Evaluating Human-LLM Representation Alignment: A Case Study on Affective Sentence Generation for Augmentative and Alternative Communication

Shadab Choudhury[1], Asha Kumar[2], Lara J. Martin[1]

[1] Computer Science and Electrical Engineering Department
[2] Information Systems Department

{shadabc1, laramar}@ umbc.edu

Motivation: AAC and LLMs

Augmented and Alternative Communication (AAC) are software or tools used by people who cannot communicate verbally.

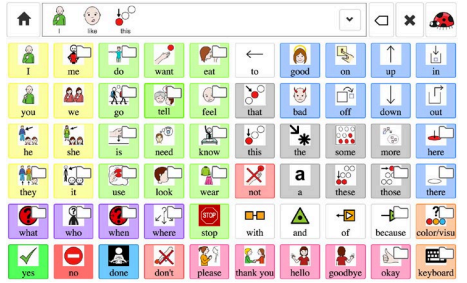
Examples of existing tools are CoughDrop, MyVoice, AlekAssist, Tobii Dynavox, etc. Each is designed for a different group of AAC users.

AAC tools struggle on two main fronts:

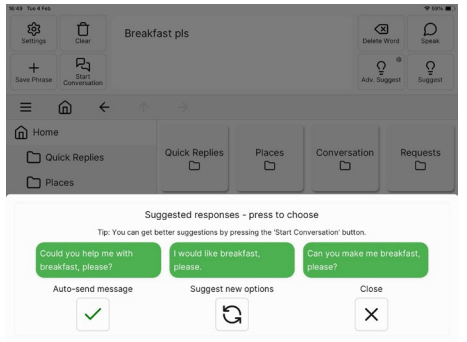
- Input speed, and
- Personalization,

And Keyword-Based Generation plus LLMs can help resolve both of them.

But- LLMs tend to ‘overwrite’ the user’s voice.



CoughDrop



AlekAssist

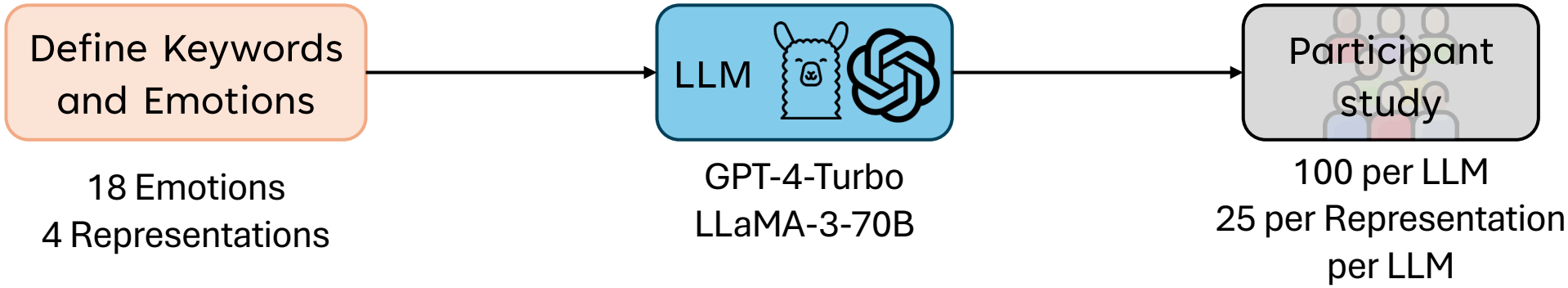
Research Questions

1. **Representation Alignment:**

   Do LLMs' use of emotion representations match humans' expectations?

2. **Accuracy and Realism:**

   Is there a preferred representation for conveying emotions when performing keyword-based sentence generation?



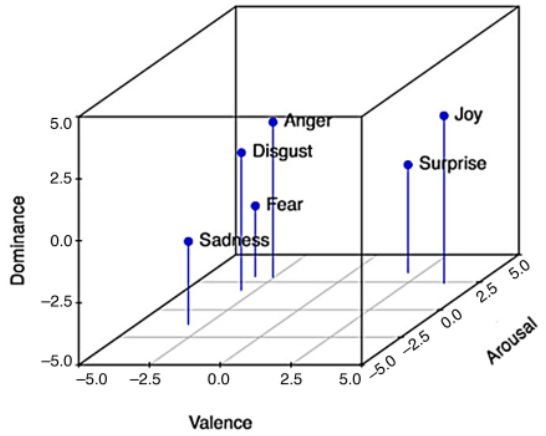
Representing Emotions

Words – English terms for the emotion

Lexical VAD – VAD scales expressed in English (Very High, High, Moderate, Low, Very Low)

Numeric VAD – VAD scales expressed in numeric terms (-5.0 to +5.0 in increments of 0.5)

Emojis



Grateful	😄	Very High Valence, Moderate Arousal, Low Dominance	V: +2.5, A: 0.0, D: -2.5
Furious	🤬	Very Low Valence, Very High Arousal, High Dominance	V: -4.0, A: +4.0, D: +1.0
Sad	😟	Very Low Valence, Low Arousal, Very Low Dominance	V: -4.0, A: -2.5, D: -4.0

Generating Sentences

Emotions Used – Grateful, Joyful, Content, Surprised, Excited, Impressed, Proud, Anxious, Afraid, Terrified, Annoyed, Angry, Furious, Sad, Devastated, Ashamed, Embarrassed, Guilty

3 Keywords per Sentence – e.g., [“Place”, “Great”, “Korean”], or [“Semester”, “Finals”, “Math”]

For **Words** and **Emojis** – Few-Shot Prompting

For **Lexical VAD** and **Numeric VAD** – Chain-of-Thought + Few-Shot Prompting

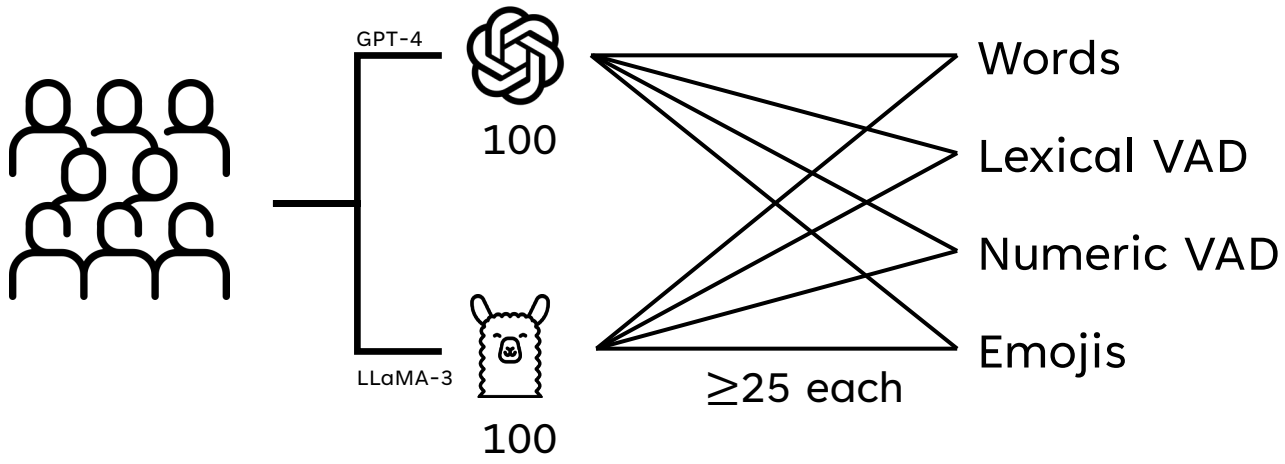
* The prompts are fixed. In AAC applications, users would only enter keywords and the emotion.

Participant Survey

Participants were recruited on Prolific, required to be fluent in English, 18 years or older, and residing in the United States. Each participant was paid at a rate of $14/hour for completing the survey.

Each participant was ‘assigned’ one representation. All ‘emotions’ shown to them in the survey were in that representation only

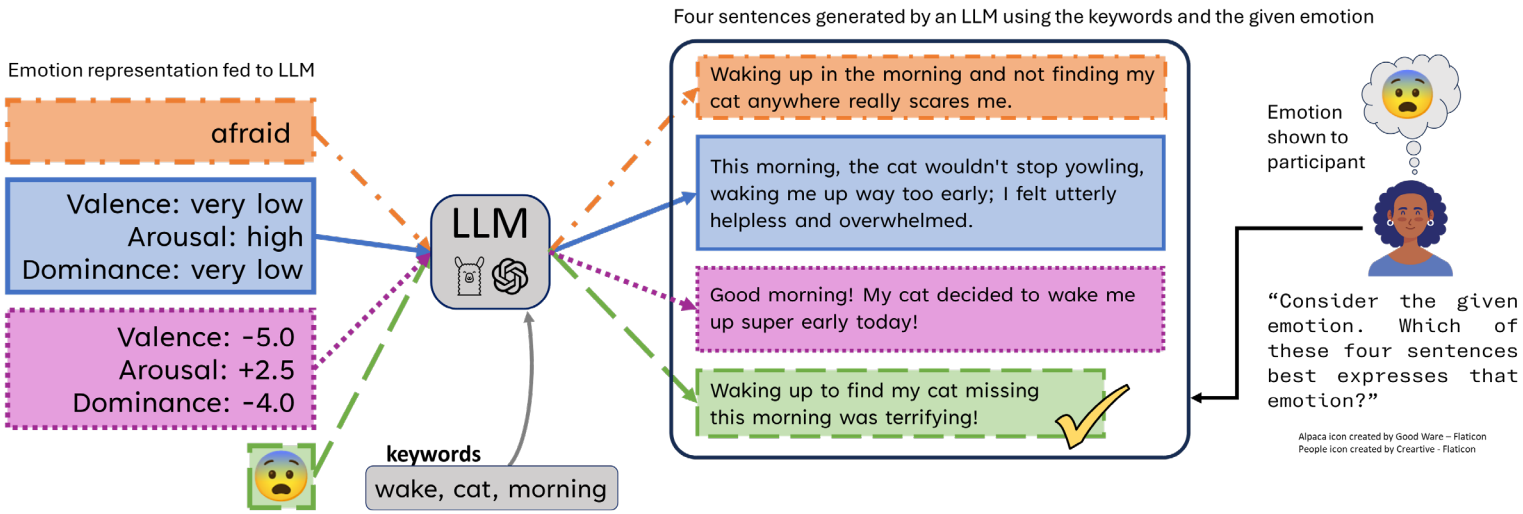
Participants were given a short tutorial on how to read Valence-Arousal-Dominance scales.



RQ1: Representation Alignment

Determining which representation was the best at conveying the emotion to the user and the LLM *as the user understood it*. Each participant was given 10 questions of this type.

Rep_A is the representation shown to the participant in the question. Rep_B is the representation used by the LLM to generate the sentences. Participants see Rep_A but are unaware of Rep_B. They simply pick the best matching sentence. In the example below, Rep_A is **Emojis**.

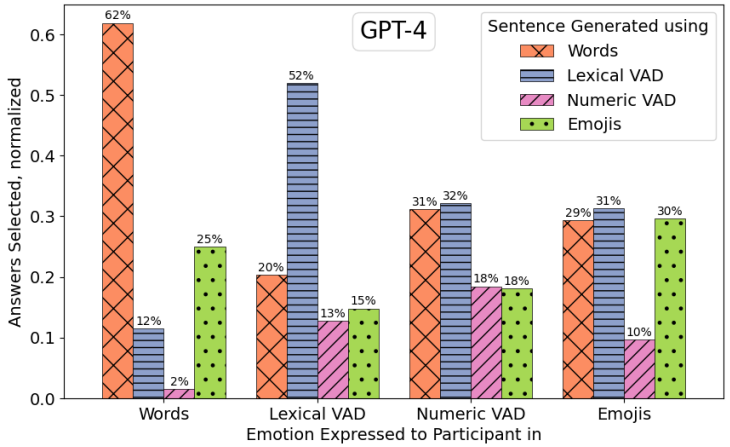


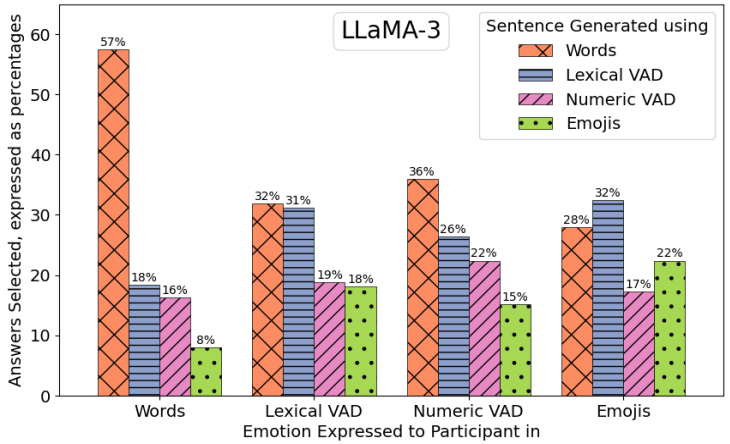
Findings: Representation Alignment

We consider high alignment between two representations when participants in Rep_A select the Rep_B sentence frequently. We also note 'Self-Alignment' when Rep_A and Rep_B are the same.

Words come out as having the best Representation Alignment, while **Lexical VAD** comes second.

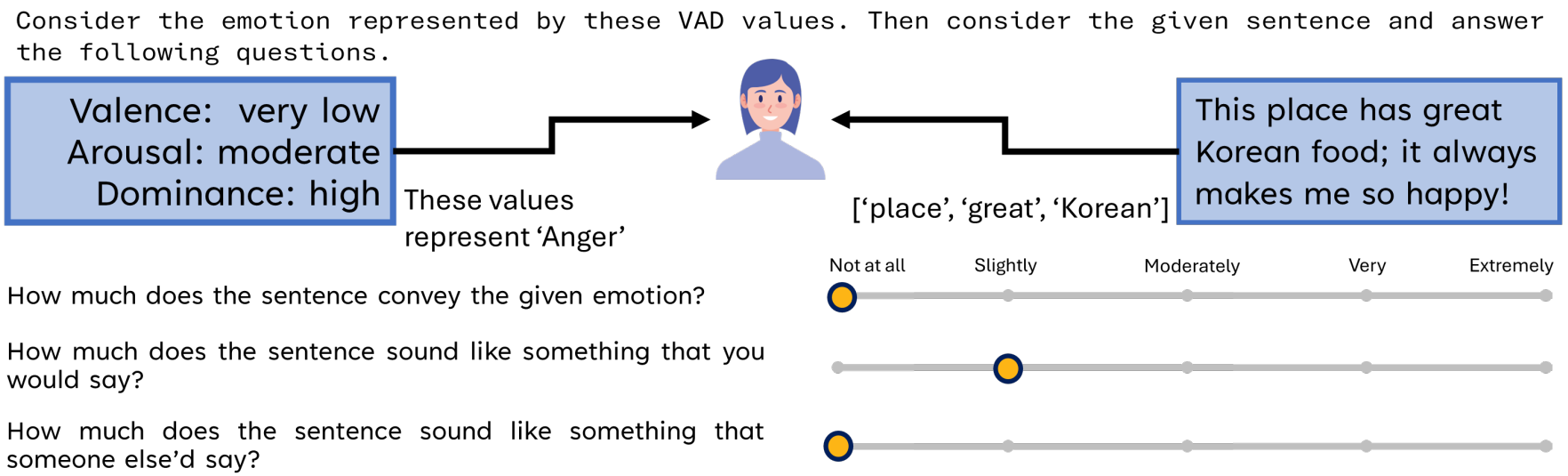
Participant's Representation	Entropy↓ GPT-4	LLaMA-3
Words	**.32**	**.42**
Lexical VAD	.61	.72
Numeric VAD	.70	.63
Emojis	.67	.52





RQ2: Accuracy and Realism

Determining which representation outputs the most realistic sentences. Each participant was given 5 questions of this type. Each question was answered by ~4 participants on average.

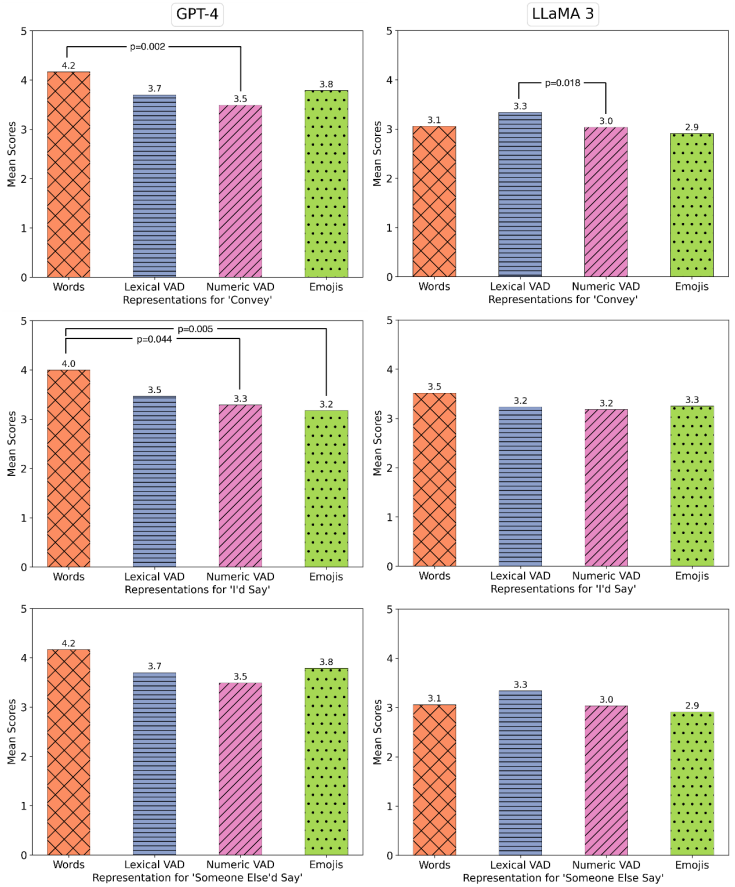


Findings: Accuracy and Realism

ANOVA and pairwise t-tests on the Likert results show the following:

- **Words** was significantly better at “Convey” than **Numeric VAD** for GPT-4 (p = 0.002)
- **Words** is significantly better than both **Emojis** (p = 0.005) and Numeric VAD (p = 0.044)
- **Lexical VAD** significantly better at “Convey” than **Numeric VAD** for LLaMA-3 (p = 0.018)

These results, plus the general higher rating of **Words** and **Lexical VAD**, show these two are the best option for realistic outputs.



Takeaways

We've shown:

- a human evaluation method for measuring the alignment between mental models of concepts (such as emotions) and how they are used by LLMs.
- We show that humans and LLMs align best when **Words**, or to a lesser extent, **Lexical VAD** are used to represent emotions, and that these two also give the most realistic outputs.

Downstream, these can be used to:

- Evaluate other concepts. Anything that is represented in distinct ways by people could be evaluated in this way.
- Improve the speed and precision of inputting emotions into text generation tools (especially if models are further optimized for VAD).

Paper Link