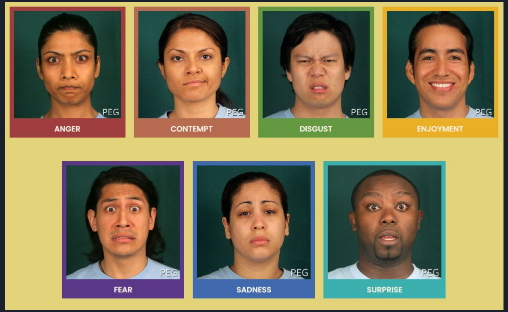
Emotional Dialog Agent

COGS 4640 Project Proposal

Ryan Carroll & Joshua Pile

Objective

- We want to create an intelligent agent to act as a companion that can provide support in a way unique to most popular chatbots. This mainly revolves around better detection and understanding of user emotion and appropriately responding through multimodal input.
- To accomplish this, we want it to
  - Classify input into emotions based on Ekman's Universal Emotions
    - Facial expression data, conversation context, tone of voice
  - Gives appropriate responses based on emotion and conversation context
    - I.e. can detect a user is sad without being told and tries to cheer them up
    - Also could include a (described) facial expression, tone, etc.
- This would make it a much more human and empathetic feeling companion. We are planning to utilize this for a topic that elicits strong emotional or requires high emotional context.
- Potentially as a therapist



Main Motivations

- Dialogue agents can only determine user emotions with explicit input
  - “I’m sad,” “I’m angry at A about X,” etc.
- Integrating other data allow agents to have implicit understanding of user’s emotions
  - Interpreting expressions, tone of voice
  - Ability to emulate emotions

System Design

- The system receives input data
  - User’s facial expression data
  - User’s audio feature data
- Data is used to describe the user’s current emotional state according to Ekman’s classifications
- Classification + the conversation context is used to generate an emotionally appropriate response to serve back to the user
- This response is then described by the system as an Ekman classification
- The diagram for the system design is presented on the following slide

System Design

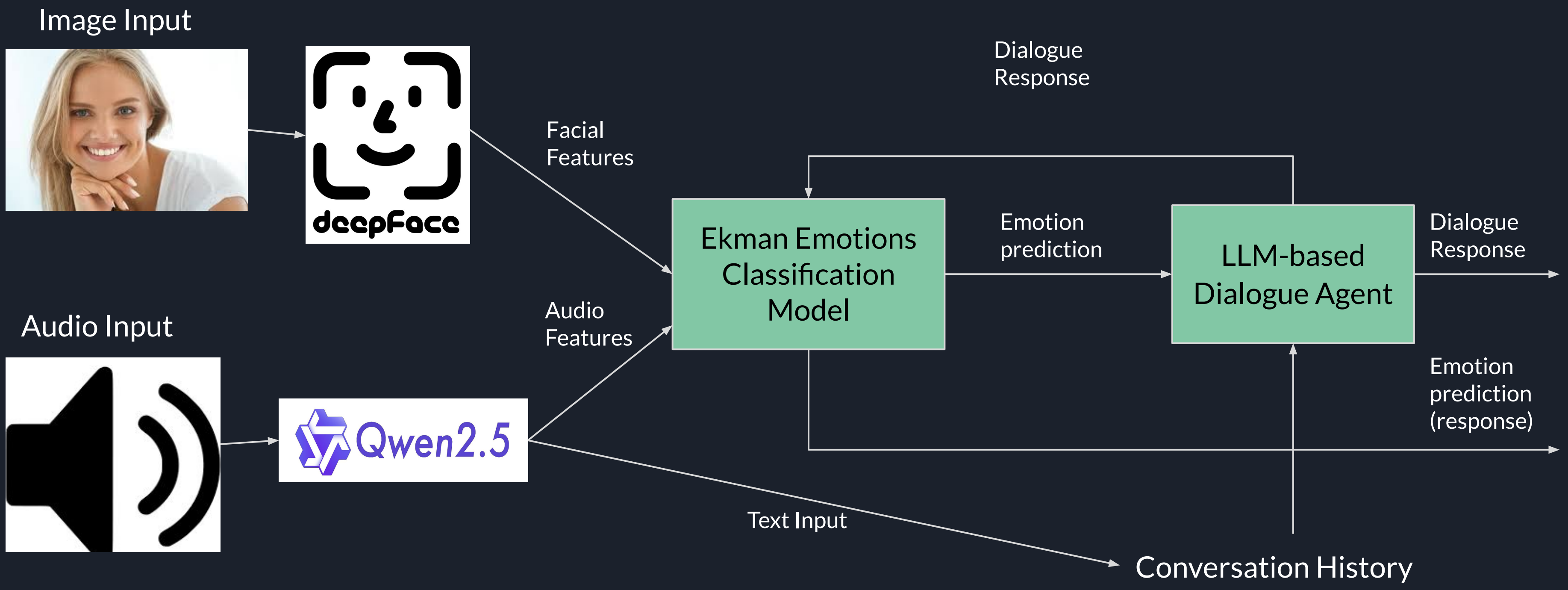


Image Input

Facial Features

Audio Input

Audio Features

Ekman Emotions Classification Model

Dialogue Response

Emotion prediction

LLM-based Dialogue Agent

Dialogue Response

Emotion prediction (response)

Text Input

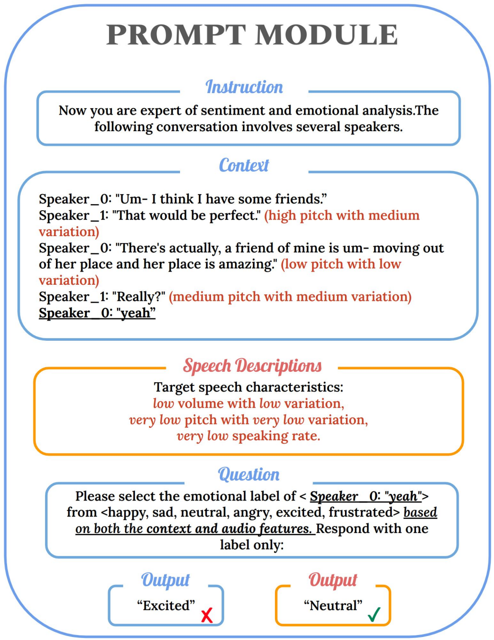
Conversation History

Evaluation Metrics

- Objective Evaluation
  - Give bot inputs that align with specifically chosen Ekman classifications
    - i.e. a genuine smile with dialog about excitement for happiness
  - Grade on accuracy of classification
- Subjective Evaluation
  - Either:
    - Record interactions with bot and have others “grade” responses
    - Have users interact with bot and fill out survey
  - In either case, ask about
    - Appropriateness of response
    - Engagement
    - Coherence

Leading Methods In The Field

- There is a plethora of other research articles that attempt to improve the emotions modeling detection of LLM based chat
- The main one that we took inspiration from an article called Beyond Silent Letters: Amplifying LLMs in Emotion Recognition with Vocal Nuances.
- LLMs lack the ability to directly process audio inputs. This limitation has hindered their potential in multimodal emotion recognition tasks that involve both textual and audio data.
- The research enabled Large Language Models to perform emotion analysis in speech by translating audio features into natural language descriptions that can be integrated into text prompts, bridging the gap between audio and text modalities without architectural modifications.
- Their approach is exemplified through the diagram on the right.



PROMPT MODULE

Instruction

Now you are expert of sentiment and emotional analysis.The following conversation involves several speakers.

Context

Speaker_0: "Um- I think I have some friends."
Speaker_1: "That would be perfect." (high pitch with medium variation)
Speaker_0: "There's actually, a friend of mine is um- moving out of her place and her place is amazing." (low pitch with low variation)
Speaker_1: "Really?" (medium pitch with medium variation)
Speaker_0: "yeah"

Speech Descriptions

Target speech characteristics:
low volume with *low* variation,
very low pitch with *very low* variation,
very low speaking rate.

Question

Please select the emotional label of < *Speaker_0: "yeah"*> from <happy, sad, neutral, angry, excited, frustrated> *based on both the context and audio features.* Respond with one label only:

Output

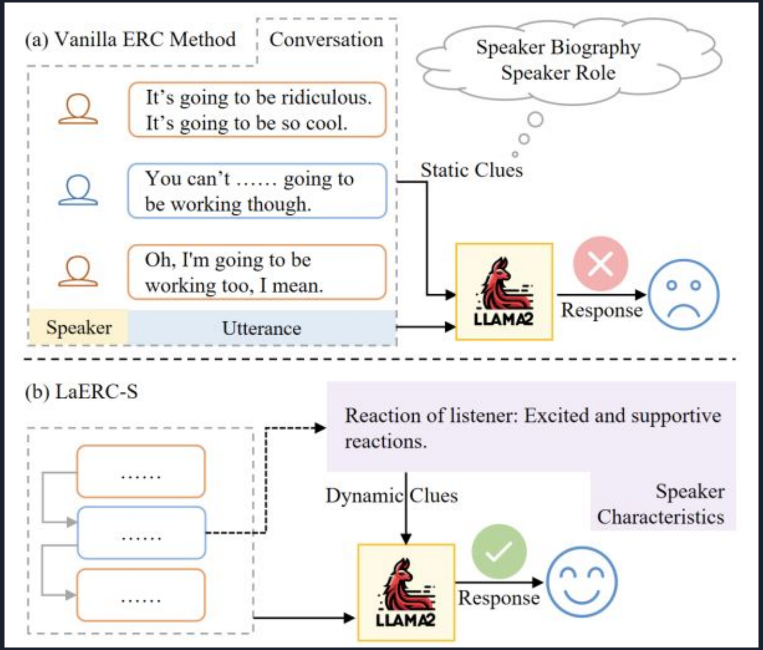
"Excited" ✗

Output

"Neutral" ✓

Leading Methods In The Field

- This is another research article we drew direction from under the name “LaERC-S: Improving LLM-based Emotion Recognition in Conversation with Speaker Characteristics”
- LaERC-S is a novel framework that stimulates LLMs to explore speaker characteristics involving the mental state and behavior of interlocutors, for accurate emotion predictions.
- They utilized a two-stage learning approach that equips LLMs with the ability to analyze speaker characteristics and monitor emotional patterns throughout complex conversational exchanges.
- This is seen in the right diagram


(a) Vanilla ERC Method
Conversation
Speaker Biography
Speaker Role
It’s going to be ridiculous.
It’s going to be so cool.
You can’t ...... going to be working though.
Oh, I'm going to be working too, I mean.
Static Clues
Speaker
Utterance
LLAMA2
Response
(b) LaERC-S
Reaction of listener: Excited and supportive reactions.
......
......
......
Dynamic Clues
Speaker Characteristics
LLAMA2
Response

Our Improvements

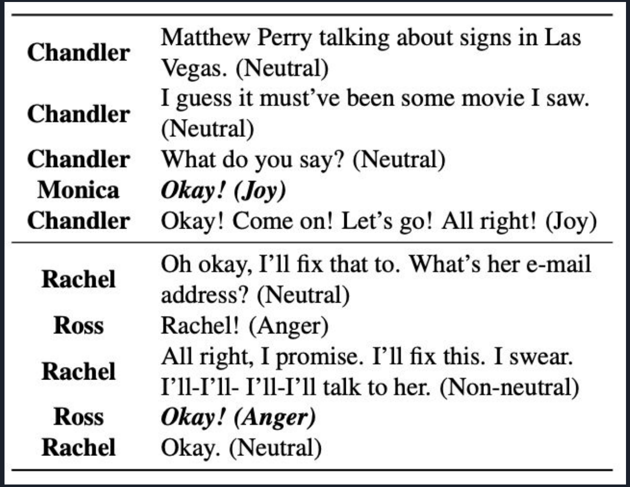
- Using these two research articles as a jumping off points we are attempting to create a LLM dialogue agents that can determine user emotions with cue from multimodal inputs like voice and visual facial expressions and respond accordingly.
- The first is the improvement is that of the utilization of the users facial visualization as features for the Ekman classification which wasn't considered in the aforementioned research article.
- We understand that this might alway seem the most practical as most people that are interacting with the chatbot will hold a neutral expression however we will choose and emotionally arousing topic for exemplify the system's potential.
- Furthermore , we also plan to leverage traditional ML models to help the classification as compared to that of feeding the features to the LLM itself leading to less ambiguity in feature interpretation and potentially higher accuracy through ensemble techniques

Project Phases

- Phase 1
  - Input: Textual description of face + conversation
  - Output: Ekman classification of input
- Phase 2
  - Input: Textual description of face + conversation
  - Output: Textual response for user + Ekman classification of output
- Phase 3
  - Input: Video and audio of user
  - Output: Textual response for user + Ekman classification of output w/ facial description

Feasibility

- While the project does seem like a large project given the short time frame that we were given we believe that we can make meaningful progress on the project we have decided on.
- One of the biggest reasons that we believe we can do this is due to the fact that there are pre existing models for the visual face emotions features as well as for turning the audio input into descriptive textual feature, both for accurate model predictions.
- The bulk of the time spent on this project will deal with finding relevant datasets or creating our own for the Ekman prediction model. Training the model itself may also pose a challenge depending on the complexity of the features. As well as fine-tuning the LLM dialogue agent to respond in a appropriate manner to the user input in addition to the given predicted Ekman emotion.



Chandler	Matthew Perry talking about signs in Las Vegas. (Neutral)
Chandler	I guess it must've been some movie I saw. (Neutral)
Chandler	What do you say? (Neutral)
Monica	***Okay! (Joy)***
Chandler	Okay! Come on! Let's go! All right! (Joy)
Rachel	Oh okay, I'll fix that to. What's her e-mail address? (Neutral)
Ross	Rachel! (Anger)
Rachel	All right, I promise. I'll fix this. I swear. I'll-I'll- I'll-I'll talk to her. (Non-neutral)
Ross	***Okay! (Anger)***
Rachel	Okay. (Neutral)

EmotionLines Dialogue Example

References

- https://qwen.readthedocs.io/en/latest/
- https://arxiv.org/pdf/1802.08379
- https://arxiv.org/html/2407.21315v1
- https://aclanthology.org/2025.coling-main.451.pdf
- https://github.com/serengil/deepface