

Pediatric Emotion Recognition Using ChatGPT, Grok and DeepSeek

Human Centered Generative AI Final Report

Project Authors: Chuanling Qin, Sreshta Shetty

Contact: chuanling.qin.gr@dartmouth.edu, sreshta.m.shetty.th@dartmouth.edu

March, 08, 2025

1. Motivation and Introduction

In recent years, accurate and timely recognition of emotional states in children has attracted growing interest from researchers and clinicians due to its critical implications for early mental health intervention and psychological development [3, 1]. Early identification of emotional difficulties can significantly enhance therapeutic outcomes by enabling proactive and tailored interventions. However, traditional approaches to emotion recognition often face challenges arising from children’s developing communication skills, evolving expressive capacities, and their varying willingness to share emotional experiences [4].

Advancements in artificial intelligence (AI), particularly through Generative AI (GenAI) models such as ChatGPT (OpenAI), Grok (xAI), and DeepSeek (DeepSeek AI), present novel avenues to address these limitations. These models have demonstrated considerable capabilities in interpreting emotional expressions by leveraging both verbal (textual) and non-verbal (visual) cues. Nonetheless, prior research indicates potential biases in GenAI models trained predominantly on adult datasets, which may lead to reduced accuracy when recognizing emotions in children, whose emotional expressions differ significantly from adults [5, 2].

This study explicitly investigates these biases, comparing the differential performance of prominent GenAI models—GPT-4o, Grok-2, and DeepSeek—on emotion recognition tasks involving adults and children. Given that major AI models primarily train on adult datasets, we hypothesize reduced emotion recognition accuracy for children relative to adults. Specifically, we aim to: (1) assess differences in emotion recognition accuracy across adult and child populations; (2) evaluate whether incorporating both verbal and non-verbal cues significantly enhances performance compared to using verbal cues alone; and (3) identify cases in which environmental or contextual factors cause misclassification despite clear emotional signals.

2. Data Collection and Data Description

Existing large scale vision and language datasets have significant limitations in capturing comprehensive emotional context. Current datasets typically offer only superficial visual descriptions or action instructions, which fail to provide meaningful insights into human emotions. Traditional captions often merely describe scenes at a surface level, lacking depth in emotional representation. We utilized advanced off-the-shelf models to accomplish several key tasks. Our approach leverages YouTube as a content repository, enabling us to collect a diverse and rich dataset. To further enhance the dataset’s emotional and contextual richness, we implemented a targeted manual annotation process for a subset of video clips. All code and data are available on GitHub: <https://github.com/antoniojustin/GenAI>.

3. Method

3.1 Data Preprocessing

We collected 24 video clips from various sources, as detailed in Section 2, to ensure a diverse range of video resolutions, environments, and daily interactions. These videos were divided into two sets:

- **Adult set:** Observable age > 20 years
- **Children set:** Observable age < 10 years

For each video, we manually selected 3 frames based on the following criteria:

- At least one frame contained a clear emotional expression from the individual that was contextually linked to their behavior.
- The individual’s age was visually distinguishable, making it easy to classify them as either a child or an adult. This classification was further validated using the video’s caption (e.g., a video titled “Recordings from an adult playground in NYC”).
- We could clearly determine whether the recorded activity was staged (e.g., from a movie) or spontaneous (e.g., real-life footage).

These criteria ensured that we could accurately assess the difficulty of classifying emotions. Each video contributed three frames featuring a primary individual, resulting in a total of 36 frames for both the children and adult sets.

We processed the videos into image frames because most generative models struggle to interpret both video and text cues simultaneously. However, models such as ChatGPT (GPT-4V), Grok-3, and DeepSeek demonstrate some capability in understanding stationary images alongside text. Therefore, we paired each selected video frame (representing non-verbal cues) with its corresponding caption (representing verbal cues) as a unified input for the AI models, prompting them to infer the primary individual’s emotion based on both verbal and non-verbal information.

Among the 12 video clips in each set, we categorized them into two subsets:

- **Actor Set:** Frames from movies or clearly staged activities.
- **Naturalist set:** Frames from spontaneous interactions.

Within each subset, we further assessed whether the verbal and non-verbal cues aligned. Cases where the cues conflicted were considered more challenging, resulting in nine video frames per subset. We then evaluated the performance of leading commercial LLMs (ChatGPT - GPT-4V, Grok-3, and DeepSeek). Since DeepSeek primarily relies on textual cues from attachments, we anticipated that it might be disadvantaged in this task, warranting further analysis of its results.

3.2. Prompt Engineering and Response Collection

We will approach emotion identification using both qualitative and quantitative methods. In the quantitative approach, we will instruct the model to classify emotions into one of the six universal emotional expressions (happiness, sadness, anger, fear, surprise, and disgust) plus a neutral expression. The neutral expression is particularly important as it serves as a baseline for emotion comparison across different scenarios. Before classification, the LLM will first define these emotional expressions using the following prompt:

```
'Provide a detailed definition of the six universally recognized human emotional
→ expressions (happiness, sadness, anger, fear, surprise, and disgust), including
→ their psychological significance and typical facial expressions. Additionally,
→ define the concept of a neutral expression, explaining its role in emotion
→ recognition and how it differs from emotional expressions'
```

Following this, the LLM will classify the emotional state of the major character by providing a single-word response based on the given verbal (text) and non-verbal (video frame) cues. The classification will be elicited using the following prompt:

```
'Based on the given verbal and non-verbal cues, determine the primary emotional state
→ of the major character. Select the most probable emotion from the following seven
→ categories: (1) Happiness, (2) Sadness, (3) Anger, (4) Fear, (5) Surprise, (6)
→ Disgust, or (7) Neutral. No explanation is needed.'
```

The model’s responses will be collected and compared to the ground truth labels created by human observers (Chuanling and Sreshta). Since we, as human observers, have a more comprehensive understanding after watching the full video and possess a natural perception of emotional states, our annotations will serve as the gold standard for evaluating the LLM’s classification accuracy.

In the qualitative approach, we will give the LLMs a second task, allowing them to freely describe their perception of the emotions. This will be done using the following prompt:

'Based on the given verbal and non-verbal cues, determine the primary emotion of the
 → individuals. Explain why in a short paragraph'

This task removes the constraint of selecting a discrete emotion category, allowing the LLMs to articulate their reasoning. The responses will be qualitatively analyzed to assess how the models perceive and justify emotional states beyond strict categorical classification.

3.3. Evaluation Metrics

For the discrete classification, we will provide a summary of overall performance metrics, including accuracy and F1-scores, which is cable of giving a general sense of performance for all 7 categories of emotions. For the qualitative analysis, we will highlight key findings and insights in the discussion section.

4. Results

The results align well with our hypothesis, showing that GenAI models perform better for adults and struggle more with children. In Section 4.1, we will evaluate their performance on the discrete classification task, where models were required to assign a single label from the seven universal emotions. In Section 4.2, we will analyze the models' responses to explore potential insights and key takeaways from our experiments.

4.1. Discrete Classification Task

The results indicate that GPT-4o outperforms grok-2-vision-1212 and DeepSeek-R1 in the discrete classification task for emotion recognition, achieving the highest overall accuracy (33.3%) and performing notably better for adults (37.1%) compared to children (30.0%). This trend is consistent across models, with all three demonstrating lower accuracy when classifying emotions in children. DeepSeek-R1, which relies solely on verbal cues, performed comparably to grok-2-vision-1212, suggesting that multimodal processing (text + image) provides only a modest advantage in this specific task.

Table 1: Accuracy Comparison of GenAI Models for Emotion Recognition Using Verbal and Non-Verbal Cues

| Model/Accuracy | Overall (%) | Children (%) | Adults (%) |
|--------------------|-------------|--------------|------------|
| GPT-4o | 33.3 | 30.0 | 37.1 |
| grok-2-vision-1212 | 28.3 | 25.9 | 30.0 |
| DeepSeek-R1* | 28.3 | 27.6 | 29.6 |

* DeepSeek-R1 model can only understand text (verbal cues)

In terms of F1-score, grok-2-vision-1212 slightly outperforms GPT-4o overall (23.1% vs. 21.8%), particularly for children (21.5% vs. 12.8%), indicating that it may achieve a better balance between precision and recall in certain cases. However, GPT-4o still achieves the highest F1-score for adults (38.4%), reinforcing the idea that GenAI models are more effective in interpreting adult emotions. DeepSeek-R1's performance remains competitive, despite its lack of image processing, further highlighting the significance of verbal cues in emotion recognition. These findings suggest that while multimodal models have advantages, text-based cues alone still contribute significantly to classification accuracy.

Table 2: F-1 Score Comparison of GenAI Models for Emotion Recognition

| Model/Accuracy | Overall (%) | Children (%) | Adults (%) |
|--------------------|-------------|--------------|------------|
| GPT-4o | 21.8 | 12.8 | 38.4 |
| grok-2-vision-1212 | 23.1 | 21.5 | 25.4 |
| DeepSeek-R1* | 21.8 | 19.5 | 28.1 |

* DeepSeek-R1 model can only understand text (verbal cues)

4.1. Free Form Answer Task

GenAI models demonstrate remarkable proficiency in identifying emotions based on facial expressions, non-verbal cues, and environmental context. We observe that models capable of processing images tend to prioritize visual information when making emotion classifications. In Figure 1, we highlight instances where the model accurately inferred emotions even from subtle cues. In image (a), the child’s facial expression alone does not clearly reveal her emotional state, yet the model correctly identified fear, recognizing that she had seen a snake. Similarly, in image (b), the child’s facial expression is ambiguous in the given frame, but based on contextual environmental cues, the model successfully inferred that she was feeling joy about receiving a dress—despite not being explicitly prompted to consider the surrounding elements. These examples underscore the model’s ability to integrate non-verbal information effectively, enhancing its emotion recognition capabilities beyond facial expressions alone.



(a) The primary emotion of the individual in the image appears to be surprise or fear. This is suggested by the wide eyes and raised hands, which are common physical reactions to encountering something unexpected or potentially frightening, such as a snake. The expression conveys a sense of alertness and readiness to respond to the situation.
-- GPT-4o



(b) The image likely conveys joy, highlighted by the young child’s bright smile and her blue princess dress with a tiara. Her outfit and expression suggest she’s excited to show off, possibly basking in the delight of family praise. The vivid colors and her cheerful demeanor create a lighthearted, happy vibe.
-- Grok2

Figure 1: The model correctly identifies fear and happy state with accurate description

Despite their strong performance, GenAI models exhibit notable limitations. One major challenge is difficulty recognizing human faces when blurred or in low resolution. In such cases, the model still generates emotion classifications, even when it struggles to recognize the subject as human—an issue evident in Figure 1(a). While models can extract contextual clues to compensate for missing facial details, this ability appears inconsistent, as it sometimes fails when the face is less defined. Additionally, we observe instances where the model detects a person but refrains from making any judgment, likely due to hardcoded constraints—as demonstrated in Figure 1(b). Lastly, an over-reliance on environmental context can lead to misinterpretations, where the model contradicts clear verbal and non-verbal cues, as seen in Figure 2(c). Nonetheless, these findings affirm that GenAI models possess impressive capabilities in analyzing images and synthesizing verbal and non-verbal information to infer human emotions effectively.

5 Discussion

Our study highlights several key findings regarding the performance of GenAI models (ChatGPT, Grok-2, and DeepSeek) in emotion recognition tasks that incorporate both verbal and non-verbal cues. First, we observed a notable performance gap between adults and children, with all three models performing significantly worse on children’s emotions. This suggests that children’s emotional expressions are more challenging to interpret, likely due to their more subtle and less exaggerated facial cues compared to adults.



(a) I'm sorry, I can't tell who this is based on the image alone.

-- GPT-4o



(b) I'm sorry, I can't help with identifying or making assumptions about people in images.

-- Grok2



(c) The primary emotion in the image appears to be excitement and enjoyment. The person in the foreground is energetically moving in the water, suggesting they are having fun. The presence of others around them in what seems to be a wave pool contributes to a lively and playful atmosphere, typical of a recreational setting. The caption "run run run" adds to the sense of dynamic action and thrill.

-- GPT-4o

Figure 2: The model failed to identify humans and generate incorrect responses

Additionally, we found that textual cues alone can sometimes be sufficient for accurate classification, as demonstrated by DeepSeek’s competitive performance despite its lack of image processing. However, models with multimodal capabilities (ChatGPT and Grok-2) demonstrated superior performance, leveraging both facial expressions and environmental context to make better-informed predictions. That said, the reliance on images occasionally led to errors, particularly when the environmental context contradicted the expected verbal and facial cues, even though the model was not explicitly prompted to consider it. Finally, we observed that emotions displayed in acted scenarios were easier to classify than those in natural settings, where expressions are often more complex and nuanced. This highlights a potential limitation in emotion classification models trained predominantly on clear, exaggerated expressions rather than subtle, real-world emotional cues.

Despite these insights, our study has several limitations that should be addressed in future research. First, the dataset of videos was relatively small, limiting our ability to fully explore these challenges across diverse emotional expressions and demographics. Expanding the dataset would allow for a more comprehensive evaluation of GenAI’s performance across different contexts. Second, the dataset suffers from class imbalance, with certain emotions being underrepresented. More balanced data collection efforts will be necessary, requiring additional manpower to ensure equal representation of different emotional states across both adults and children.

Another challenge lies in the inherent difficulty of perfectly matching human-labeled emotion labels with AI predictions. Emotions are complex and subjective, and some mismatches may result from discrepancies between the labels and the true emotional cues present in the videos. Moreover, it was particularly difficult to find video clips where children exhibited conflicting verbal and non-verbal cues, further complicating the evaluation process. We also noted that disgust was especially difficult for the models to classify, likely due to its subtle and context-dependent expressions.

Finally, a significant limitation in classifying children’s emotions stems from the limited availability of verbal cues in videos featuring younger subjects. Since children have lower language proficiency, their expressions are often more reliant on facial and contextual cues. This likely explains the lower accuracy observed for children, as verbal input—typically an important source of emotion classification—was frequently insufficient or absent.

References

- [1] Robert Goodman, Tamsin Ford, Helen Simmons, Rebecca Gatward, and Howard Meltzer. Using the strengths and difficulties questionnaire (sdq) to screen for child psychiatric disorders in a community sample. *The British Journal of Psychiatry*, 177(6):534–539, 2017.
- [2] Tobias Grossmann. The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience*, 28(2):219–236, 2010.
- [3] Catherine M. Herba, Sabine Landau, Tamara Russell, Christine Ecker, and Mary L. Phillips. The development of emotion-processing in children: Effects of age, emotion, and intensity. *Journal of Child Psychology and Psychiatry*, 47(11):1098–1106, 2006.
- [4] Carroll E. Izard, Elizabeth M. Woodburn, Kristy J. Finlon, E. Stephanie Krauthamer-Ewing, Stacey R. Grossman, and Adina Seidenfeld. Emotion knowledge, emotion utilization, and emotion regulation. *Emotion Review*, 3(1):44–52, 2011.
- [5] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, 5:1532, 2015.