semantic search using embeddings retrieves the most relevant Q&A pairs for LLaMA to ground its question generation

3. TRAINING DATA

WAV2VEC 2.0

Was fine-tuned on
emotional speech datasets

- RAVDESS
- TESS
- EMO-DB
- CREMA-D

&

RAG

Job-role-specific Q&A pairs
Embedded and indexed with
[Qdrant](#) to enable accurate
retrieval in RAG-based
question generation.

- Total Questions: **3,809**
- Unique Job Roles: **26**

5. IMPLEMENTATION WHAT HAPPENS BEHIND THE SCENES ?!

INTERNET

LOCAL DEPLOYMENT & DEVELOPMENT ENVIRONMENT

DEVELOPED IN PYTHON, OPTIMIZED FOR LOCAL MACHINES WITH
CONSUMER-GRADE CPU/GPU.

VS CODE

development and
writing code

GITHUB

for version control and
team collaboration

KAGGLE / GOOGLE COLAB

for model training and
experimentation

5. IMPLEMENTATION

IMPLEMENTATION FLOW

THE USER ENTERS HIS/HER INFORMATION:

Missing Information

Name

Job Role

Seniority
Entry-level

Skills

Submit

Candidate Information

Upload CV

Drop File Here
or
Click to Upload

Job Description

Continue

Missing Information

Seniority
Senior

Submit

IMPLEMENTATION FLOW

INTERVIEW_LOOP() DRIVES THE SESSION:



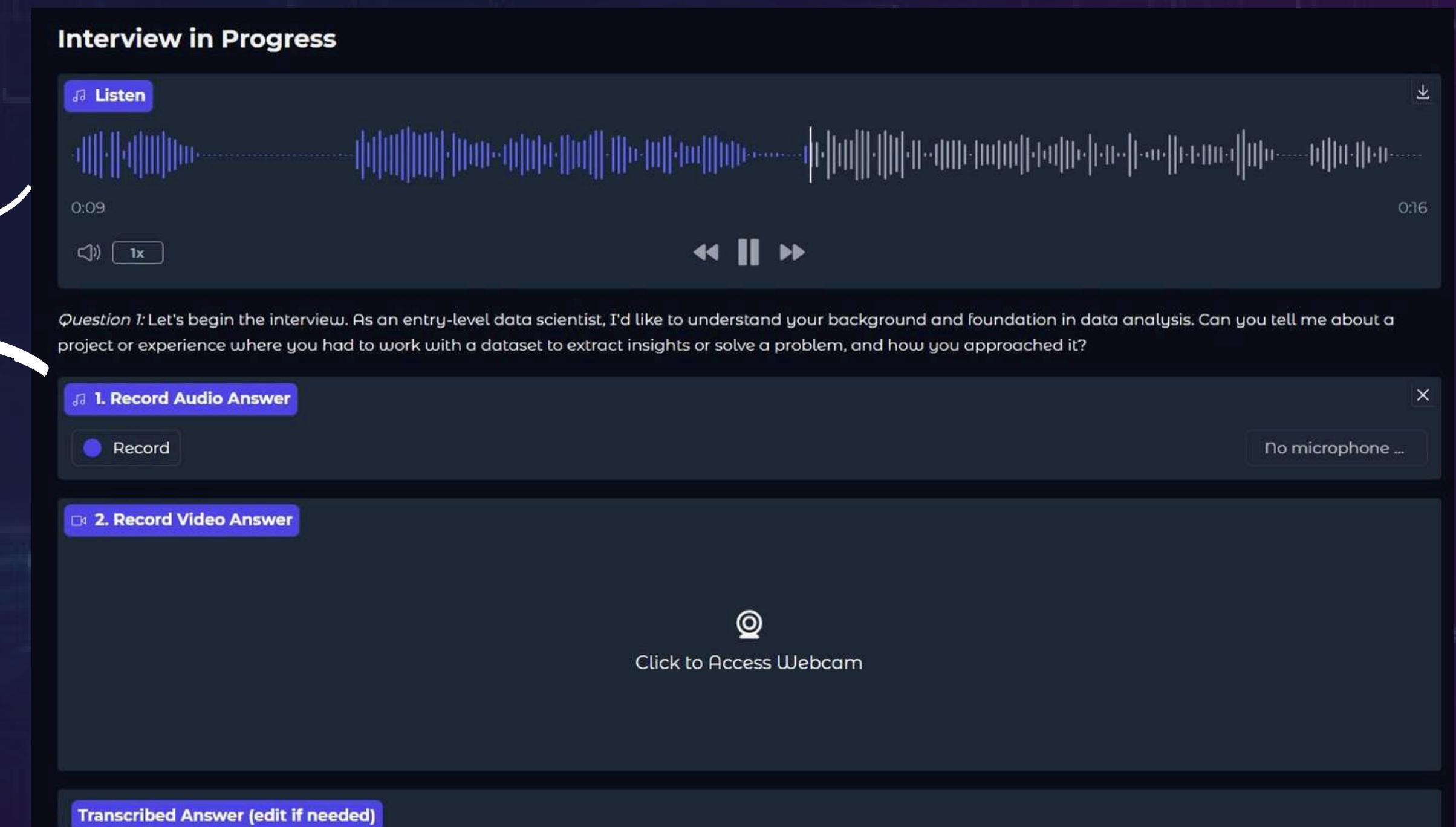
Start Interview

LETS SHOW HOW IMPLEMENTATION IS DONE
IN THE FOLLOWING STEPS >

IMPLEMENTATION FLOW

3. Uses Bark to vocalize it → the generated text question is passed to Bark, which transforms it into expressive **audio tokens**, then decodes them into natural-sounding speech using EnCodec.

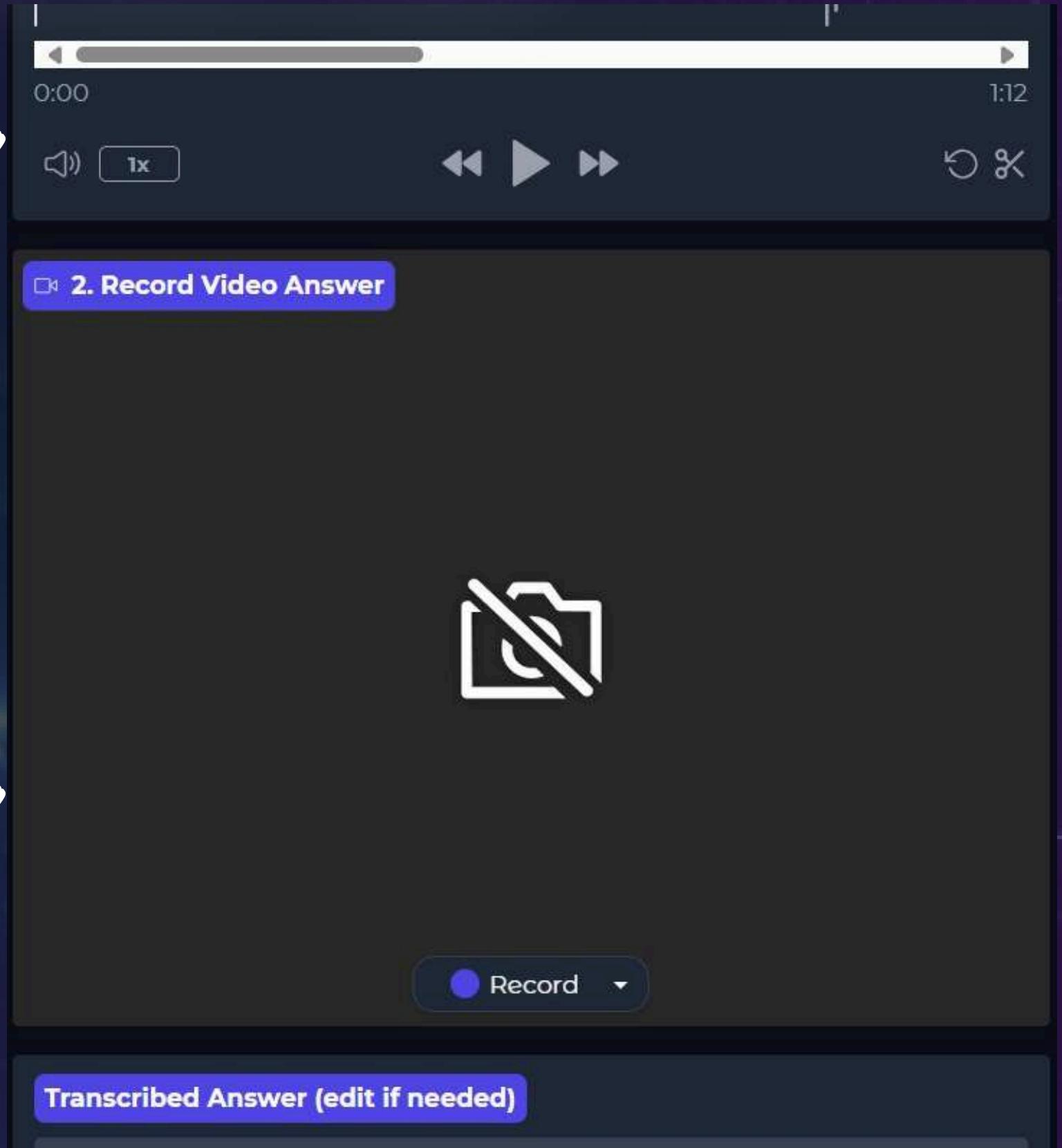
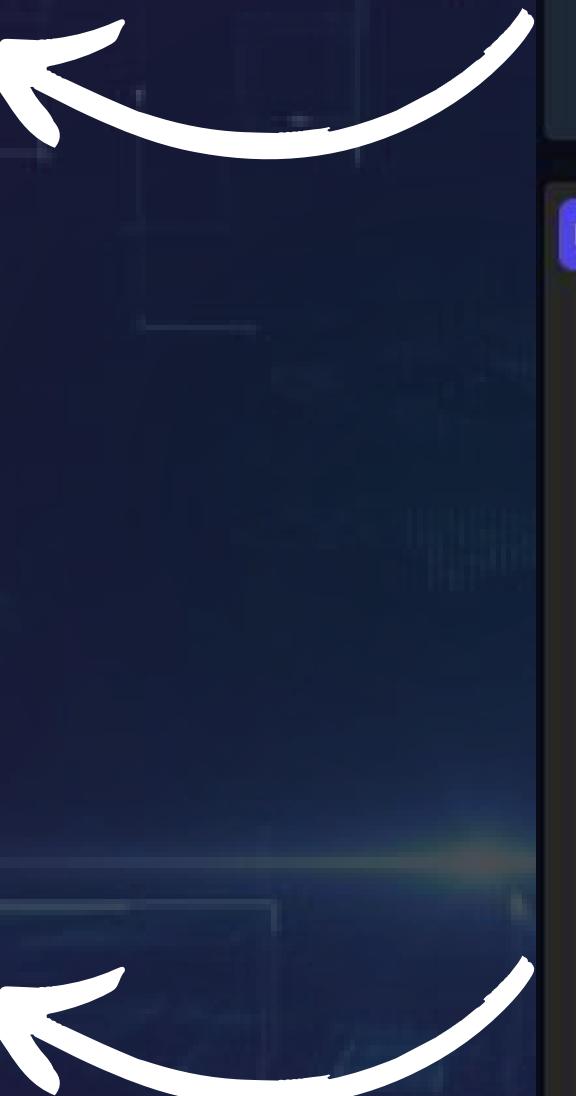
1. Retrieves contextual data from Qdrant
2. Builds a prompt for LLaMA to generate a tailored question



IMPLEMENTATION FLOW

4. Waits for verbal user response
→ transcribed using Whisper
→ audio is captured via microphone, resampled to 16kHz using [Torchaudio](#), and processed by the HuggingFace.

5. Records Video Answer →
The user's webcam is accessed via [OpenCV](#), capturing frames in real-time. Every 15th frame is passed to [DeepFace](#), which analyzes the [facial expression](#) and extracts the [dominant emotion](#).



5. IMPLEMENTATION

WHAT HAPPENS NEXT ?

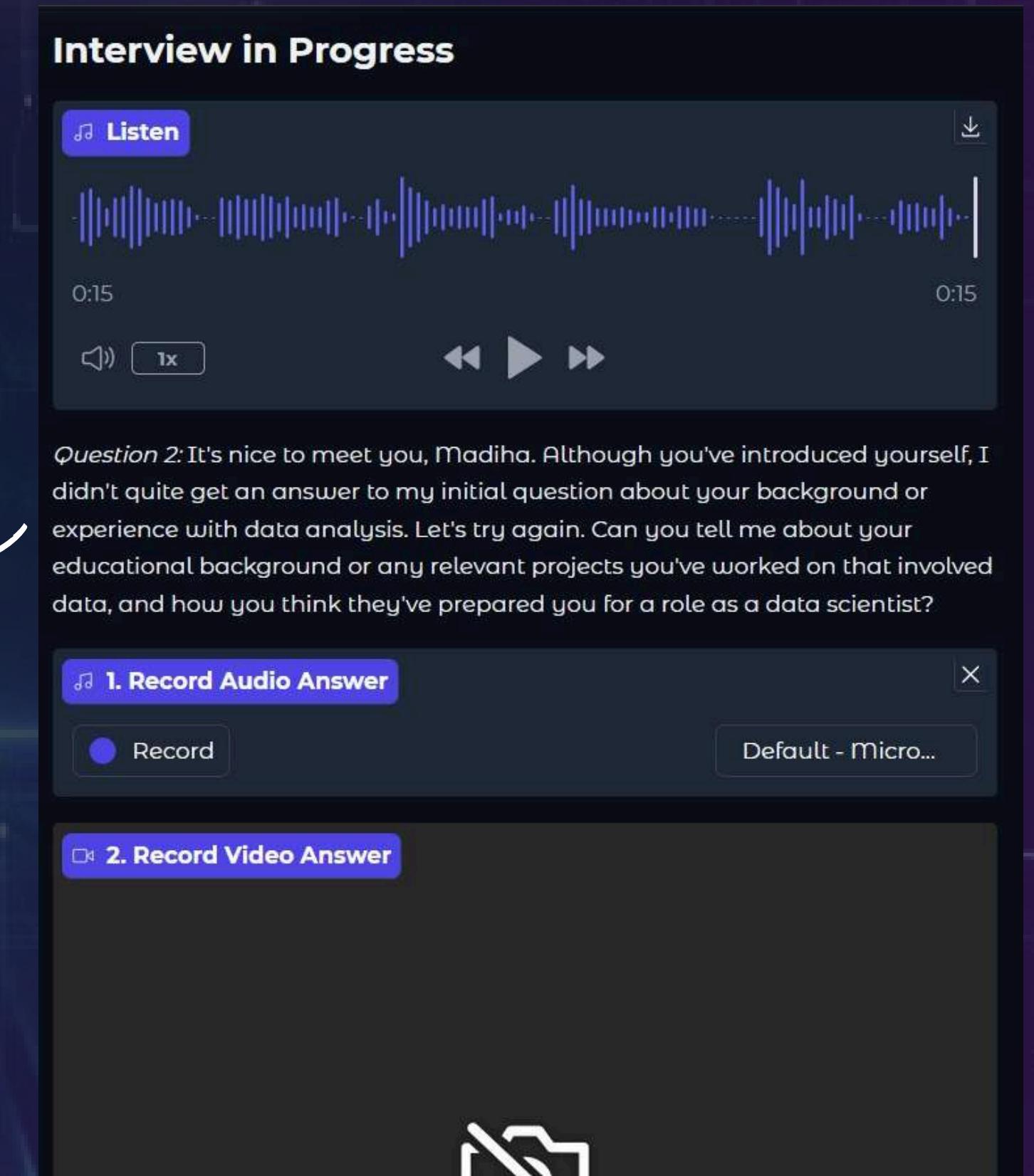
6. EVALUATE THE ANSWER

- The user's response (text from Whisper + emotion from video/audio) is passed to Mistral-7B-Instruct.
- It scores the answer on 5 dimensions: Relevance, Correctness, Clarity, Completeness, and Emotion.
- The system also calculates an effective confidence score using:

```
0.5 * answer_score + 0.22 * voice_emotion + 0.18 * facial_emotion + 0.1 * control_bonus
```

7. GENERATE FOLLOW-UP QUESTION (E.G., Q2)

- Based on the score, It then builds a fresh prompt (with feedback + context) and asks a new tailored question using LLaMA 3.3.
- This question is again converted to speech using Bark, and the loop continues.



8. SMART INTERVIEW SUMMARY - FULLY AUTO-GENERATED

- Every Q&A pair is logged with the user's exact answer.

LLM EVALUATION:

- Each response is scored by Mistral-7B, including:
- A qualitative score (Excellent, Good, Medium, Poor)
- Reasoning behind the score
- Concrete improvement suggestions

EMOTION DATA IS INTEGRATED:

- facial & vocal tones are recorded per answer.

EFFECTIVE CONFIDENCE SCORE

- is calculated and logged.

Interview Summary

Q1: Let's begin the interview. As an entry-level data scientist, I'd like to understand your background and foundation in data analysis. Can you tell me about a project or experience where you had to work with a dataset to extract insights or solve a problem, and how you approached it?

- *Answer:* hey, my name is madiha
- *Q Eval:* {'Score': 'Good', 'Reasoning': 'This question is somewhat relevant as it asks for a practical example of data analysis, which aligns with the role of an entry-level data scientist. However, it could benefit from more specificity regarding the type of datasets, problems, or techniques involved.', 'Improvements': ['- Specify the types of datasets (structured/unstructured) and their sources', '- Provide examples of specific problems or business cases', '- Request details on the analytical methods used (e.g., statistical tests, machine learning models)']}
- *A Eval:* {'Score': 'Medium', 'Reasoning': 'The answer provides some context and mentions a project, but lacks detail in describing the approach taken, the specific insights extracted, or the methods used.', 'Improvements': ['Describe the specific steps taken in the analysis process', 'Provide more details about the insights extracted and their significance', 'Explain the specific methods used for data cleaning, exploration, and modeling']}
- *Face Emotion: no_face, *Voice Emotion: sad
- Effective Confidence: 0.429
- Time: 245.33s

Q2: It's nice to meet you, Madiha. Although you've introduced yourself, I didn't quite get an answer to my initial question about your background or experience with data analysis. Let's try again. Can you tell me about your educational background or any relevant projects you've worked on that involved data, and how you think they've prepared you for a role as a data scientist?

6. EVALUATION AND RESULTS

INTERNET

READY TO EVALUATE? LET'S BREAK IT DOWN

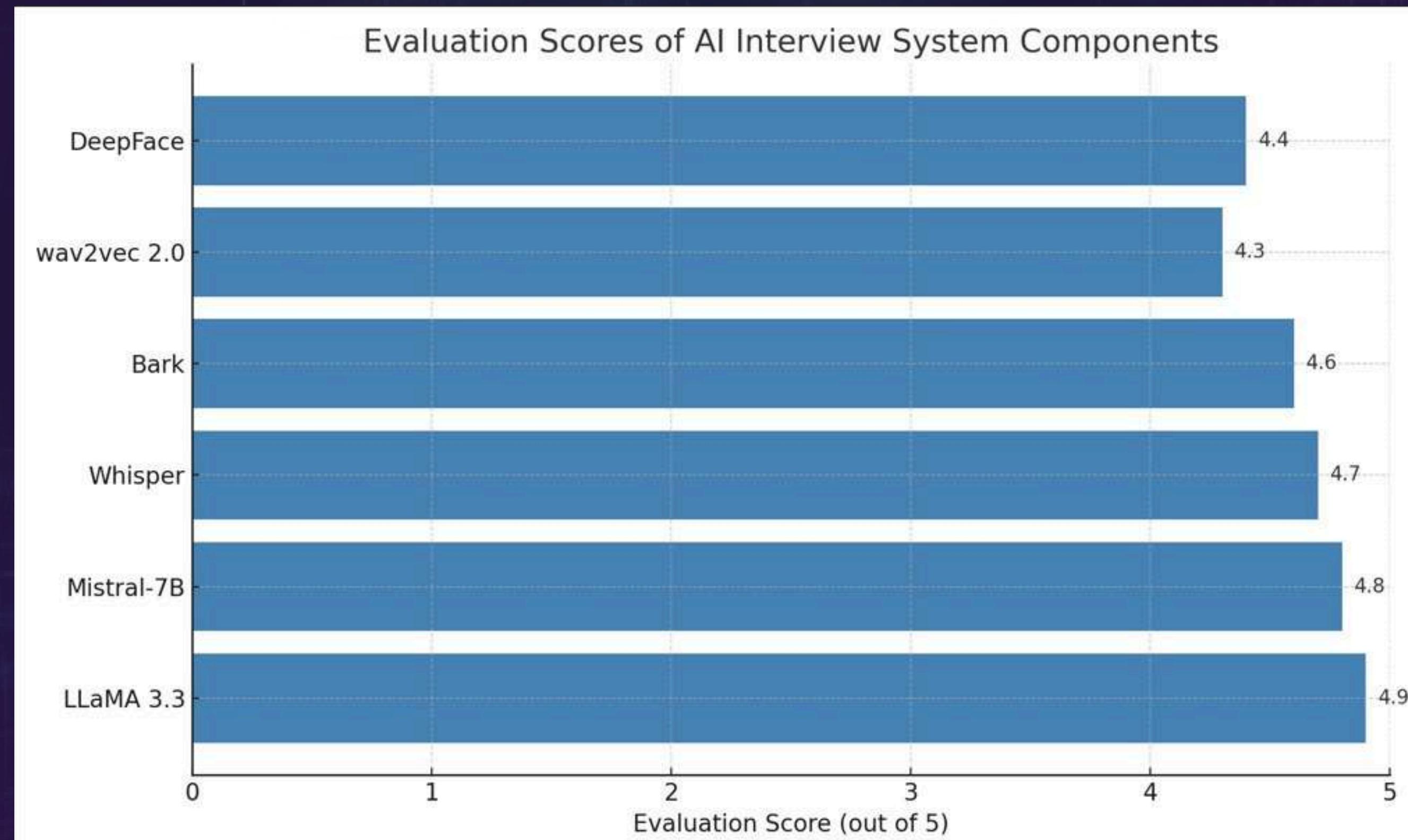
EACH MODEL WAS EVALUATED BASED ON ITS
CONTRIBUTION TO REALISM, ADAPTIVITY, AND HUMAN-
LIKENESS, NOT ONLY TECHNICAL ACCURACY.

6. EVALUATION AND RESULTS

HOW DO WE KNOW EACH MODEL PERFORMS WELL?

Model	Function	Evaluation Score	Key Insights
LLaMA 3.3	Question + Reference Answer Generation	4.9 / 5	Strong role-specific adaptation; high diversity & clarity
Mistral-7B	Answer Scoring + Feedback	4.8 / 5	Consistent scoring; slight generosity; improvement suggestions
Whisper	Speech-to-Text Transcription	4.7 / 5	Accurate with noisy input; editable for fairness
Bark	Text-to-Speech Synthesis	4.6 / 5	Natural-sounding, emotion-rich speech with prosody
wav2vec 2.0	Voice Emotion Detection	4.3 / 5	Robust across 7 emotions; strong generalization
DeepFace	Facial Emotion Detection	4.4 / 5	Reliable dominant emotion detection; consistent session trends

HOW DID THE SYSTEM ACTUALLY PERFORM(GRAPHICALLY) BASED ON GPT SCORES ?



HOW WE CALCULATED THE MODEL EVALUATION SCORES?

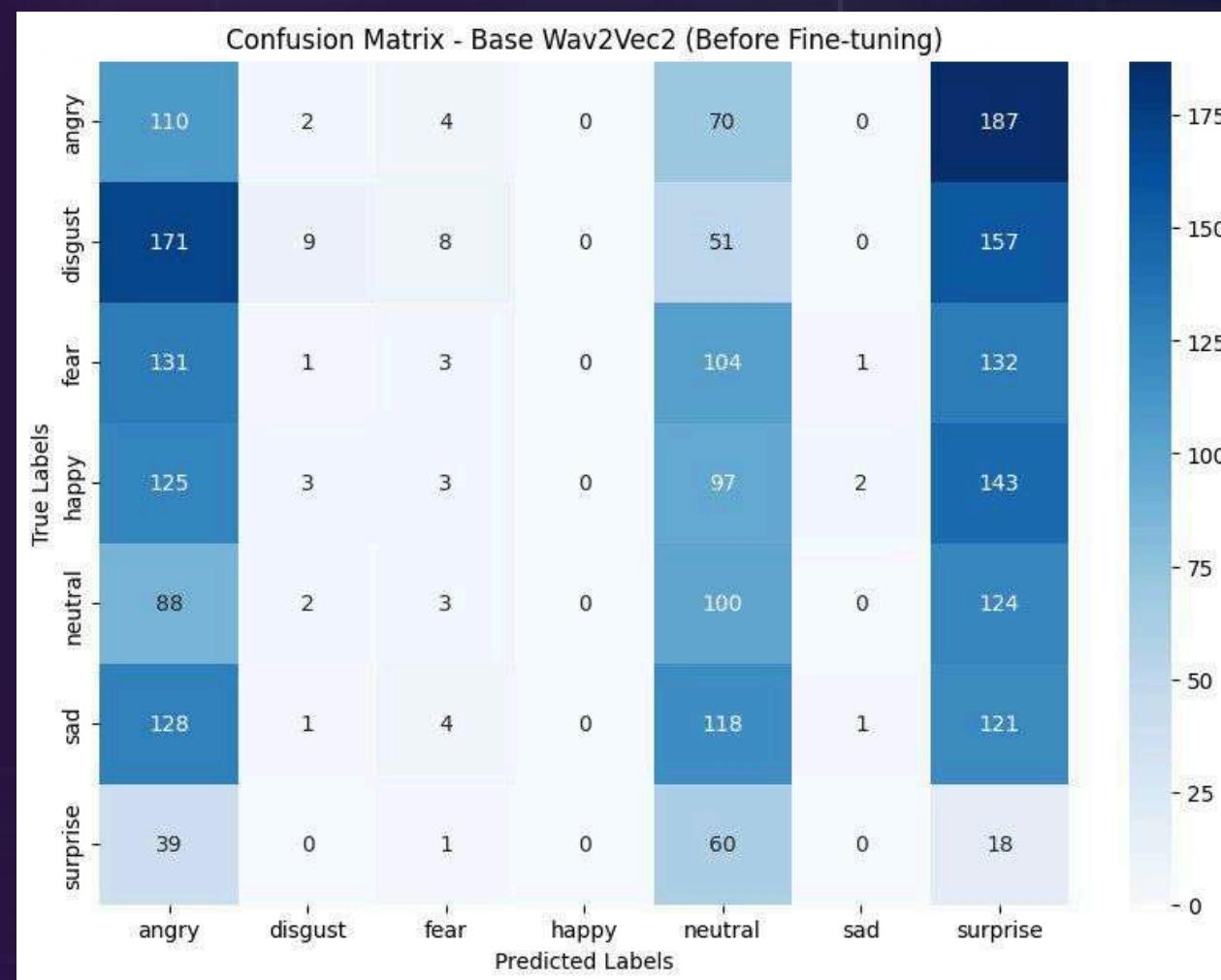
KEY EVALUATION RESULTS (5-POINT SCALE) USING A GPT-4 RUBRIC:

- **Internal Testing Performance:** For models like wav2vec 2.0, we measured classification accuracy on known emotion datasets (e.g., RAVDESS, EMO-DB), then normalized the results to a 5-point scale (e.g., 85% \approx 4.3/5). For Whisper, accuracy was verified by comparing transcriptions to ground truth on real and noisy inputs.
- **Qualitative Assessment:** For generation models (e.g., LLaMA 3.3, Mistral-7B, Bark), we conducted manual reviews: Relevance and diversity of LLaMA's questions, Clarity and fairness of Mistral's feedback, Expressiveness of Bark's audio output
- **Rubric-Based Scoring:** We used a 5-point rubric aligned with evaluation dimensions like: Accuracy, Naturalness, Role-specificity, Response consistency, Emotional sensitivity (for Bark, DeepFace, wav2vec)
- **Reference Model Comparison (GPT-4):** In some cases, we compared our outputs to evaluations generated by a higher-capability LLM (GPT-4) to check alignment with expert-like scoring.

6. EVALUATION & RESULTS

BASE WAV2VEC

BEFORE FINETUNING



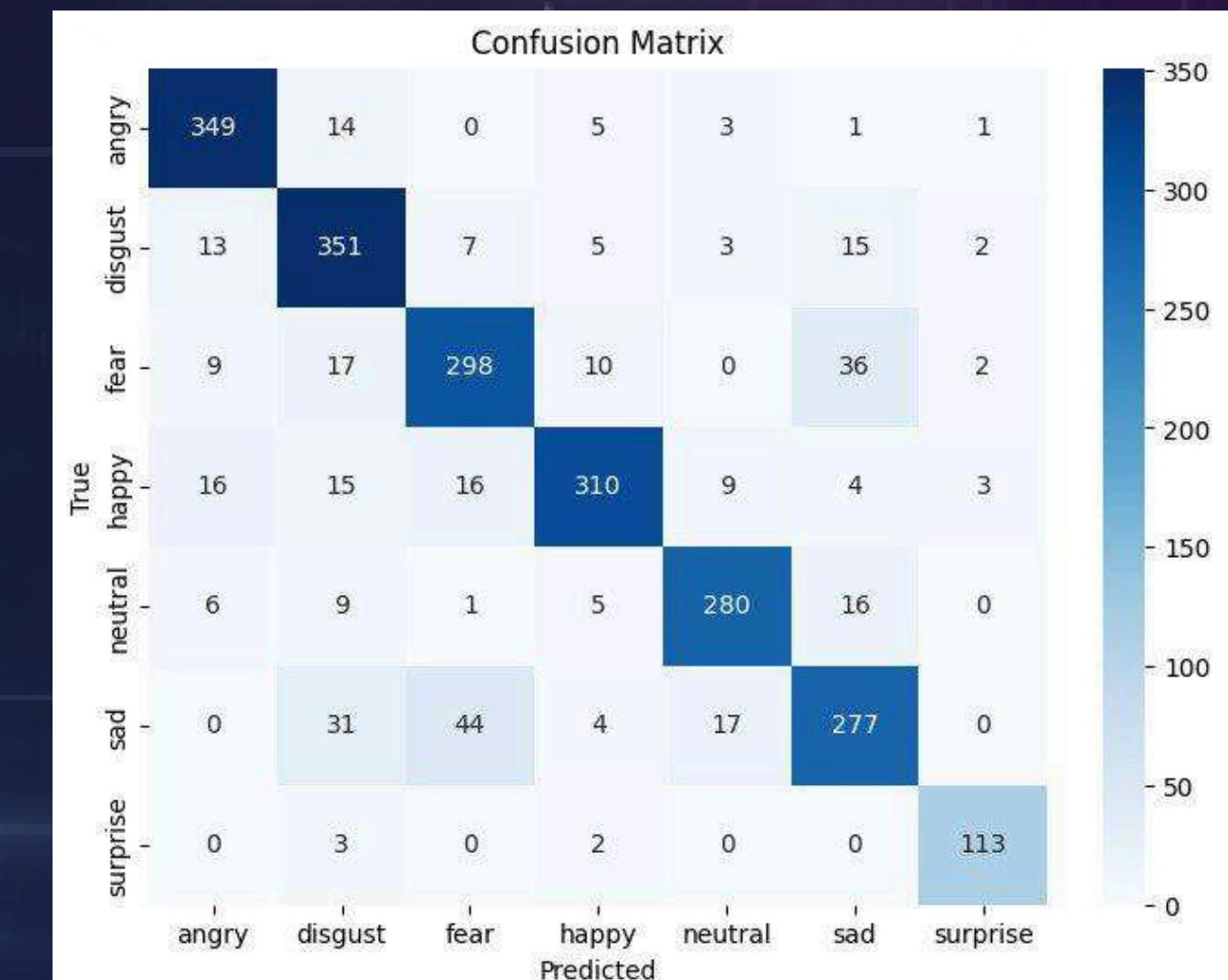
accuracy

0.17

0.10

0.07

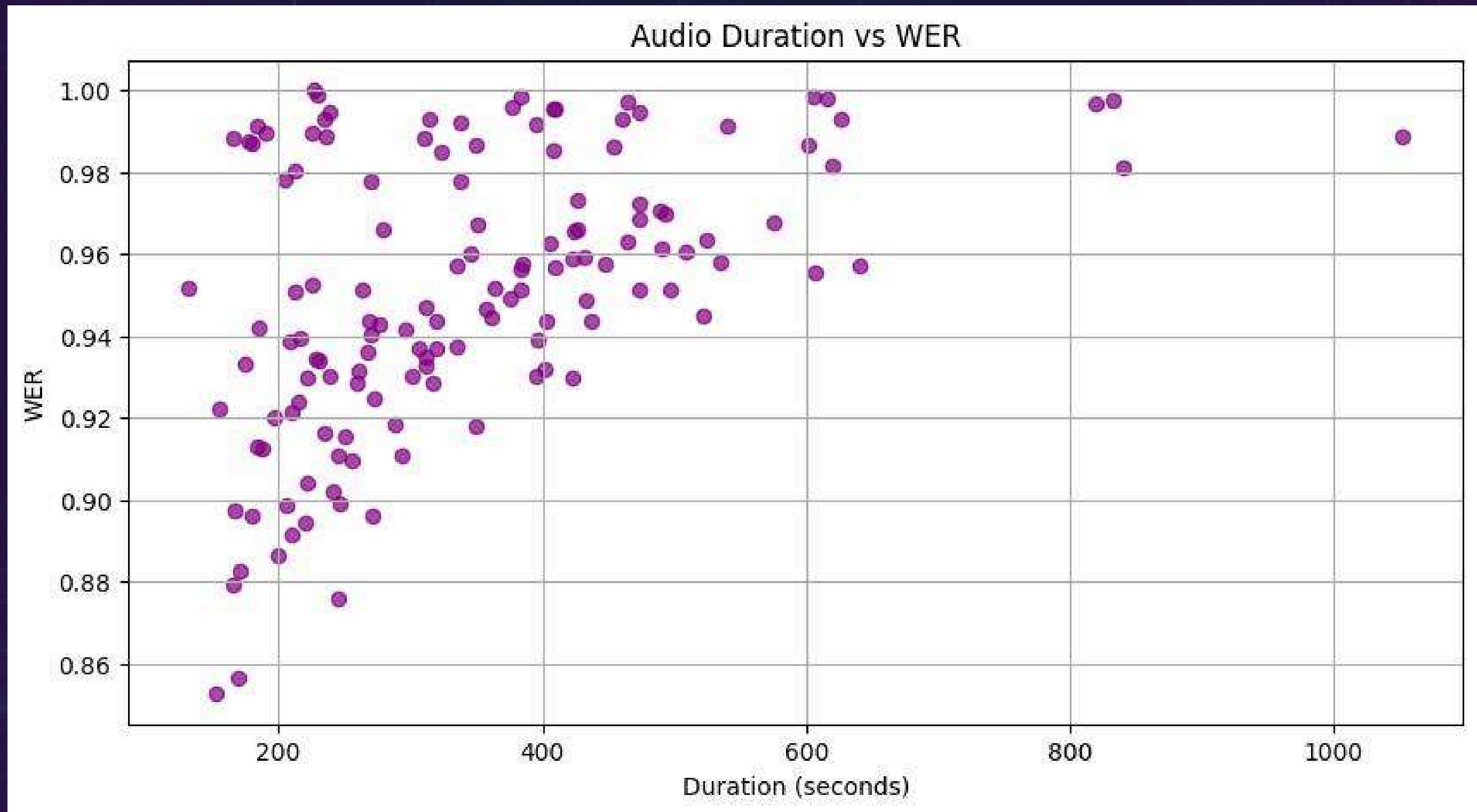
AFTER FINETUNING



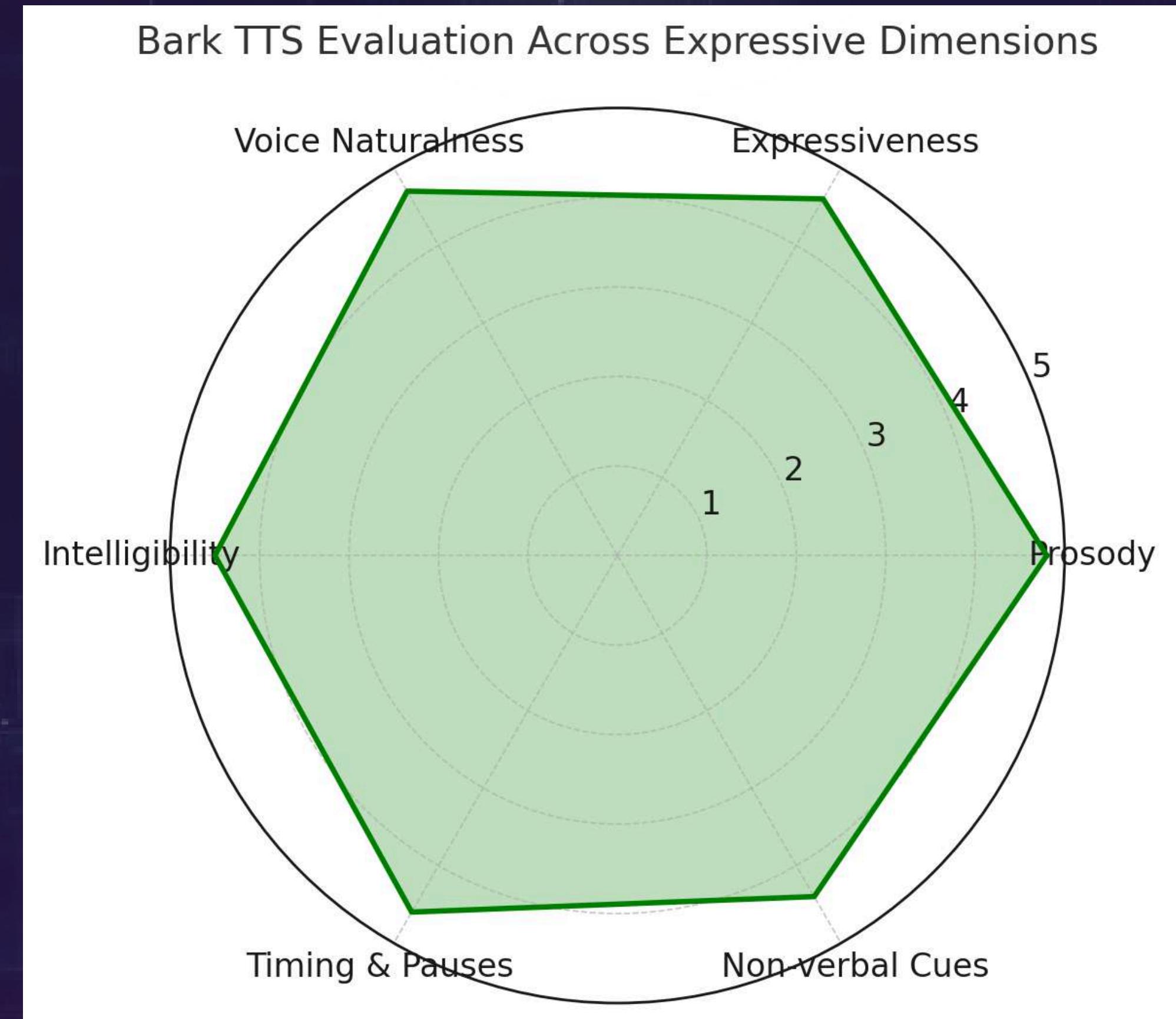
accuracy

0.85

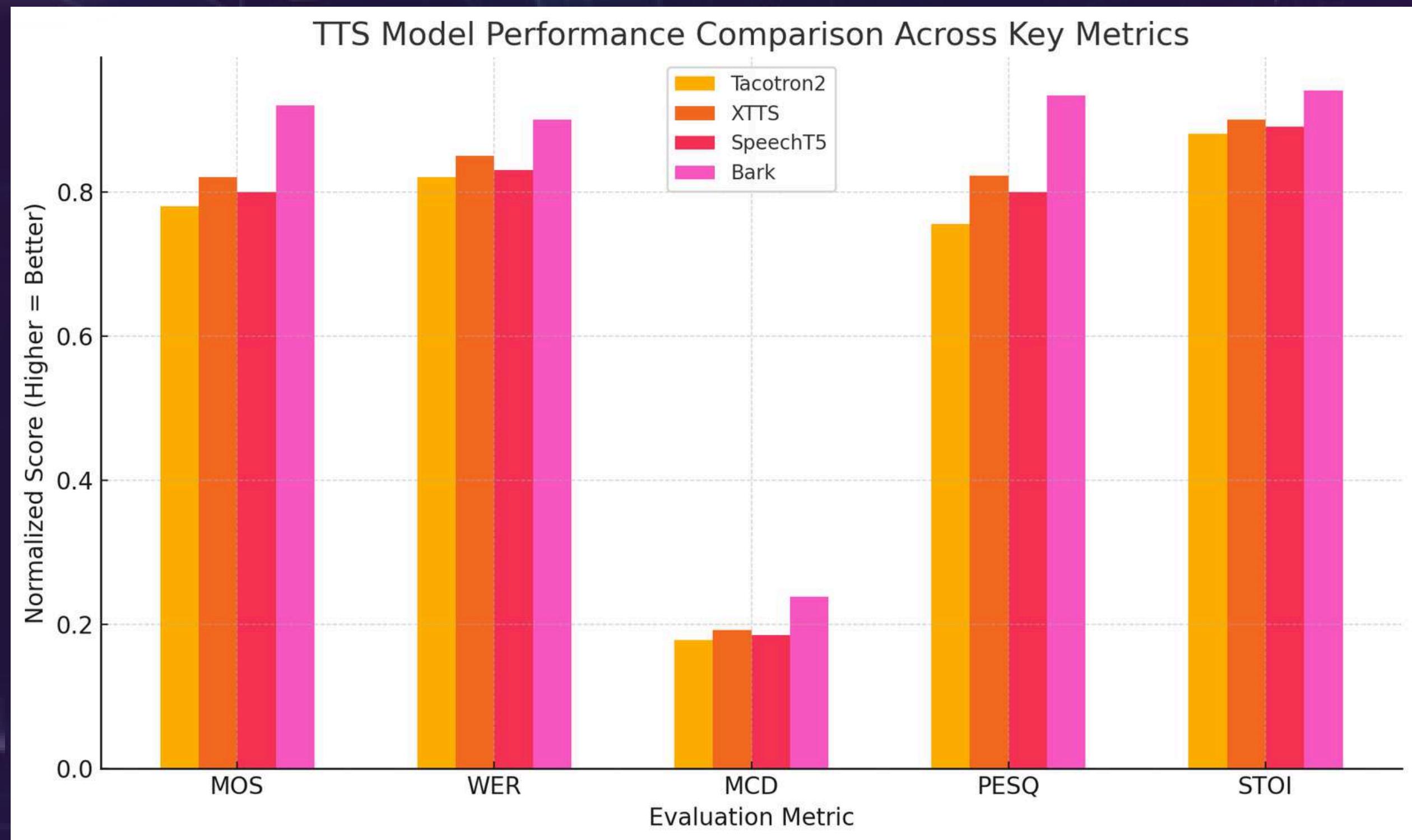
WHISPER



BARK



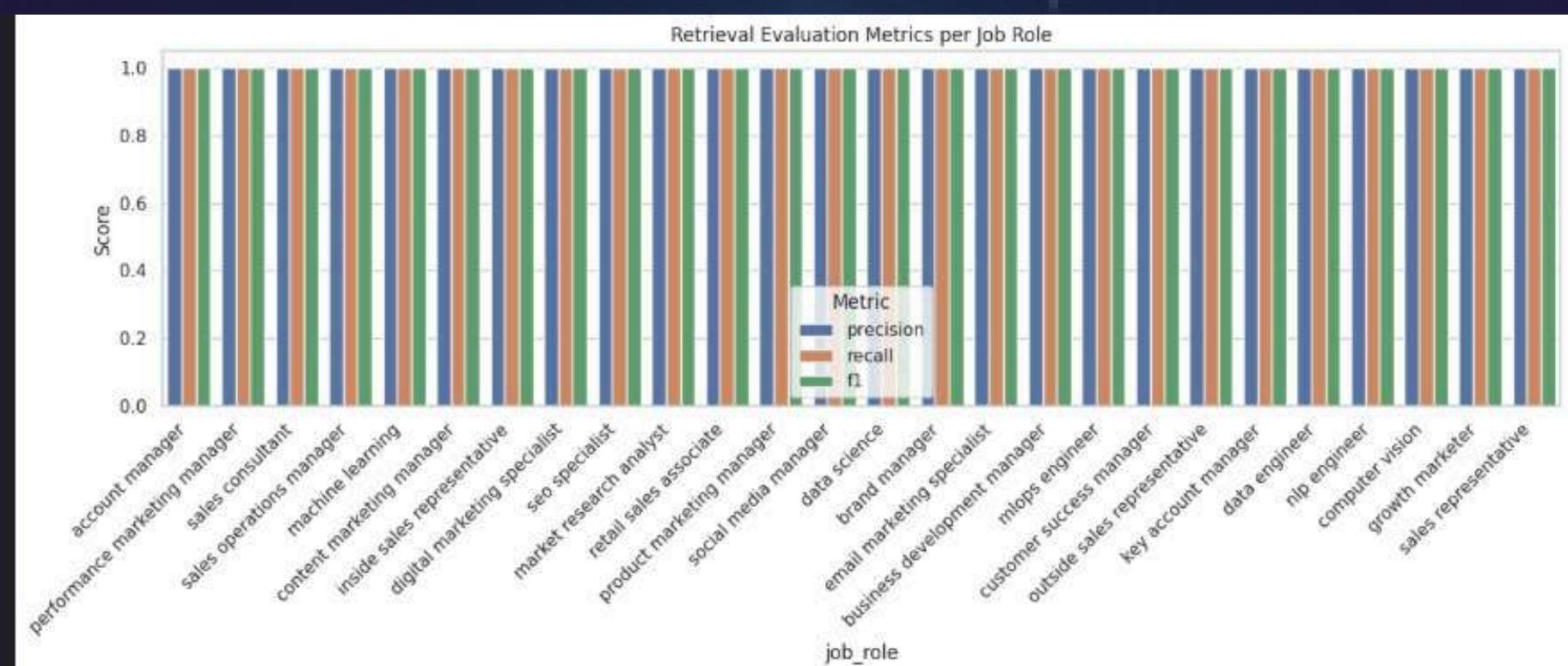
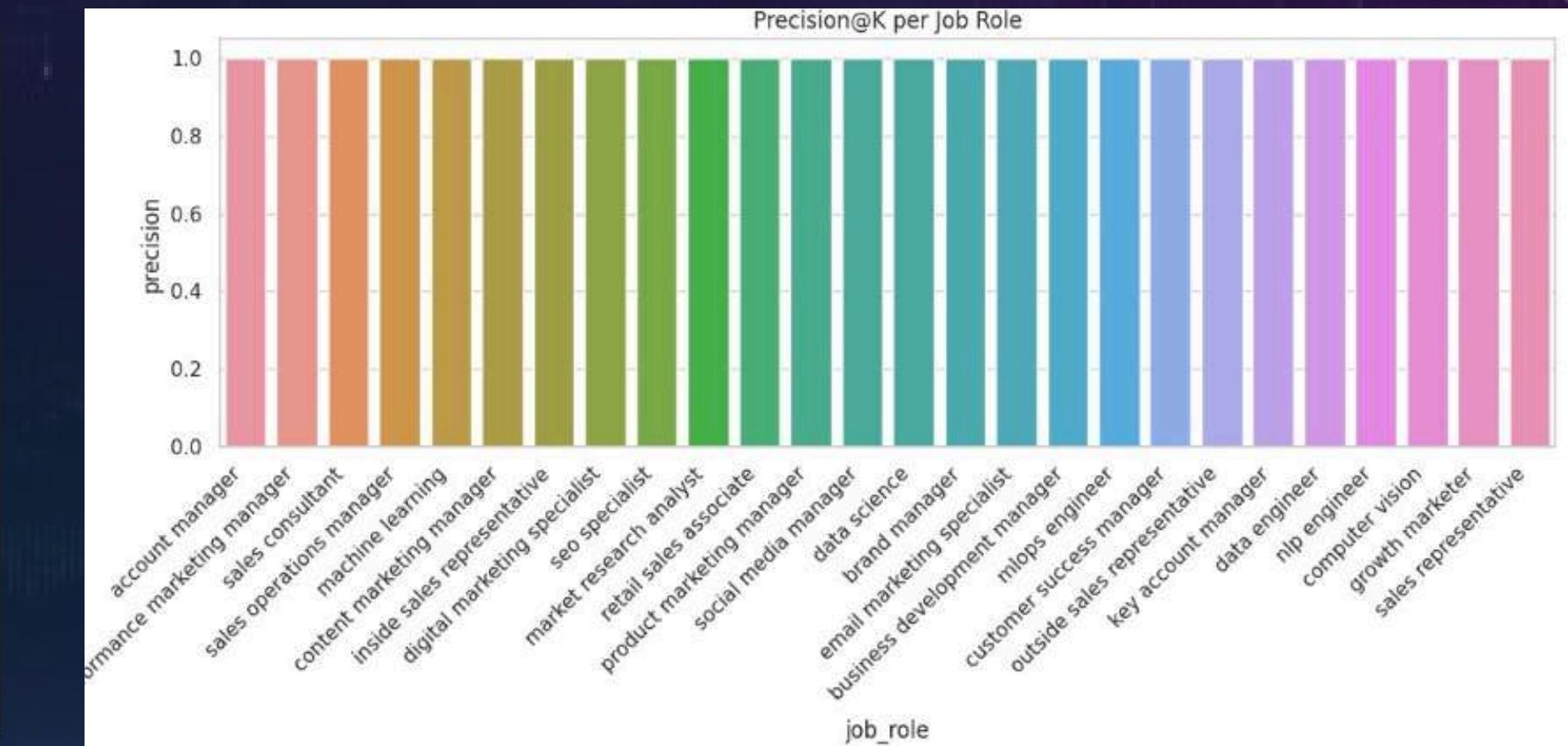
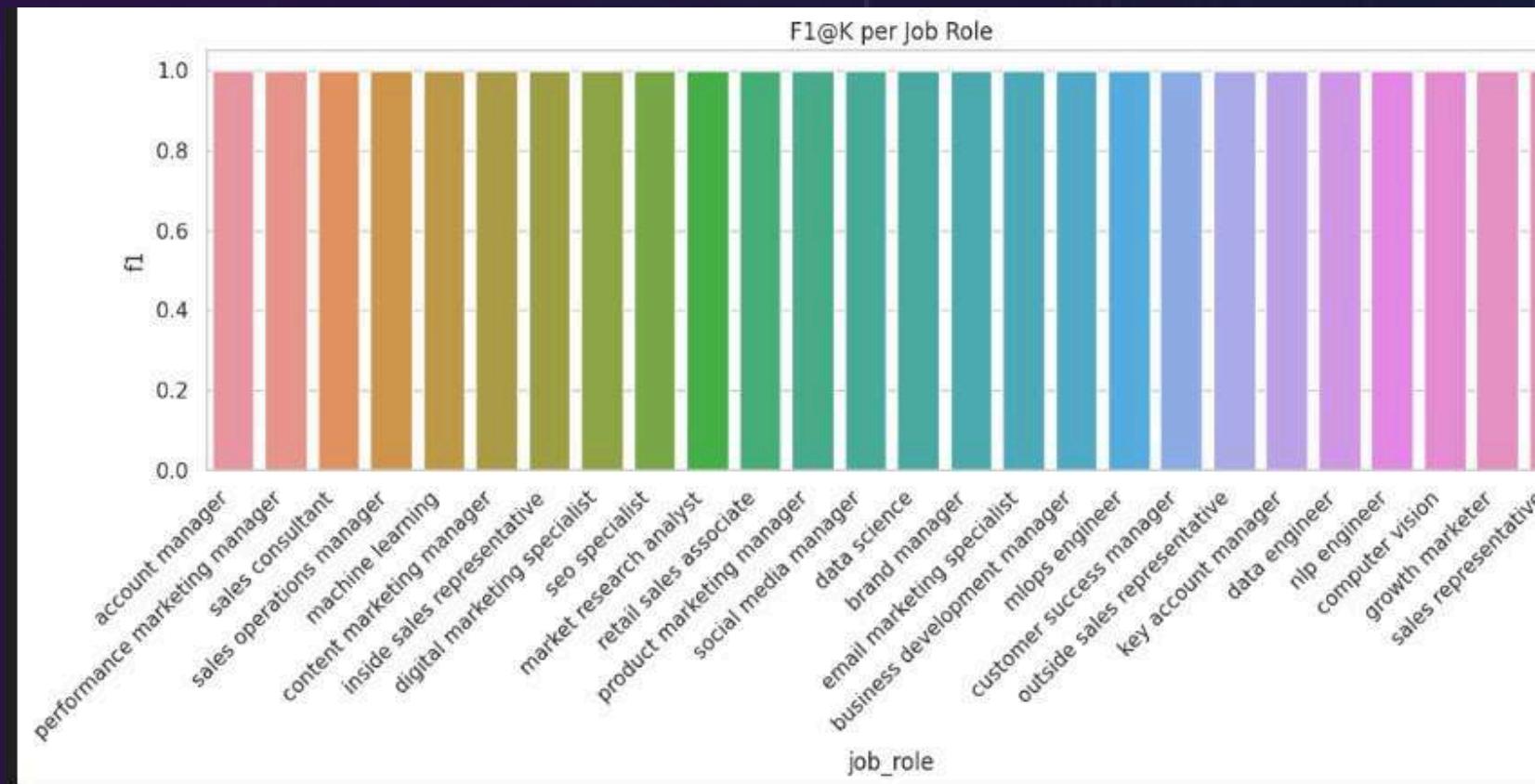
6. EVALUATION & RESULTS



- MOS (Mean Opinion Score): Human-rated naturalness (1-5)
- WER (Word Error Rate): Lower is better (inverted here for visualization)
- MCD (Mel-Cepstral Distortion): Lower is better (inverted here)
- PESQ (Perceptual Evaluation of Speech Quality): Higher = better
- STOI (Short-Term Objective Intelligibility): Clarity (0-1)

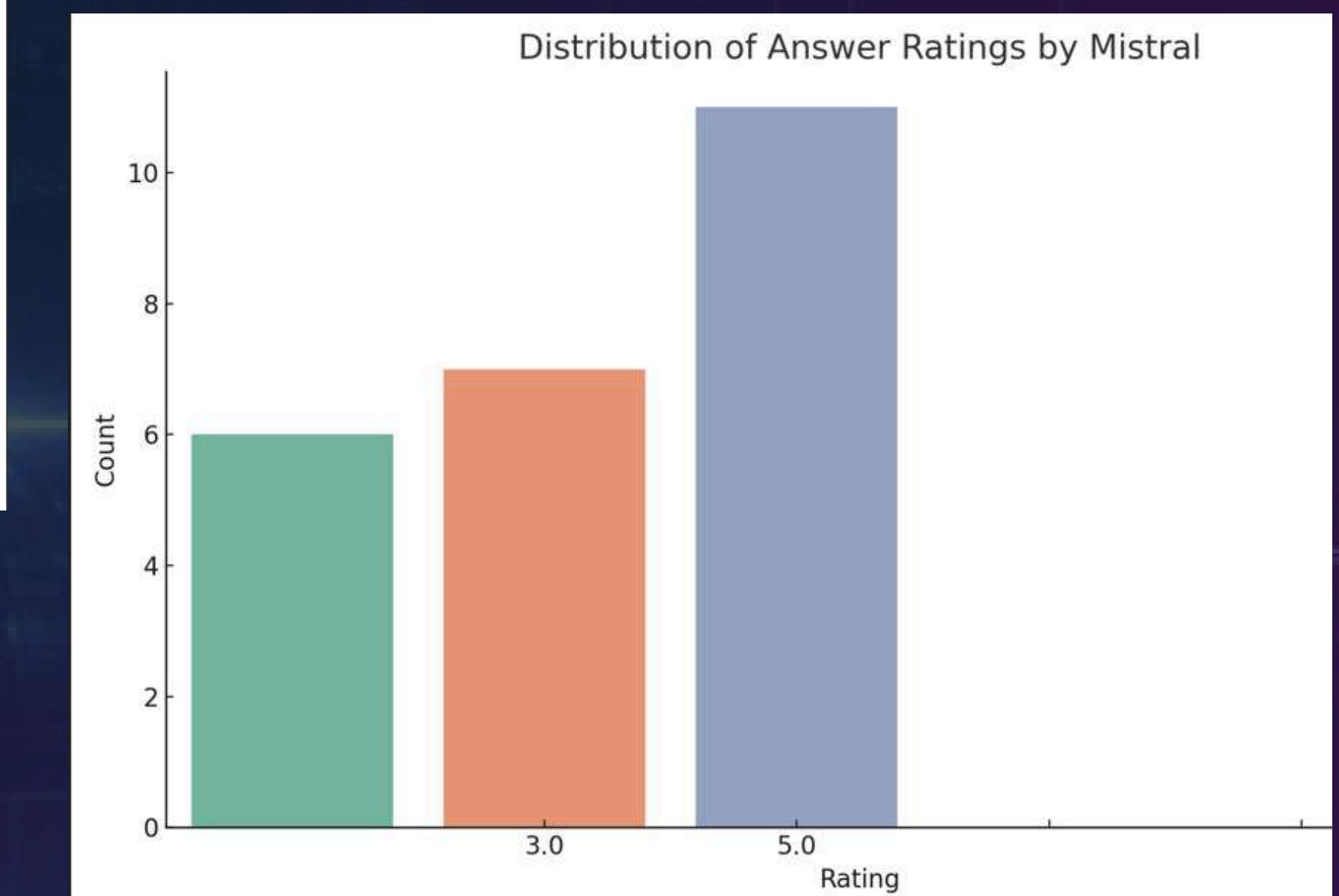
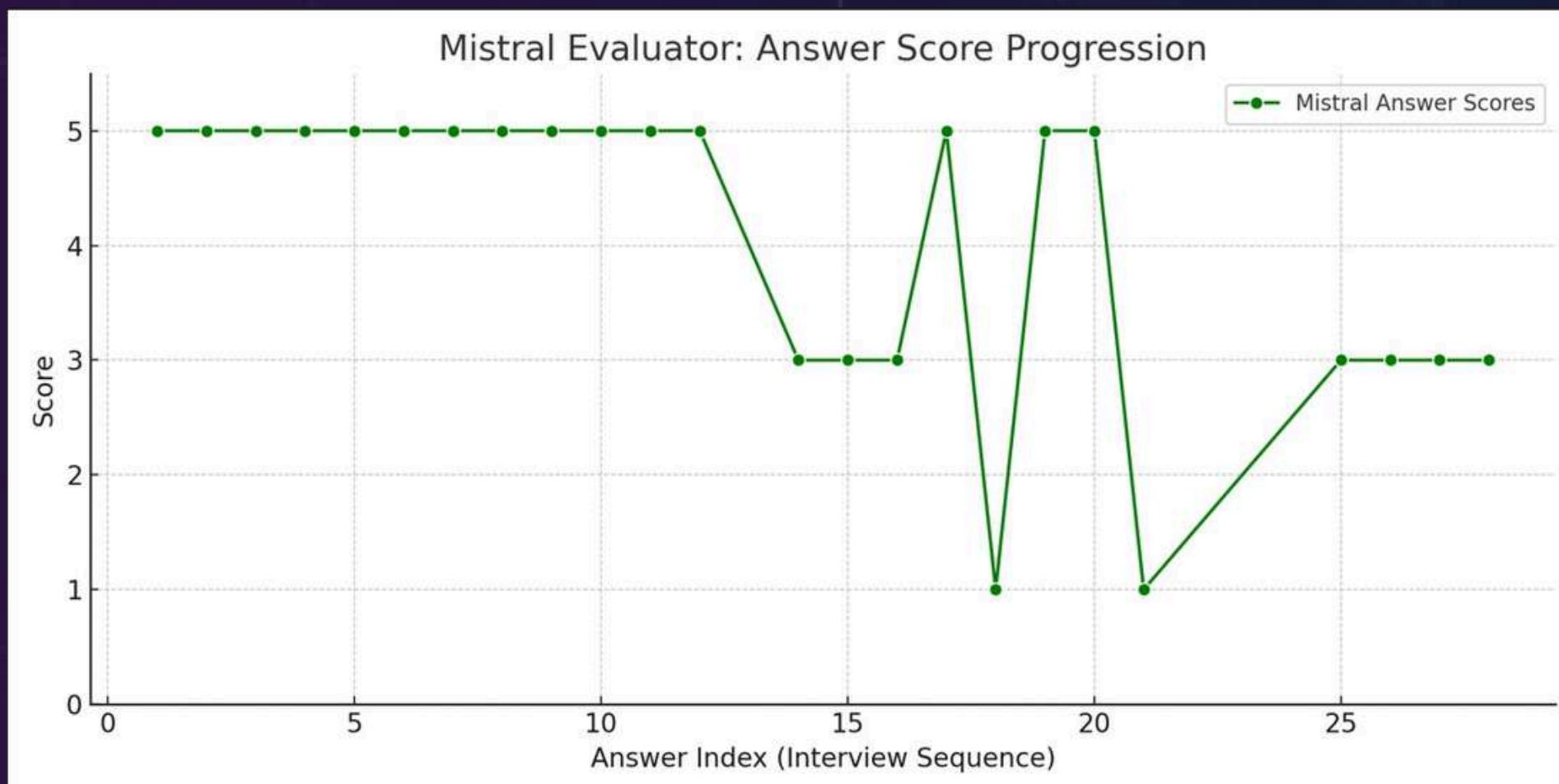
6. EVALUATION & RESULTS

HOW EFFECTIVE IS OUR RETRIEVAL ENGINE?



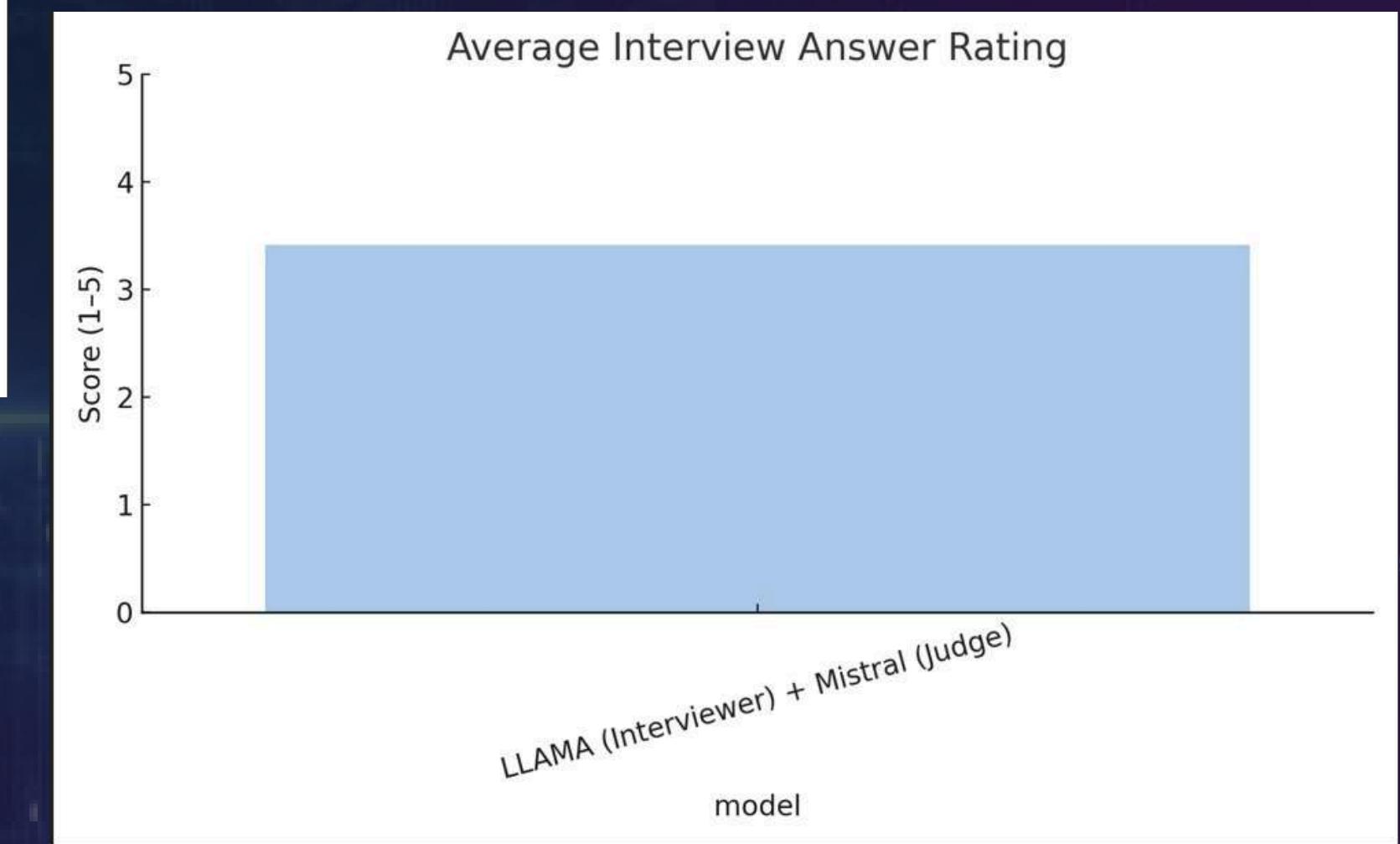
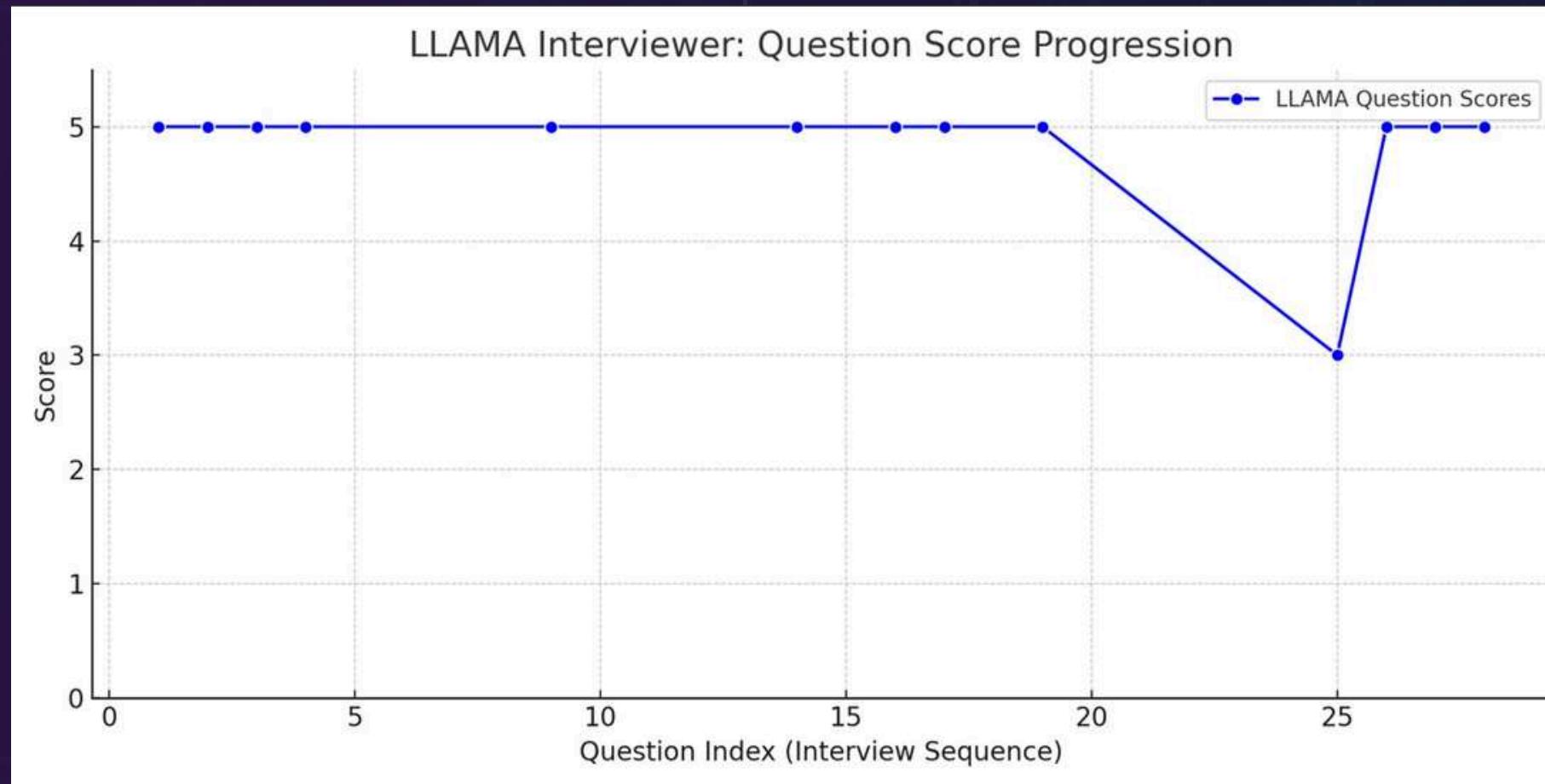
6. EVALUATION & RESULTS

WHAT ABOUT MISTRAL EVALUATOR GRAPH ?



6. EVALUATION & RESULTS

LLAMA 3.3



IMPORTANT NOTE FOR INTERPRETATION

THE HIGH EVALUATION SCORES FOR MODELS LIKE LLAMA 3.3, MISTRAL-7B, AND THE RETRIEVAL ENGINE ARE BASED ON STORED INTERVIEW SESSIONS CONDUCTED DURING THE IMPLEMENTATION PHASE.

THESE SESSIONS WERE USED TO SIMULATE REAL USER INPUT AND VALIDATE PERFORMANCE UNDER CONTROLLED TEST CONDITIONS - AS EXPLAINED IN THE SYSTEM IMPLEMENTATION.

6. EVALUATION AND RESULTS

HOW DID THE SYSTEM ACTUALLY PERFORM? KEY EVALUATION RESULTS (5-POINT SCALE) USING A GPT-4 RUBRIC:

Category	Score	Insight
Question Relevance	4.9	Strong alignment with job roles
Contextual Responsiveness	4.7	Adaptive follow-ups, slight transition rigidity
Scoring Consistency	4.8	Fair and well-reasoned, though slightly generous
Language Coherence	4.6	Formal and clear, lacks conversational warmth
Behavioral Realism	4.4	Professional tone, needs emotional flexibility

7. CONCLUSION & FUTURE WORK

INTERNET

CONCLUSION

- **AI-Powered Evaluation:** The system addresses bias, inconsistency, and inefficiency in traditional interviews using intelligent automation.
- **Multimodal Assessment:** It integrates speech, facial emotion, and answer content to evaluate both technical and behavioral aspects of candidates.
- **Adaptive Questioning:** LLaMA 3.3 with RAG ensures each question is role-specific, context-aware, and progressively challenging based on user performance.
- **Explainable & Transparent:** Scoring and feedback are backed by explainable AI, offering clarity to both candidates and recruiters.
- **Dual-Purpose Impact:** Serves as a tool for recruiters (screening aid) and for candidates (training simulator with personalized feedback).

FUTURE WORK

1. **Production Deployment:** Transition the prototype into a fast, secure, real-time web application for wide adoption.
2. **Custom Candidate Filtering:** Allow companies to define evaluation rules and weights for tailored hiring recommendations.
3. **Multilingual & Cultural Support:** Expand to multiple languages while ensuring fair, culturally aware emotion analysis and scoring.
4. **Continuous Learning (RLHF):** Use anonymized interview data to refine models over time via reinforcement learning from feedback.
5. **Dataset & Role Expansion:** Grow the dataset to support more job roles, industries, and emerging fields with limited data.

Our Team



MOHAMMAD
TAREK



MUHAMMAD
YASSER



FARIDA KHALED



HANIA RUBY



ESRAA MOHAMED



MADIHA SAIED



THANKS FOR YOUR TIME
