Analysis of Multi-Modal LLMs on Facial Expression Recognition

Jonathan Liu

liujonathan24@gmail.com

Abstract

- **Motivation:** Facial expression recognition (FER) is a field that has far-reaching applications in lecturing, security, medical rehabilitation, safe driving, and assistive technology for both the low-vision/blind and neurodivergent communities.
- **Background:** The recognition of facial expressions is a vital human skill for establishing affective relationships and in gaining and utilizing social skills.
- **Problem Statement:** Current models use support vector machines on extracted features such as linear binary patterns (LBP) and Oriented and Rotated Brief (ORB) or CNN and RNN architecture neural network models. However, their inexplainability and inaccessibility prevent their applications for day-to-day cases [3]. They also require specific training datasets and compute [3].
- **Hypothesis and Proposed Solution:** The author proposes an end-to-end prompt sequence for public, multi-modal LLMs that could become the basis for FER assistive technology. One-shot/Multi-shot prompting as well as Tree of Thought techniques are expected to improve accuracy. [4, 5]

Methods

Materials

- **Dataset:** Extended Cohn-Kanade Dataset [2] with the facial emotions anger, contempt, disgust, fear, happiness, sadness, and surprise.
- **Multi-Modal LLM:** Bakllava, a multimodal model created by SkunkworksAI consisting of the Mistral 7B base model augmented with the LLaVA architecture.

Procedure & Proposed Models

1. **Feature Extraction & SVMs:** An RBF kernel SVM was trained on 90% of CK+ ORB and LBP vectors and tested on the remaining data. This process was repeated 10 times, and the accuracy percentages are averages from all trials.
2. **LLM Zero-/One-/Multi-shot & Tree of Thought:** Specified prompts and vector-embedded CK+ images were fed into Bakllava with Langchain and the output emotion was recorded.

Tree of Thought Model

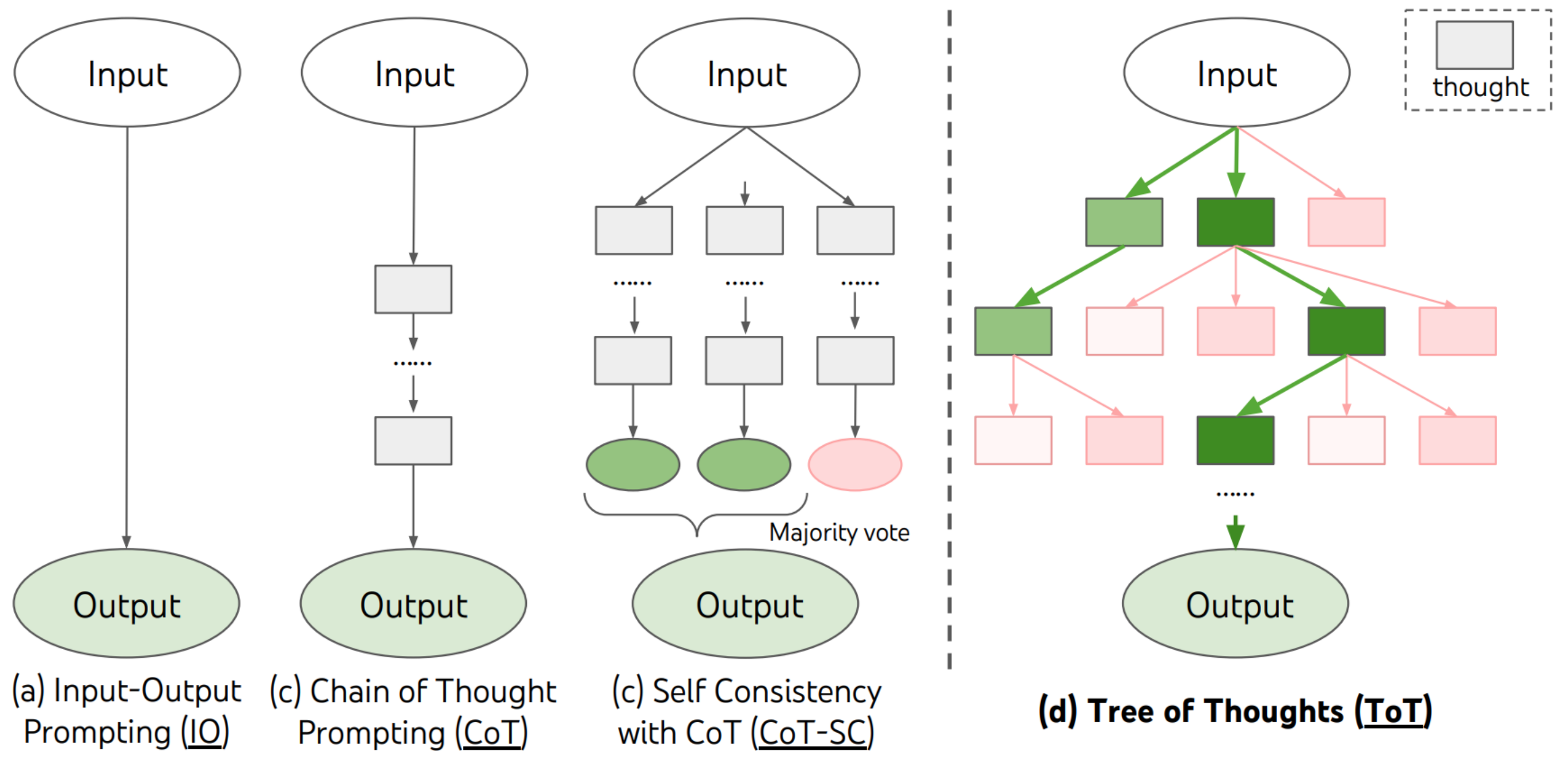
Input
Input
Input
Input
thought
Majority vote
Output
Output
Output
Output
(a) Input-Output Prompting (IO)
(c) Chain of Thought Prompting (CoT)
(c) Self Consistency with CoT (CoT-SC)
(d) Tree of Thoughts (ToT)

Figure 1. Schematic of problem-solving approaches with LLMs. Each rectangle box represents a thought, which is a coherent language sequence that serves as an intermediate step toward problem solving. [5]

Results

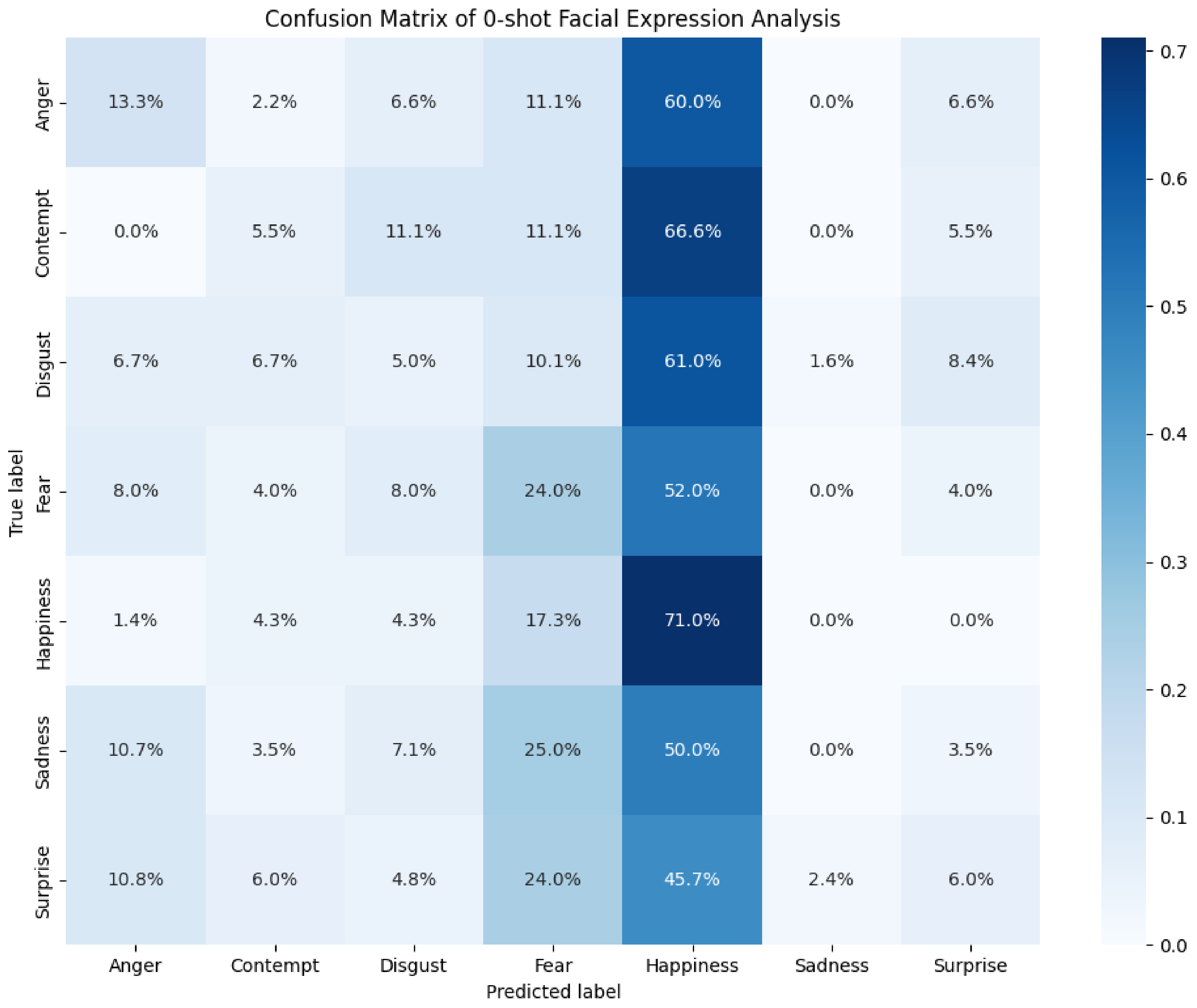

Figure 2. Confusion matrix from prompting the bakllava LLM with the processed prompts with an overall accuracy of 21.4%

Dataset	Zero-shot	One-shot	Multi-shot	ToT	SVM
CK+	21.4	15.2	15.3	12.8	91.5

Table 1. Accuracy of the traditional Baseline SVM model on LBP and ORB patterns compared to that of the experimental Zero-shot, One-shot, Multi-shot, and Tree of Thoughts prompts on the Bakllava LLM

Sample Prompt Sequence

Zero-/One/Mult-shot Prompts:

- Optional One/Multi-shot Examples
- Prompt:

  Answer the question based only on the following image: <Image>
  Question: Choose the answer that best represents the emotion on the person's face in the picture:
  0: anger 1: contempt 2: disgust 3: fear 4: happiness 5: sadness

Tree of Thought Prompt:

1. ... <image>. Could you brainstorm three distinct emotions ... consider ... the position of the person's eyebrows, mouth, eyes, and nose.
2. For each of the three proposed solutions, evaluate how true they are to the input image. ... Assign a probability ... to each option ...
3. For each solution, ... verify that the described [features] accurately describe the predicted emotion.
4. Based on the evaluations and scenarios, rank the emotions in order of promise ...

Discussion

The One-shot, Multi-shot, and Tree of Thought prompts designed to shore up the LLM's knowledge base and to improve accuracy instead resulted in significantly decreased accuracies. However, clear patterns and realizations emerge:

1. **Very Low Accuracy:** The experimental, LLM-based classifiers' maximum accuracy of 21% is not comparable to the baseline classifier, which had a 91.5% accuracy.
2. **Biased Predictions:** The Zero-shot classifier was extremely biased towards ”Happiness.” The addition of examples in the One- and Multi-shot classifiers were able to decrease the fixation at the expense of accuracy. Unexpectedly, the Tree of Thoughts model also displayed a similarly large bias, but towards ”Anger.”

Future Work

The results indicate that the promise of LLMs are not yet realized. However, several limitations of this study can be improved in future studies.

- **Alternative Datasets:** Future work by the author will consider the JAFFE, MMI, and FER-2013 datasets to generalize the patterns discovered.
- **Private LLMs:** Due to the problem statement and resource limitations, only open source multi-modal LLMs were used, however, the need for a higher accuracy indicates that private LLMs such as Open AI's GPT-4, Google's Gemini Advanced, or Anthropic's 3 Claude should be used in further experiments.
- **Prompt Engineering:** Prompt Engineering is still a new field, to the point that many LLM APIs are not fully implemented. As the field develops, documentation clarity will allow for faster development and more advanced prompt engineering techniques for in-context learning will be developed.

Acknowledgements & Disclosure of Funds

I would like to thank the Yale College Admissions Office for sponsoring this poster. This research is part of the efforts to continue developing the author's project, Hear the Room, awarded by HackDisability sponsored by Amazon, MIT, and the Perkins School for the Blind.

References

[1] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski.
A survey on facial emotion recognition techniques: A state-of-the-art literature review.
Information Sciences, 582:593–617, 2022.

[2] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews.
The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression.
In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[3] Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, and Joel J.P.C. Rodrigues.
A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines.
Alexandria Engineering Journal, 68:817–840, 2023.

[4] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, Enze Shi, Yi Pan, Tuo Zhang, Dajiang Zhu, Xiang Li, Xi Jiang, Bao Ge, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang.
Review of large vision models and visual prompt engineering.
Meta-Radiology, 1(3):100047, 2023.

[5] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan.
Tree of thoughts: Deliberate problem solving with large language models.
In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc., 2023.

YES Scholars Poster Session