LARA Lab
Language, Aid, and Representation AI Lab
@ UMBC

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

Shadab Choudhury, Asha Kumar, Lara J. Martin
{shadabc1,laramar}@umbc.edu

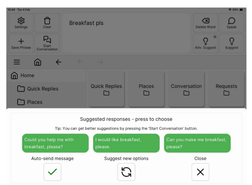
arXiv Link

Augmentative and Alternative Communication (AAC)

Tools or software to help supplement or replace speech for people who cannot communicate verbally. Two examples of other already-existing AAC software are given on the right.

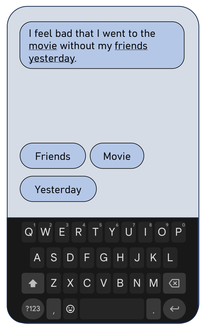
Can we improve AAC software using LLM text generation?


CoughDrop


https://alekassist.com/

Keyword-Based Generation

Get a full sentence from a few keywords. Balances input speed with personalization for AAC applications. Also called lexically-constrained generation.



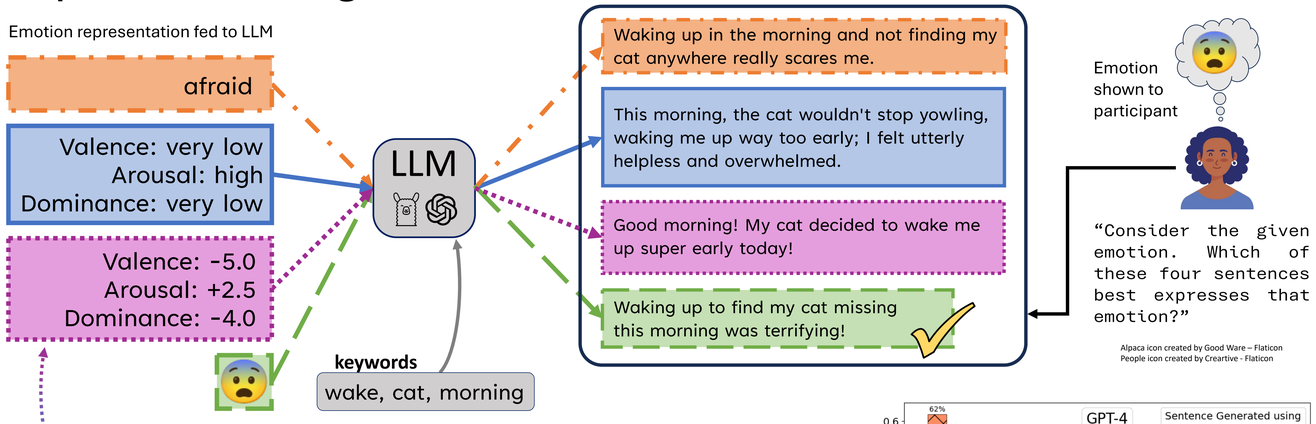
Designing **AAC software** or other Human-AI Collab tools?

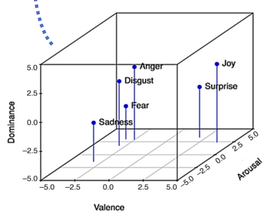
Words to describe emotions are the best option for alignment and quality. If you need finer, less ambiguous outputs, use VAD, but quantify them in words (low, moderate, high, etc).

Deciding how to **represent a concept** for Human-LLM Collab?

Use our representation alignment paradigm! Give a user one representation, then ask them to pick from multiple LLM outputs, each using a different representation, then check representation alignment.

Representation Alignment





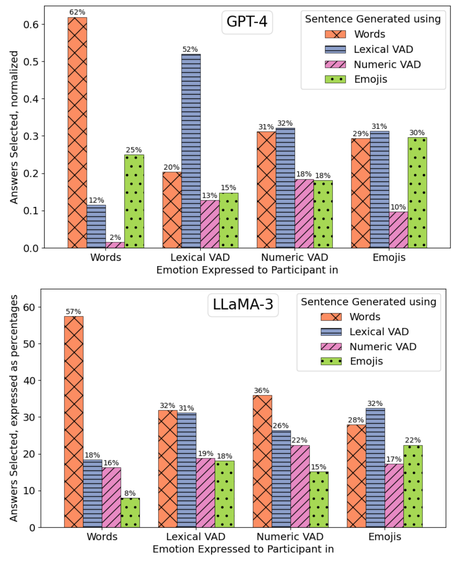
Valence – Positive (High) or Negative (Low) feelings.

Arousal – How much energy's behind the feeling.

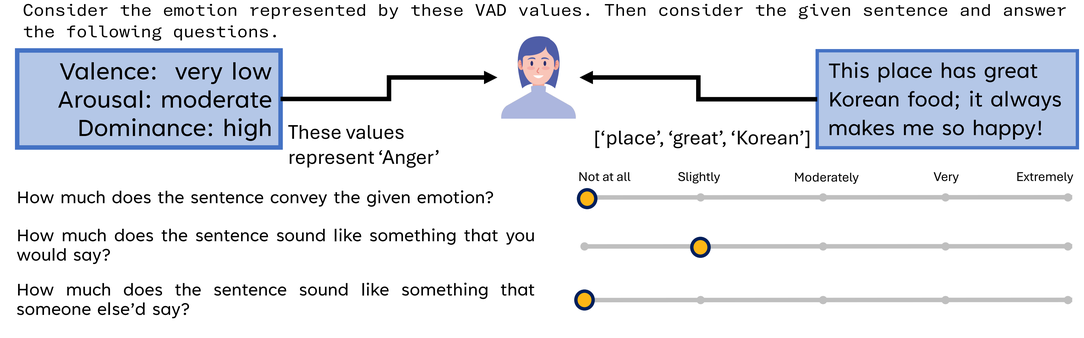
Dominance – How sure and concrete is this feeling?

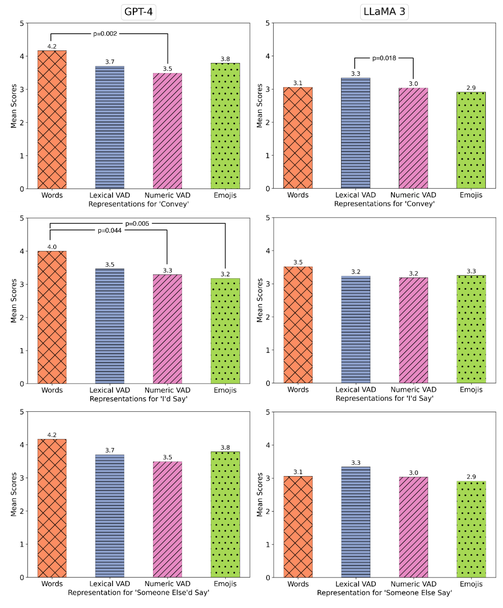
Words had the best Representation Alignment through self-alignment. When participants were shown an emotion in Words, they picked the sentence generated using Words.

Lexical VAD had the second-best Representation Alignment. People were uncertain of the best sentence when the emotion was given in Numeric VAD. Emojis also had poor alignment, even if most people are used to using emojis daily.



Accuracy and Realism





Lines generated using Words were the most accurate and realistic. Lines generated using Lexical VAD came second.

Both Numeric VAD and Emojis gave much poorer results overall.