




AceInterview

Eric Jung, Sean Wei, Parker Brailow, Naveen Sukumar



Introduction



Team



Eric Jung
erijung@berkeley.edu
5th Year MIDS



Sean Wei
seanwei2001@berkeley.edu
5th Year MIDS



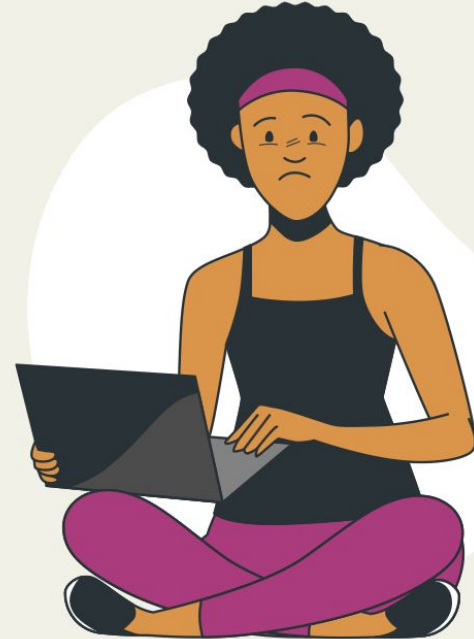
Parker Brailow
brailowparker@berkeley.edu
5th Year MIDS



Naveen Sukumar
naveensukumar@berkeley.edu
5th Year MIDS

Customer Problem

- **Target User:** New college graduates or those in their early career.
- *"How can I effectively practice the behavioral elements of virtual interviews?"*
- *"How can I quickly and easily get feedback on behavioral aspects of my interview performance?"*



Mission

- Provide **informational & behavioral** feedback
 - Content/structure of answers
 - Vocal presentation
 - Non-verbal expressions
- Help candidates secure their dream jobs and set them up for long-term success.





Demo





Technical Approach






Dataset

- MIT Interview Dataset by Dr. Ehsan Hoque.
- 138 video interviews scored by human raters on performance indicators:
 - RecommendHiring
 - Engaging
 - Focused
 - Etc.
- Obstacles:
 - Not representative of typical interview angle
 - Camera angles and zoom change.

Pictures removed for subject privacy



Baseline Feedback Model



GPT-4o mini

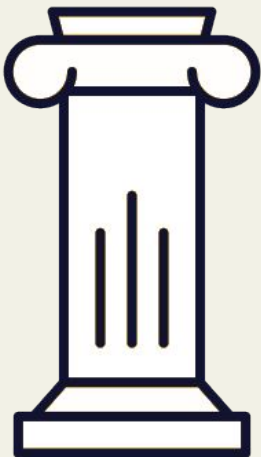
- Most widely known AI product
- 128k token context window
- Process up to 250 frames
- No Audio
- Ranked #10 Video-MME benchmark



3 Pillars of Impression Formation

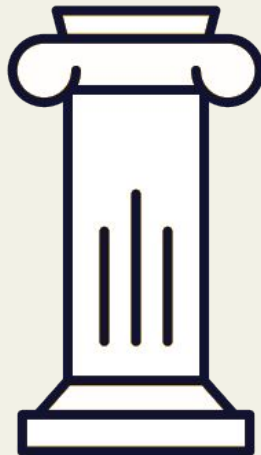
Visual Features

- Where am I looking?
- How is my body positioned?



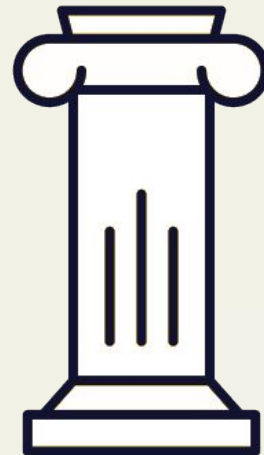
Verbal Features

- How fast am I talking?
- What am I saying?

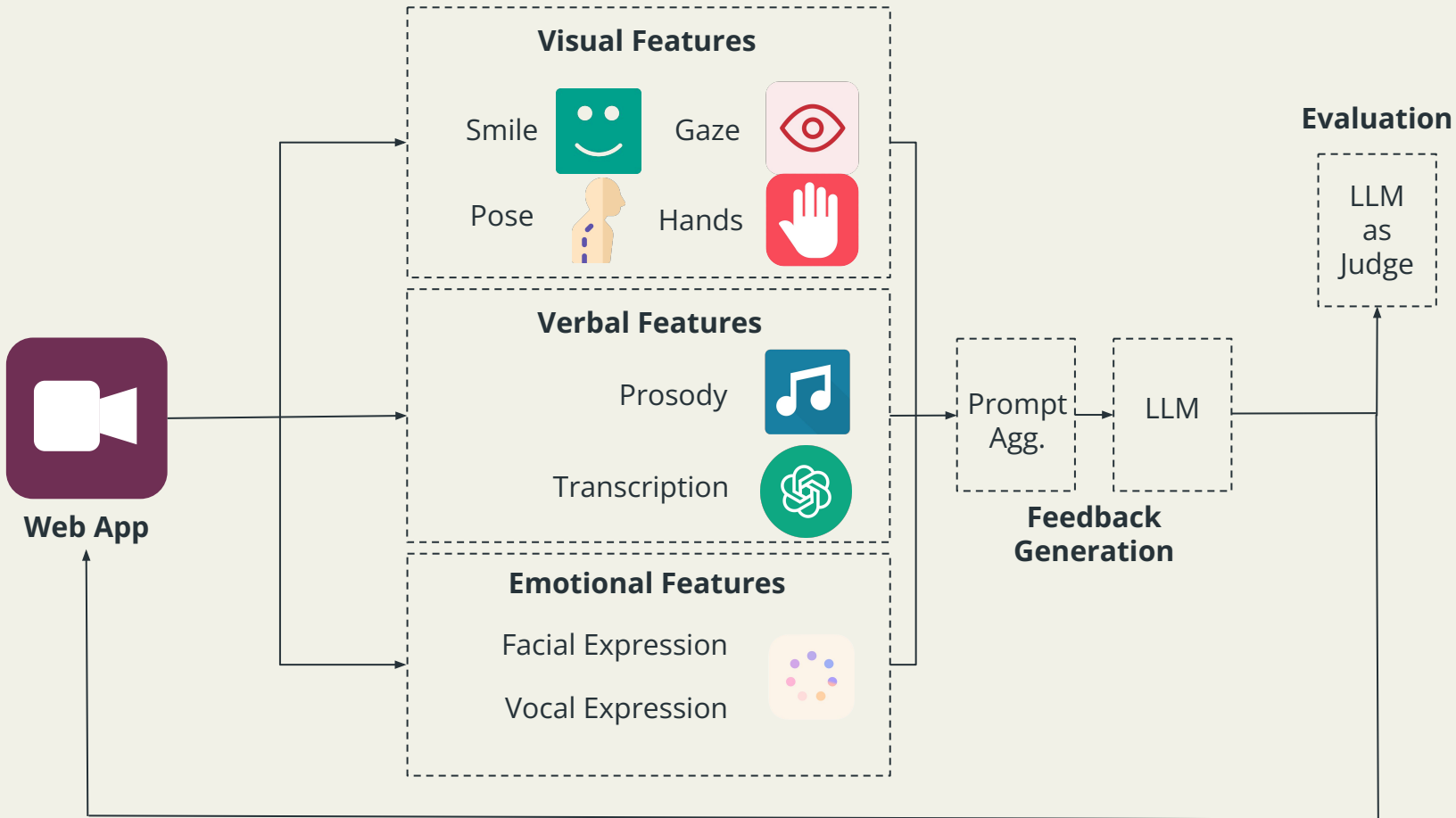


Emotional Features

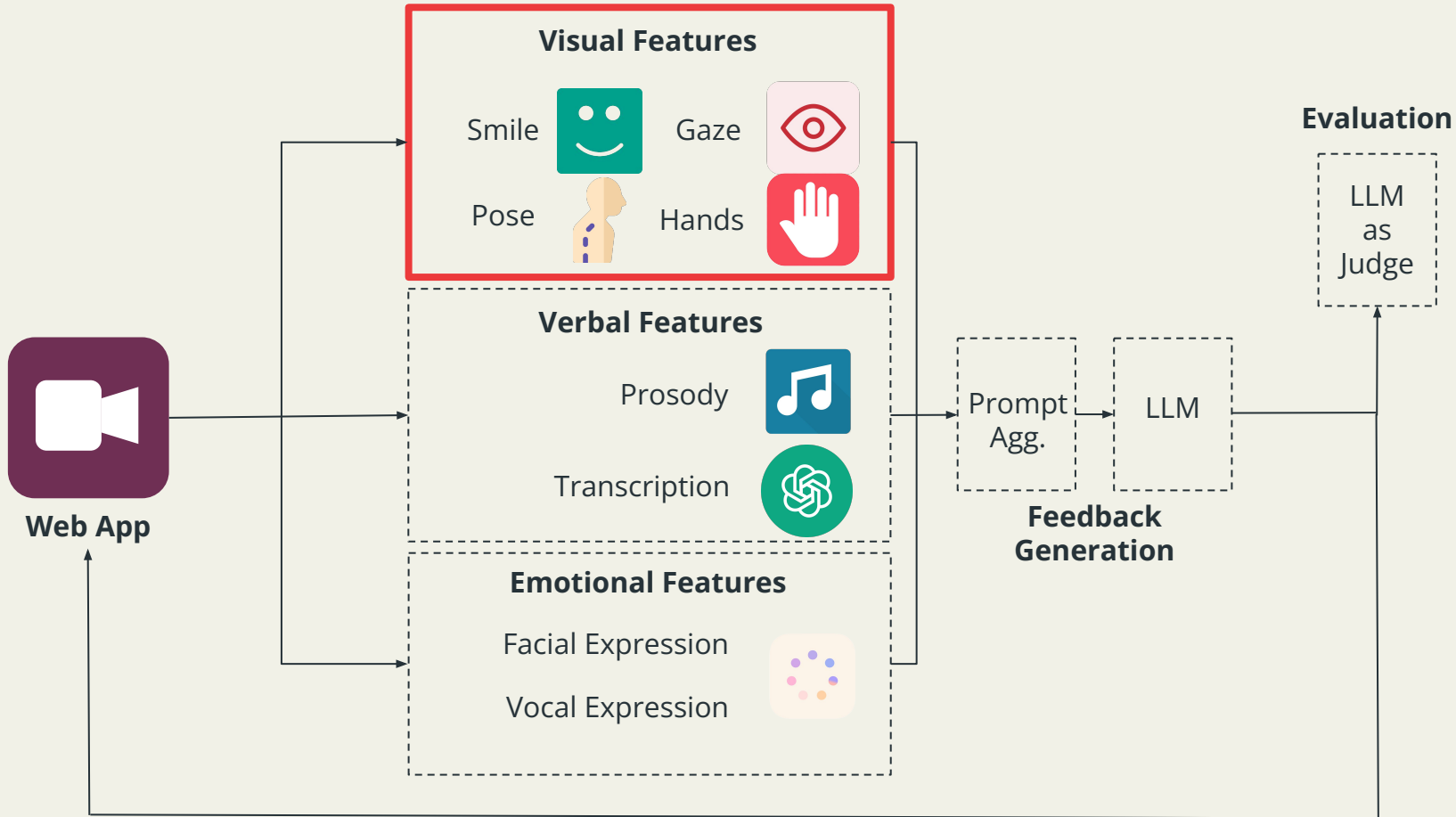
- Am I showing happiness on my face?
- Do I sound confident?



Advanced Modeling Pipeline



Visual Features



Visual Features: Pose and Hands

MediaPipe Models

- Detect landmarks like shoulders and eyes.
- Calculate body positioning to identify:
 - **Slouching**
 - **Head tilt**
 - **Face touching**

Evaluation

- Unit Testing:
 - Trying different angles, poses, etc.
 - Camera angles can impact features.



Visual Features: Smile Detection

- Previously Tried:
 - Pre-trained NN, Detectron2
- Using Haar Cascades
 - Accuracy: >90%
- EDA & Lit Review
 - Correlation: $r = 0.49$
 - People who smiled in interviews were more likely to receive offers.

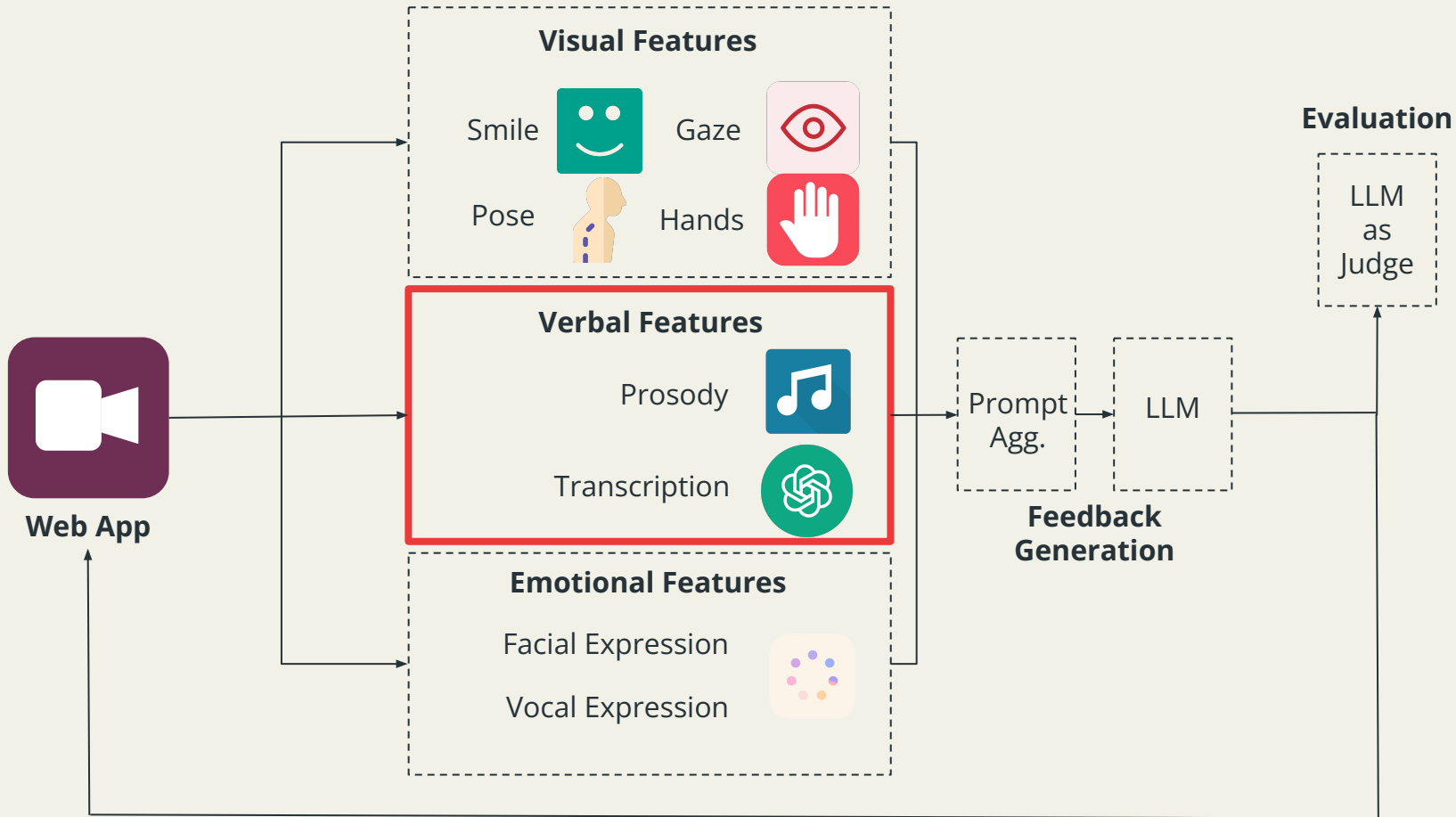
*Pictures removed for subject
privacy*

Visual Features: Gaze Detection

- Utilizing pre-trained CNN
 - (>96% accuracy)
- EDA & Lit Review
 - Our dataset correlation: $r = .59$
 - Study: Virtual interview away-from-camera gaze negatively associated with evals.

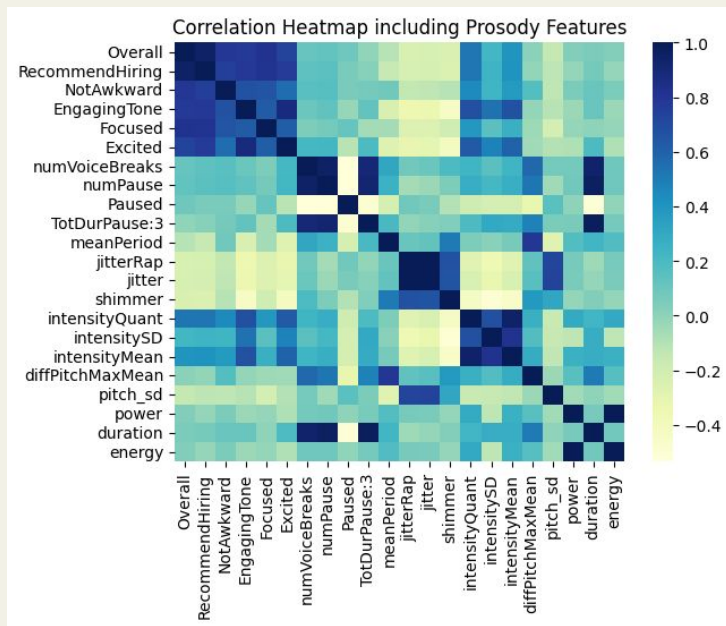


Verbal Features



Verbal Features: Prosody

- Utilizing parselmouth library



- EDA (Correlation):

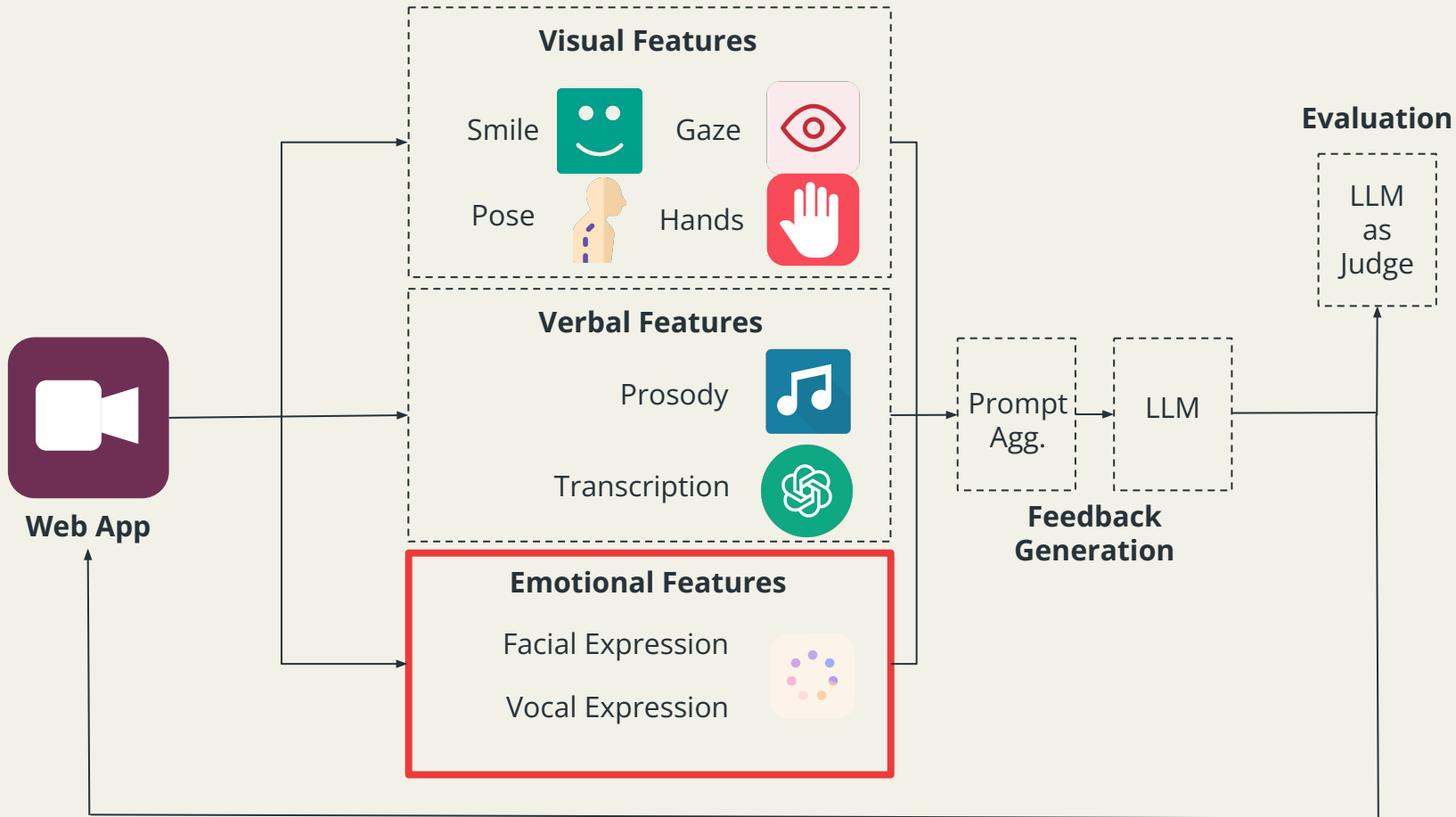
Feature 1	Feature 2	Correlation (r)
EngagingTone	RecommendHiring	0.78
Jitter	EngagingTone	-0.33
Jitter	RecommendHiring	-0.22
Shimmer	EngagingTone	-0.44
Shimmer	RecommendHiring	-0.25
IntensityQuant	EngagingTone	0.68
IntensityQuant	RecommendHiring	0.52

Verbal Features: Transcription

- Eloquent and articulate delivery are important to presentation skills.
- CrisperWhisper transcription model:
 - Filler words
 - Pauses
 - Stutters
 - False starts
 - Ex: *"Its nice to meet you. [UM] I am doing great!"*
- Model struggles when audio quality is low or muffled.



Emotional Features

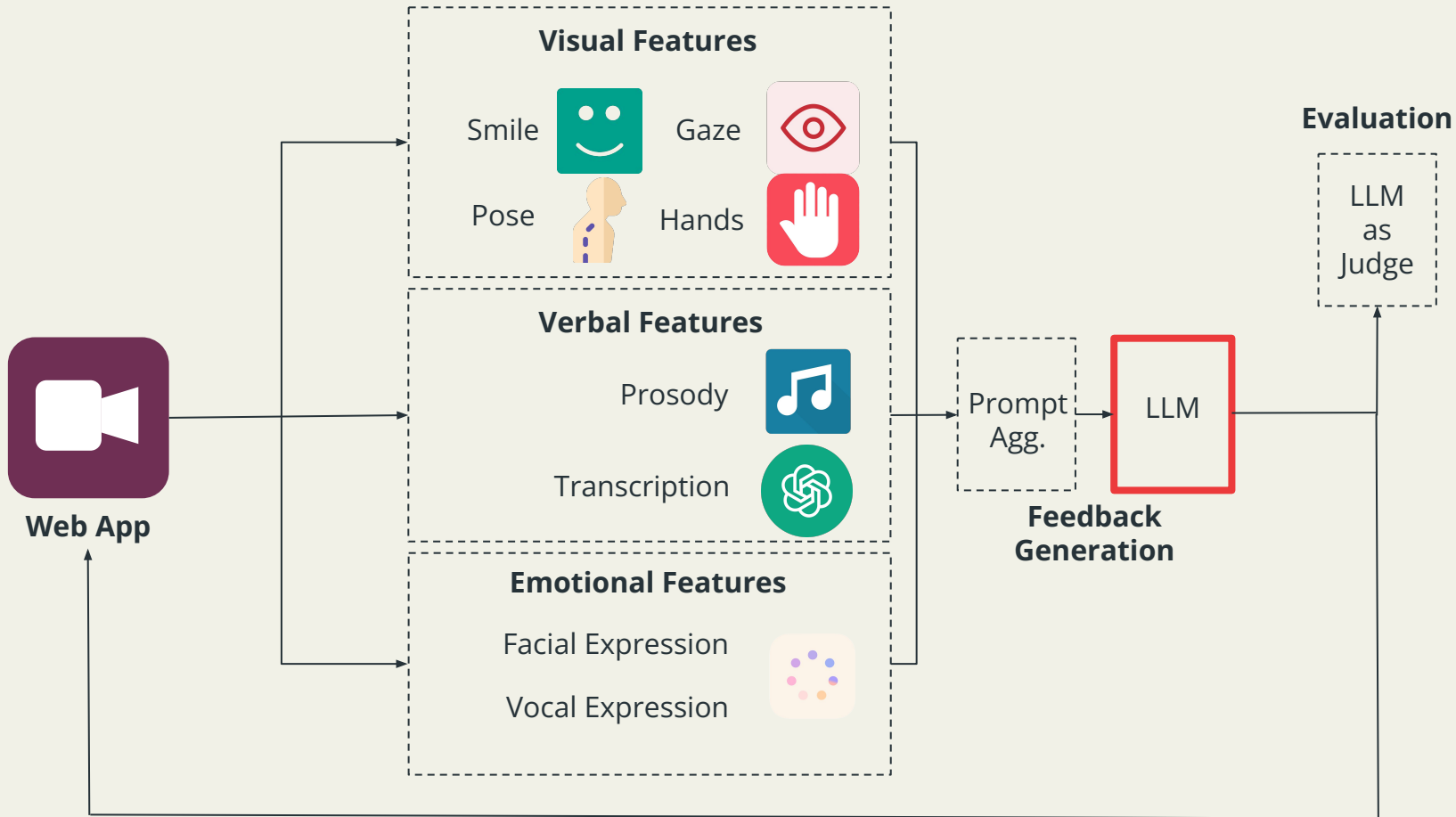


Emotional Features: Hume

- Hume has state-of-the-art models to **predict 48 distinct dimensions of emotion**.
- Used their expression measurement models to predict candidate's facial and vocal expressions.
- Example:

Beginning Timestamp	Ending Timestamp	Text	Facial Expressions	Vocal Expressions
0.0	4.85	"Its nice to meet you. [UM] I am doing great!"	"Joy", "Enthusiasm", "Happiness"	"Anxiety", "Doubt", "Fear"
...

LLM





Feedback Generation: Chosen LLM

- 2 million token context window
- Process up to 1 hour of video
- \$300 of free credits
- Process up to 50 minutes of video w/ audio
- Current SOTA in video understanding

Gemini

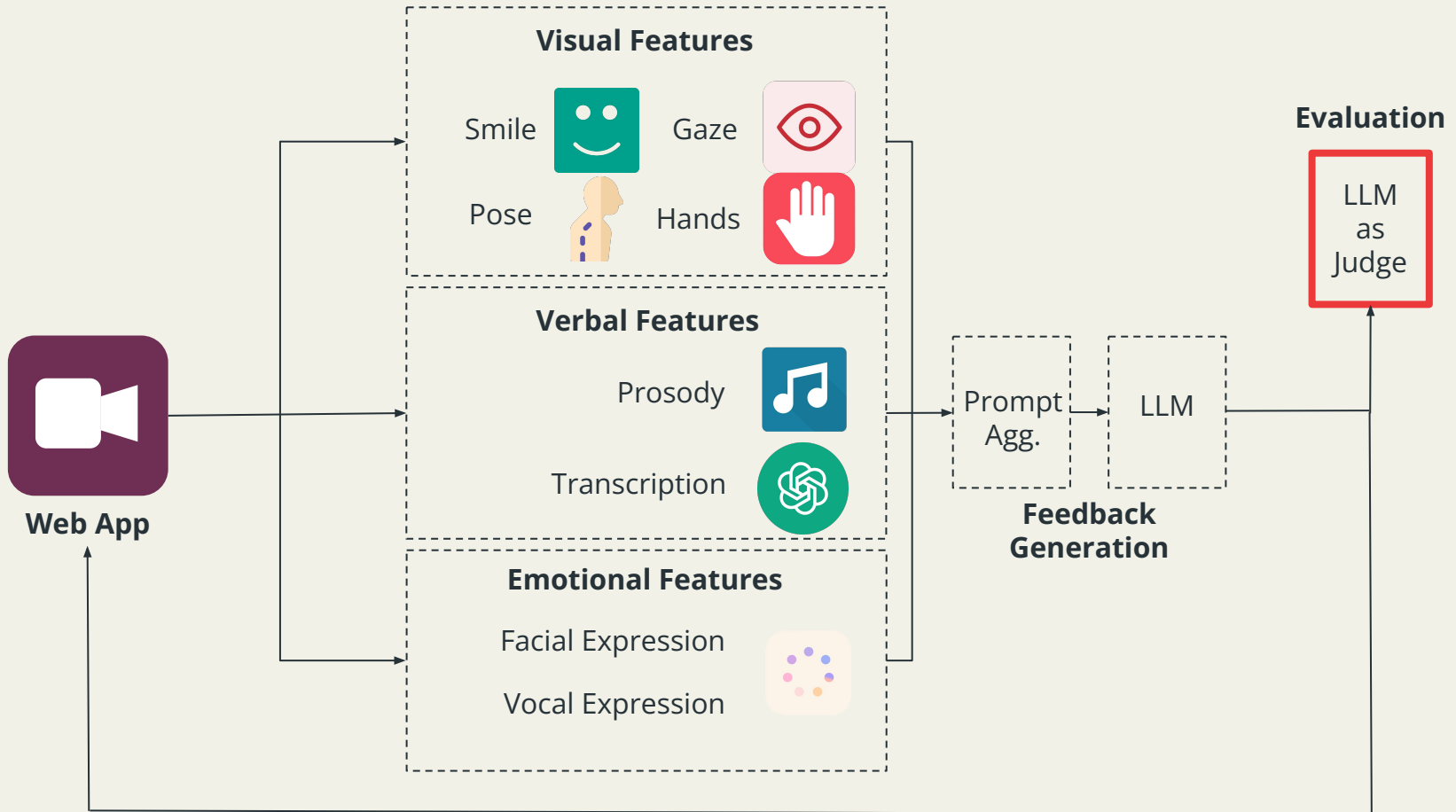




Evaluation and Discussion

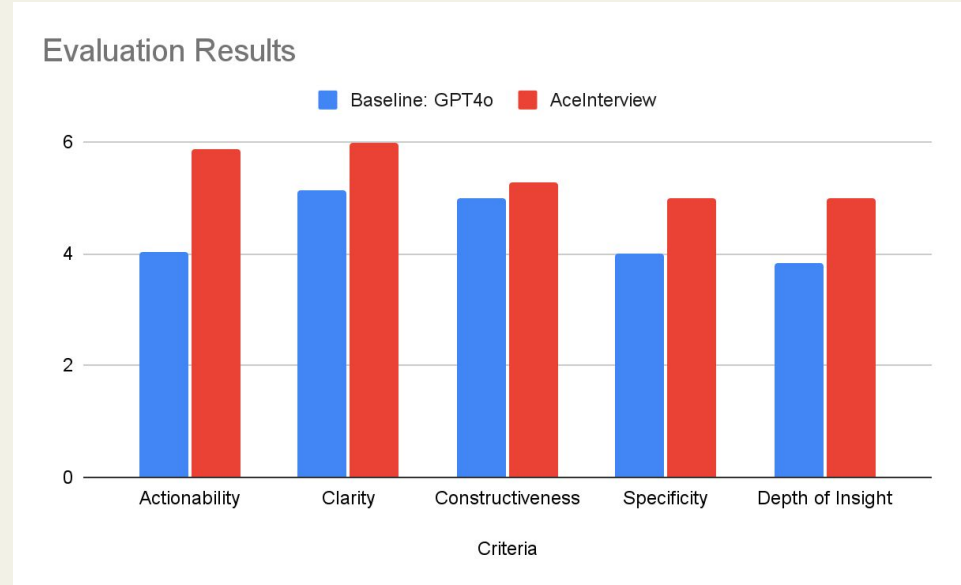


Evaluation



Evaluation: LLM as a Judge

- Gemini evaluated feedback on:
 - **Actionability**
 - **Clarity**
 - **Constructiveness**
 - **Specificity**
 - **Depth of Insight**
- Scored on a Likert scale from 1-7
 - 1 is poor feedback
 - 7 is excellent feedback



Takeaways and Summary

Challenges	Innovations
Video dataset not representative of typical online interview.	Visual feature extraction + evaluation tested on multiple camera angles and subject placements.
No traditional method for evaluation of LLM output.	LLM as a judge with generated Likert scale scores for human evaluation.
Difficulty hosting and deploying to AWS.	Deployed part of pipeline to AWS Lambda and Sagemaker. Other features requiring less processing done locally in-app.

Future Roadmap



Virtual Avatar



Knowledge Graph



Human Evaluation



Record Directly on App

Conclusion and Mission

AceInterview is an accessible, easy-to-use AI practice interview platform that analyzes and provides feedback on user posture, speaking tone, eye contact, facial expressions, and other key interview behaviors.

Appendix



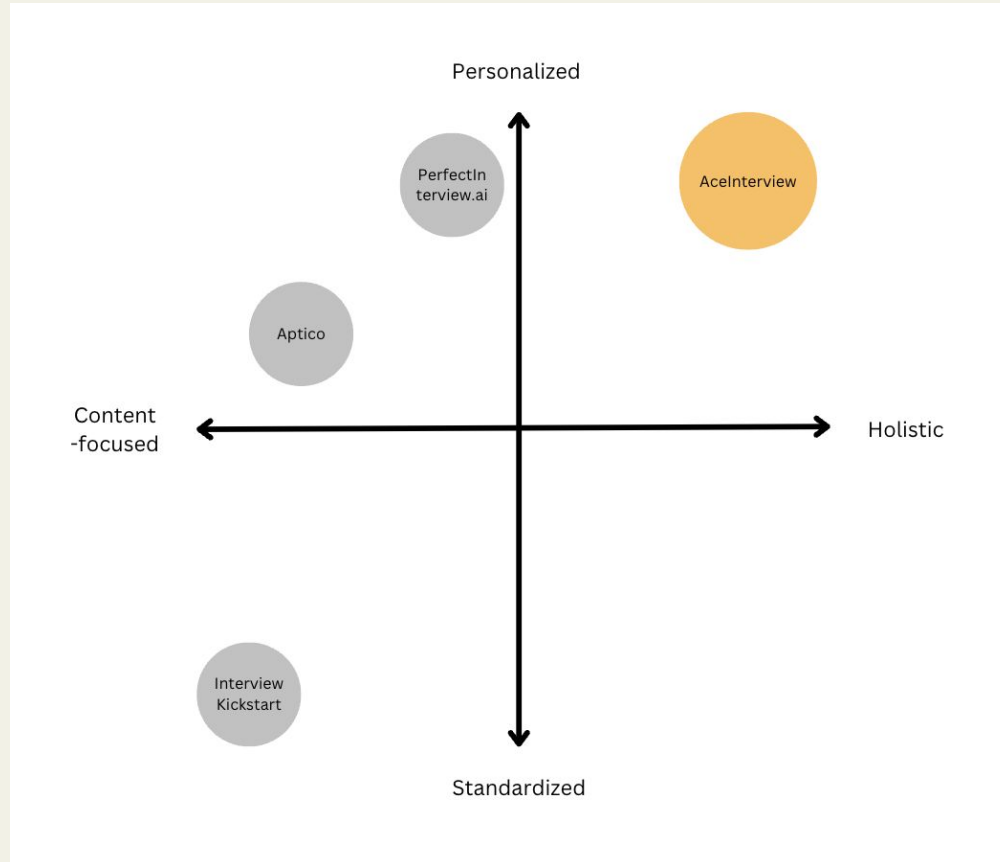
Acknowledgements

- Ehsan Hoque, PhD. Professor of Computer Science, University of Rochester
- MediaPipe
- OpenCV Haar Cascades
- GazeTracking: <https://github.com/antoinelame/GazeTracking>
- Hume
- Google Gemini

Literature review:

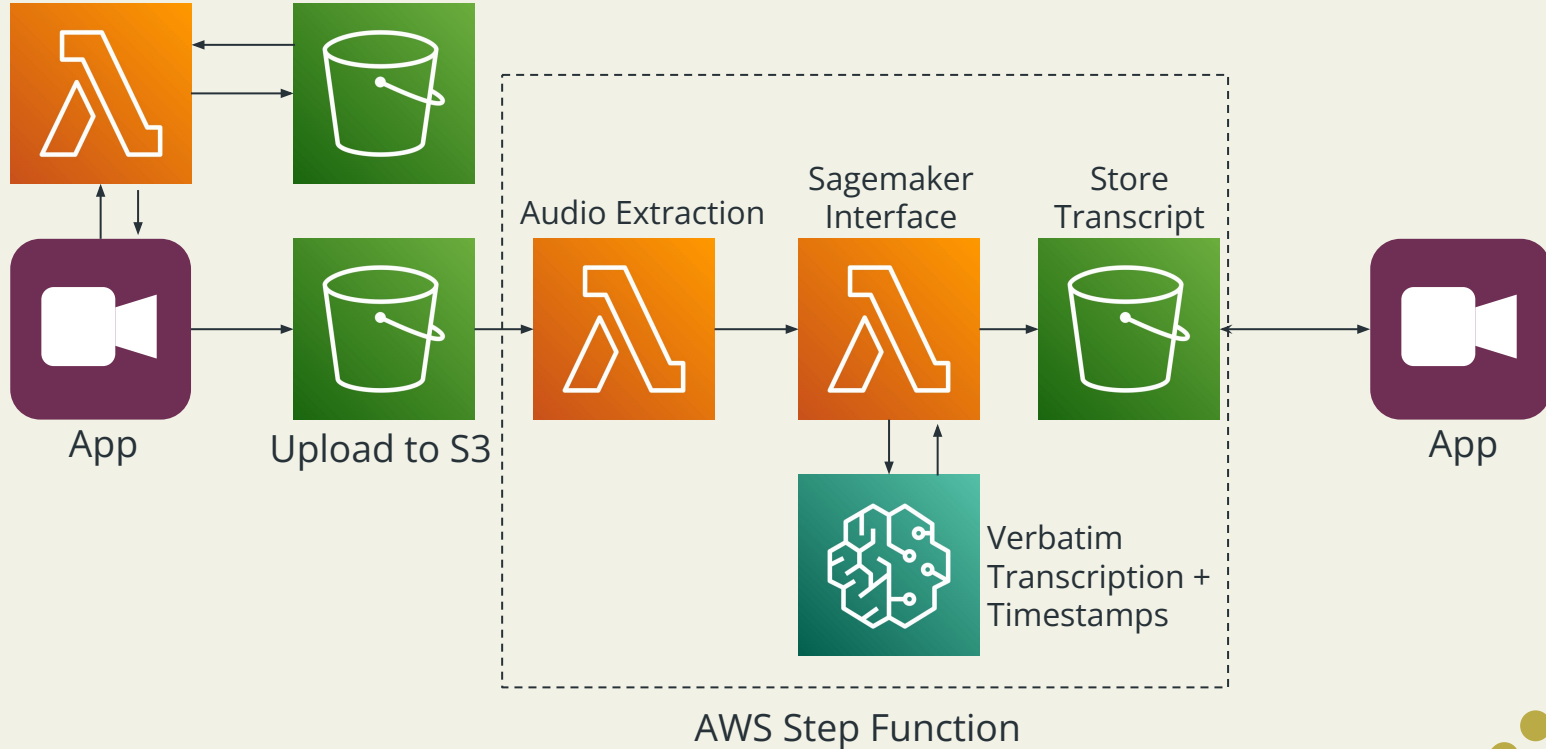
- Voice: https://www.researchgate.net/publication/319432789_The_effect_of_voice_quality_on_hiring_decisions
- Automated video interview judgment on a large-sized corpus collected online <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8273646>
- What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. <https://psycnet.apa.org/buy/2009-21033-003>
- Gaze: <https://www.nature.com/articles/s41598-024-60371-5>
- Authenticity & Focus: <https://link.springer.com/article/10.1007/s10869-024-09949-4>

Market Landscape



AWS Data Pipeline

Get Storage URL





Visual EDA/Cleaning


Attempts to adjust our video set to be more generalizable towards other orientations:

Initial Approach: Image Rectification

Pros:

- Angle more straight on

Cons:

- Significantly altered subject proportions
 - Extra black space due to image plane rotation
 - Have to hard code original and desired coordinates for each frame
- 


*Pictures removed for subject
privacy*



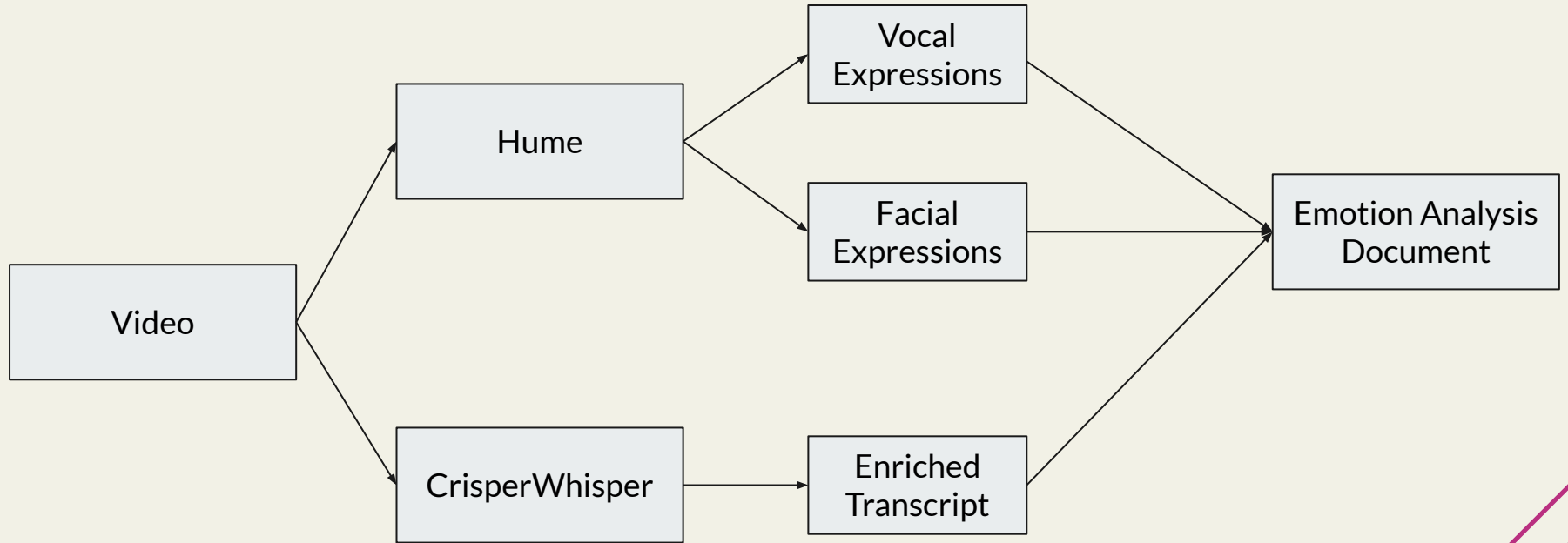
Bounding Box

Attempts to adjust our video set to be more generalizable towards other orientations:

*Pictures removed for subject
privacy*

- Initially used YoloV8 model
 - Realized we didn't need it for most feature extractors
 - -> Still needed it for smile
 - Switched to Haar Cascade
 - Much faster
 - More accurate on tilted subjects
- 

Vocal and Facial Expressions Classification



Feedback Generation



- 2 million token context window
- Process up to 2 hours of video
- Process up to 19 hours of audio
- Enhanced multimodal capabilities and cost effective
- Current SOTA in video understanding
- #1, #5 Video-MME benchmark
 - Pro: Overall 75%
 - Flash: Overall 70%

- Most widely known AI product
- 128k token context window
- Process up to 250 frames
- No Audio
- #3, #10 Video-MME benchmark
 - *with 384 frames
 - 4o: Overall 72%
 - 4o-mini: Overall 65%

Standard Metrics

- We calculate the RMSE for each attribute
- Example attributes:
 - RecommendHiring
 - EyeContact
- Most RMSEs were similar between Baseline and Video Only

Attribute	Baseline	Video Only
Authentic	0.70095234	0.84781956
Calm	1.05882295	1.13392491
EngagingTone	0.89481859	1.29687651
Excited	1.36397133	1.7089065
EyeContact	1.04308636	1.39333704
Focused	0.71726284	1.05557266
Friendly	1.09923989	1.46917645
NoFillers	1.54014271	1.39872312
NotAwkward	1.46059734	1.11646268
NotStressed	1.15015621	1.20286205
Overall	0.9920722	1.05212523
Paused	0.78559499	1.56994116
RecommendHiring	1.19927898	1.17980714
Smiled	1.29376486	1.59261389
SpeakingRate	1.06631201	1.01139593
StructuredAnswers	1.21834727	1.13813156

Rater Agreement

We calculate Krippendorff's alpha using ordinal squared distance

- Measures Rater Agreement
 - Does our model agree with the humans?
- Compared Model Raters to the Mean of the Human Raters
- GPT-4o and Gemini perform similarly with low agreement with the mean rater.
- Shows that out of the box the models are still decently far away from closely approximating Human Raters on subjective self-presentation.

Attribute	Baseline	Video Only
Authentic	0.05101822	0.07458904
Calm	0.21985141	0.02114472
EngagingTone	0.00130443	0.01811107
Excited	0.06401514	0.02314798
EyeContact	0.20818629	0.08554797
Focused	0.01250786	0.00916526
Friendly	0.16173645	0.16794741
NoFillers	0.09121396	0.08806791
NotAwkward	0.29478507	0.1652248
NotStressed	0.23325931	0.04204097
Overall	0.26733474	0.0809744
Paused	-0.0016599	0.34733895
RecommendHiring	0.3217199	0.12064629
Smiled	0.15836168	0.01430296
SpeakingRate	0.54622691	0.0828559
StructuredAnswers	0.27476163	0.01931648



Feedback Example (Structure)


However, some responses could benefit from more structure and depth. For example, when discussing a time you demonstrated leadership (1:04), you provided a good overview of your role in the Camp Kesem fundraising event, but you could strengthen this response by using the STAR method (Situation, Task, Action, Result).

* **Situation:** Briefly describe the context. For instance, “As the fundraising lead for Camp Kesem, we needed to raise \$50,000 to run the camp, which provides free summer experiences for children whose parents have cancer.”

* **Task:** Explain your specific responsibility. “My task was to spearhead the annual SAE date auction and organize the fundraising efforts.”

* **Action:** Detail the steps you took. “I organized our members into committees, delegated tasks such as soliciting donations and securing gift cards, managed logistics, and ensured everyone stayed on track.”

* **Result:** Describe the outcome. “Through these efforts, we successfully raised [amount] for Camp Kesem, exceeding/meeting our goal and enabling the camp to continue its mission.”






Feedback Example (Vocal Presentation)

You maintain consistent eye contact with the interviewer, which conveys engagement and attentiveness.

However, there are areas for improvement in your delivery. You used filler words like “um,” “uh,” “so,” “like,” and “I guess” frequently


(e.g., 0:08, 0:10, 0:14, 0:23, 1:51). While occasional fillers are natural, excessive use can distract the listener and diminish the impact of your message. Your vocal tone was mostly steady and calm, but injecting more enthusiasm and variation in your pitch could make your responses more engaging and captivating. For instance, when describing your passion for computational biology (0:20), elevating your tone and speaking with more energy could better convey your excitement and genuine interest. Additionally, varying your pace and using pauses strategically can add emphasis to key points and improve clarity.





Feedback Example (Body Language)


Your posture is generally good, with no slouching observed. However, you frequently tilted your head (see `pose_analysis` data), which, while not inherently negative, could be perceived as a sign of uncertainty or hesitation in some contexts. Aim for a more neutral head position to project confidence. Your hand gestures, though not excessive, sometimes seemed somewhat repetitive or unfocused (e.g., 0:19). Practice using more purposeful hand movements that reinforce your message and add visual interest to your communication. While the limited number of smiles (see `smile_count` data) isn't necessarily a problem, incorporating more smiles, especially when discussing positive experiences or your passion for the field, can create a warmer and more engaging impression. Your facial expressions generally align with the content of your responses, but incorporating more expressive nonverbal cues could further enhance your communication.





Feedback Example (Overall Impression)

You presented yourself as a thoughtful and engaged candidate with a clear interest in computational biology and relevant experiences. Your involvement in community service and leadership roles showcases valuable transferable skills. However, refining your communication skills by minimizing filler words, modulating your tone and pace, and using more purposeful nonverbal cues can significantly elevate your overall presentation and create a more compelling impression.





Feedback Example (Recommendation)

* **Filler Words:** Practice pausing briefly before speaking to gather your thoughts. Replace filler words with silence. Record yourself answering common interview questions and identify your most frequent fillers. Focus on eliminating one or two at a time.

* **Vocal Delivery:** Practice varying your pitch and pace. Record and listen to your responses, paying attention to your tone and energy levels. Practice emphasizing key points with changes in your voice.

* **Nonverbal Communication:** Practice your responses in front of a mirror or record yourself to observe your posture, head movements, and hand gestures. Aim for a neutral, confident posture and use deliberate, expressive hand movements. Practice incorporating genuine smiles at appropriate moments.

* **Structured Responses:** Practice the STAR method for behavioral questions. Prepare specific examples that showcase your skills and accomplishments, ensuring you quantify your results whenever possible.

