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
University of Maryland, Baltimore County
{shadabc1,laramar}@umbc.edu

Abstract

Gaps arise between a language model's use of concepts and people's expectations. This gap is critical when LLMs generate text to help people communicate via Augmentative and Alternative Communication (AAC) tools. In this work, we introduce the evaluation task of Representation Alignment for measuring this gap via human judgment. In our study, we expand keywords and emotion representations into full sentences. We select four emotion representations: Words, Valence-Arousal-Dominance (VAD) dimensions expressed in both Lexical and Numeric forms, and Emojis. In addition to Representation Alignment, we also measure people's judgments of the accuracy and realism of the generated sentences. While representations like VAD break emotions into easy-to-compute components, our findings show that people agree more with how LLMs generate when conditioned on English words (e.g., "angry") rather than VAD scales. This difference is especially visible when comparing Numeric VAD to words. Furthermore, we found that the perception of how much a generated sentence conveys an emotion is dependent on both the representation type and which emotion it is.

1 Introduction

Augmentative and Alternative Communication (AAC) tools help people who cannot communicate verbally to hold conversations. Speed of communication is one of the biggest pain points that users point out about their AAC tools (Trnka et al., 2007). As a result of this lag, users are often talked over, ignored, or otherwise disrespected in conversations (Kane et al., 2017). Fortunately, due to their medium, high-tech AAC options such as phone or tablet applications present the opportunity to incorporate NLP techniques to improve communication speed.

To work toward speeding up AAC use, keyword-based sentence generation and word prediction

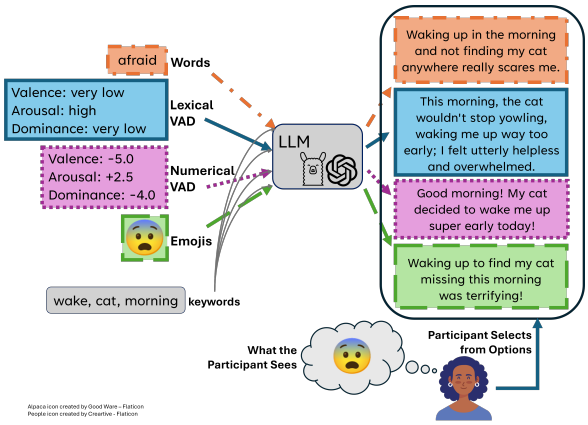

Figure 1: Representation Alignment experiment. Three keywords and an emotion from one of the four representations are used to generate a sentence. Participants are shown the emotion in only one of the representations and select the sentence that best fits that emotion.

have been explored in AAC technologies for decades (Trnka et al., 2007; Wiegand and Patel, 2012; Garcia et al., 2014; Fontana de Vargas and Moffatt, 2021; Shen et al., 2022). Unsurprisingly, we have also seen a recent uptick in the use of LLMs in AAC tools (Valencia et al., 2023; Kitayama and Hirotomi, 2024; Francis et al., 2024; Bailey et al., 2024).

There have been fairly successful attempts for personalized generation (Zhang et al., 2023), such as through prompt engineering, keyword-based generation (Hokamp and Liu, 2017; Yao et al., 2019) or style transfer (Liu et al., 2024). While LLMs can generate human-like text better than their predecessors, there are concerns about putting them into AAC applications. Namely, how much control users have over the LLM, whether the LLM accurately captures the user's way of speaking, or if the LLM is generating text that is appropriate to the conversation's context (Valencia et al., 2023; Martin and Nagalakshmi, 2024). Part of this context includes the current emotional state of the AAC

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1618–1637
December 20-24, 2025

user. As such, we wanted to see how well LLMs and humans aligned in their understanding of various emotion representations. Our hope is that representations used in future AAC applications will generate sentences that match the user's intended emotion.

In this paper, we introduce the task of REPRESENTATION ALIGNMENT, illustrated in Figure 1, where the alignment between the LLM's output and the human's mental model of a concept is measured. Note that this is different from Theory of Mind, as the LLM is not trying to ascertain the current mental state of the user. Representation Alignment looks at how well an LLM's use of concepts aligns with human understanding. Specifically, we will be focusing on the Representation Alignment of *emotions*.

There are multiple ways an emotion could be represented in the input. One representation could be easier for an LLM to parse, but a different representation might be more accessible to users. Hence, to assess emotion representations in the context of keyword-based sentence generation, we investigate the following research questions:

1. Do LLMs' use of emotion representations match humans' expectations?
2. Is there a preferred representation for conveying emotions when performing keyword-based sentence generation?

We look at four ways of representing emotions: with Words (the English word for a particular emotion), via emojis – which have become an effective way to express emotions visually over text (Kaye and Schweiger, 2023), and two types of Valence-Arousal-Dominance scales. We will explain the representation implementation in Section 3.

Valence-Arousal-Dominance (VAD) (Mehrabian, 1980) scales measure emotions on three axes: Valence (or pleasure) indicating whether its a positive or negative feeling, Arousal indicating how much energy is behind the feeling, and Dominance indicating how much control the user has over that feeling. Although VAD has origins in psychology research, it has been useful for NLP-related areas such as affective computing (Mohammad, 2018; Guo and Choi, 2021; El-Haj and Takanami, 2023) and social computing (Hutto and Gilbert, 2014; Khosla, 2018; Garg, 2023). Thus, we were curious about the effectiveness VAD for text generation.

In this work, we addressed the research questions by generating sentences using each of the four emotion representations by few-shot prompting two large language models (GPT-4 and LLaMA-3). We ran a human participant study to determine how well the representations aligned to participants' expectations and how accurately & realistically LLMs can generate sentences using a particular representation. Our contributions are as follows:

1. We introduce a human evaluation paradigm for measuring the alignment between mental models of concepts (such as emotions) and how they are used by LLMs.
2. We show that humans and LLMs align best when Words, or to a lesser extent, Lexical VAD are used to represent emotions.
3. We also show that either Words or Lexical VAD, when used in a prompt, give realistic sentences that most participants agree sound like something they would say.

2 Related Work

2.1 Keyword-based Sentence Generation

Keyword-based sentence generation, also referred to as *lexically-constrained generation*[1], is a subtopic of controlled text generation where the input to a model is a set of keywords and the output is a sentence that uses those keywords. Early on, Kasper (1989); Uchimoto et al. (2002) used grammar systems to build sentences using keywords as anchors. More recently, Mou et al. (2016); He and Li (2021) used sequential models and revised the output repeatedly at each sampling step. Yao et al. (2019); Wang et al. (2020); Ammanabrolu et al. (2020) similarly generated sequential outputs by prompting with a list of keywords. With transformer-based language models, it's also possible to use the decoder or special tokens to control keywords, style or sentiment (Kumar et al., 2021; Samanta et al., 2020; Nie et al., 2023; Krause et al., 2021; Sasazawa et al., 2023).

However, most of these works focus on smaller language models. Chen and Wan (2024) showed that LLMs have strong baseline lexically-constrained generation ability. To our knowledge, not much other work has been done in this area using LLMs.

2.2 Emotion-conditioned Sentence Generation

Emotional sentence generation has been widely studied as part of style-transfer or empathetic

[1]We use the two terms interchangeably.

dialogue generation problems. Early work on emotional sentence generation, such as Polzin and Waibel (2000), relied on rule-based systems. Ghosh et al. (2017); Song et al. (2019) both conditioned the output of a recurrent neural network on specific emotional words, while Singh et al. (2020) also conditioned on words but used GPT-2 (Radford et al., 2019) as the base model. Colombo et al. (2019); Lubis et al. (2018) utilized the VAD space, adding an emotional vector to the internal representation of the text. Zhou and Wang (2018) conditioned a variational autoencoder on Emojis instead.

LLMs have also been used for emotional text generation. A variety of methods like chain-of-thought, retrieval-augmented generation, prompt tuning, etc. have all been successfully used (Li et al., 2024; Mishra et al., 2023; Rasool et al., 2025; Yang et al., 2024; Resendiz and Klinger, 2025; Zhang et al., 2024). However none of these works deal with non-emotional constraints, nor cover the use of VAD scales.

2.3 Value Alignment

There have been multiple methods for integrating values into AI such as learning normative behavior from children's stories (Nahian et al., 2020), integrating logic (Kim et al., 2021), using actor-critic models(Liu et al., 2022), integrating situated annotations in reinforcement learning from human feedback (Arzberger et al., 2024), or matching behavior to underlying ethics rather than human actions (Rigley et al., 2025). By finding and comparing explicit representations for ethical values, these methods could potentially be evaluated using our Representation Alignment evaluation technique. There have been recent efforts to measure the value alignment of existing models (Norhashim and Hahn, 2024), but only for a binary concept (moral or not moral).

3 Sentence Generation

We prompted LLMs to generate sentences using three words to denote the content of the sentence—which we will refer to as the *keywords*, in addition to the emotion we wanted the sentence to express.

We selected four representations of 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), and
- *Emojis*[2].

The numeric values for VAD were generated by normalizing values from Guo and Choi (2021) (which were in the range of 0.00 to 1.00) to a -5.0 to +5.0 scale, then rounding to the nearest 0.5. We also recognize that there will be loss from converting Guo and Choi's scale to a *discrete* numeric representation. However, we found it to lead to easier comprehension. Before the rounding step, the Numeric VAD values were converted to Lexical VAD. Lexical VAD was mapped to a 5-point scale for simplicity. Due to the rounding step, in some cases there may be an overlap (such as Surprise's +2.5 Arousal and Fear's +2.5 Arousal corresponding to 'Very High' and 'High' respectively). However, this does not pose an issue as Lexical VAD and Numeric VAD are independent representations even though they both capture VAD. We hypothesize different strengths from each representation. Lexical VAD may carry extra semantic information from being linguistically derived but Numeric VAD denotes more meaningful separation between points by placing them on an interval scale.

When the LLMs were prompted to generate sentences using the Words representation, we explicitly forbid the models from treating the Emotion as a keyword[3], as LLMs have a tendency to "tell" instead of "show" if not provided enough guidance. This was also not to bias the participants in the Words condition into selecting sentences generated using the Words representation simply because it was using the same word.

We limited each input to three content keywords to give LLMs sufficient context to generate. The keywords were sets of arbitrarily-chosen, common words like [`Place`, `Great`, `Korean`], [`Finals`, `Semester`, `Math`]. We generated sentences using GPT-4-Turbo-2024-04-09 (OpenAI, 2024) and LLaMA-3.3-70B (Grattafiori et al., 2024), referred to as GPT-4 and LLaMA-3 respectively in this paper. Both models were used with default parameters. GPT-4-Turbo cost less than $5 in total to generate the sentences. All generations were con-

[2]Emojis were embedded as unicode so that participants would see the set that they were used to seeing on their device, although we recognize this introduces some disparity across participants.

[3]"Do not use the word '{emotion}' in the response and express the sentiment in a different way." was added to the prompt. The full prompts can be found in Appendix B.

Category	Words	Emoji	Valence	Arousal	Dominance
Happy	Grateful	😊	Very High (+2.5)	Moderate (0.0)	Low (-2.5)
	Joyful	😂	Very High (+4.0)	High (+1.0)	High (+1.0)
	Content	😌	Very High (+4.0)	Moderate (0.0)	Very High (+4.0)
Surprise	Surprised	😯	High (+1.0)	Very High (+2.5)	Low (-2.5)
	Excited	😃	Very High (+2.5)	Very High (+4.0)	High (+1.0)
Pride	Impressed	🙂	High (+1.0)	High (+1.0)	Very Low (-4.0)
	Proud	🤩	Very High (+4.0)	High (+1.0)	Very High (+2.5)
Fear	Anxious	😟	Low (-1.0)	High (+2.5)	Low (-2.5)
	Afraid	😰	Very Low (-5.0)	High (+2.5)	Very Low (-4.0)
	Terrified	😱	Very Low (-5.0)	Very High (+4.0)	Very Low (-4.0)
Anger	Annoyed	😒	Low (-2.5)	Moderate (0.0)	Moderate (-1.0)
	Angry	😡	Very Low (-5.0)	High (+2.5)	Moderate (0.0)
	Furious	🤬	Very Low (-4.0)	Very High (+4.0)	High (+1.0)
Sadness	Sad	😕	Very Low (-4.0)	Low (-2.5)	Very Low (-4.0)
	Devastated	😭	Very Low (-4.0)	High (+1.0)	Low (-2.5)
Shame	Ashamed	😳	Low (-3.0)	Moderate (-1.0)	Very Low (-4.0)
	Embarrassed	😓	Very Low (-4.0)	High (+2.5)	Low (-2.5)
	Guilty	😬	Very Low (-4.0)	Moderate (0.0)	Very Low (-4.0)

Table 1: All 18 emotions used in the study and how they were represented in words, emojis, and Valence-Arousal-Dominance (VAD) scores with values represented lexically and numerically (in parentheses). The category was used to determine which emotions to use but were not integrated into the prompt nor shown to participants.

sistent with the models' indended use.

Due to compute limitations, we used a LLaMA-3 model that was quantized (Huang et al., 2024) using 8-bit weights[4]. Computing resources for LLaMA-3's generation can be found in Appendix A. The prompt, used for both models, can be found in Appendix B. In the system prompt, we instructed the model to minimize inserting extraneous information while still generating sentences that clearly express an emotion. Although prior work has shown that LLMs are particularly good at generating sentences from keywords (Chen and Wan, 2024), we did verify this as well (results in Appendix E).

For each LLM, we generated a total of 360 sentences, 90 per representation, expressing 18 different emotions. We determined this list of emotions based off Demszky et al. (2020), ultimately selecting for emotions that were distinct but could also be easily grouped. Two sentences were selected randomly for each of the 18 emotions for use in the evaluation, resulting in 36 Representation Alignment questions per condition (i.e., the emotion representation that the participant saw). Table 1 shows all the emotions used, their grouping, and their VAD values.

Overall, 72 sentences for each representation were selected and shown to participants; half for the Representation Alignment questions (Section 4.1), and the other half for Accuracy and Realism questions (Section 4.2). We did this in order to ensure each question had answers from multiple participants.

For generating a sentence with emotions represented as Words or Emojis, we used plain few-shot prompting (Brown et al., 2020), selecting exemplars from across the range of positive and negative emotions. For prompting using either form of VAD, we used step-back chain-of-thought prompting (Zheng et al., 2024). First, we prompted the model to give an explanation of VAD, then convert it to scale from -5 to +5 if using Numeric VAD, and then gave it few-shot examples. For all representations, we converted the emotions in the exemplars to the representation we intend to generate in.

4 Evaluation

The generated sentences were evaluated via crowdsourcing. We recruited 100 participants on Prolific for each model for a total of 200 participants, pay-

[4] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct#use-with-bitsandbytes

ing them $14/hr for 15 minutes ($3.50 each) — for a total of $700. To participate, people were required to agree to the consent form approved by the university's IRB. Participants needed to be 18+ years of age, living in the United States, and fluent English speakers. Prolific automatically assigns random identifiers to participants and no identifiable information was collected from participants. Each participant was randomly assigned one of the four conditions (the four emotion representations) which was the emotion representation that they saw for all questions.

Since we assumed participants would not be used to VAD, people in either VAD condition were shown a short explanation of VAD and given questions with feedback to train them on how to use it (Appendix C). Participants were shown two sets of questions. The first set (Representation Alignment) was used to determine if the representation that the human saw aligned with the LLM's representation. The second set (Accuracy & Realism) asked the participant to consider how accurately a generated sentence captured a given emotion and how realistic the generated sentence appeared. Both the emotion and sentence from this second set of questions used the same representation. That is, the LLM was given the same representation to generate the sentence as what the user was seeing for the emotion. We found this setup compelling since the representation that produces the best output from the LLM does not have to perfectly align with the user's internal representation but rather how the user imagines that representation to be expressed. For example, if people believe that Lexical-VAD–prompted LLM output produces sentences that match their interpretation for how a certain Words representation would be expressed, then that is just as valuable information.

A final open-ended question was asked at the end of the survey as an attention check, asking participants to name a piece of media that they have consumed recently and to describe the emotion they felt watching/reading it. Participants were blocked from copying and pasting text. Any responses that were written in poor English or did not answer the question were removed, and we replaced their data with new responses. We ended up with 26, 25, 28 and 29 participants for Words, Lexical VAD, Numeric VAD and Emojis respectively for GPT-4, and 25 participants for each of the four conditions for LLaMA-3. Due to this difference, we normalized the counts during evaluation and discussion. The data was analyzed in Python using the packages pandas & scipy and visualized with matplotlib[5].

4.1 Human Evaluation 1: Representation Alignment

To evaluate Representation Alignment, we provided an emotion in the representation the person was assigned and showed them sentences generated across all four conditions. The participant was then told to select the sentence that is the best fit for the emotion, as illustrated in Figure 1. Here, the goal was to determine which representation was most effective at conveying the emotion to the user *as the user understood it* across multiple emotions.

For example, someone in the Words condition would see an emotion ("Anxious") followed by four sentences. Each of these sentences were generated using a different representation for the emotion "Anxious". The sentence order was shuffled before being presented to the participant. In the example below, 1 was generated using an English Word for the emotion, 2 with Lexical VAD, 3 with Numeric VAD, and 4 with an Emoji.

Anxious

1. I feel so nervous about my math finals this semester.
2. I can't believe the semester is almost over, and we've got that big math final coming up soon; it's really time to buckle down and study hard!
3. I'm really stressed about the math finals this semester.
4. I'm so happy I passed my math finals this semester!

Participants were required to answer all questions and were shown 10 of these questions randomly selected from the 36 questions within the participant's condition. The questions were evenly distributed using Qualtrics's "evenly display questions" feature. These questions were each answered by 5-9 people (median of 7). Some questions were seen more or less often due to participants not qualifying and the counts not being reset in between participants. Full participant instructions can be found in Appendix D.

For a user-presented representation rep_A and an LLM-presented representation rep_B, we defined rep_A as having a strong REPRESENTATION

[5]Code for visualizations was generated with help from a mixture of ChatGPT, Claude, and Gemini. All other code was written by hand.

ALIGNMENT with rep_B if

- participants were most likely to pick rep_B (*selection rate*), and
- the average Shannon Entropy for selections rep_A was low, relative to other representations,

where rep_A and rep_B may be the same or different representations. For example, if participants who were shown Lexical VAD emotions were more likely to select sentences generated using Emojis (despite not knowing how the sentence was generated), *and* the entropy for Lexical VAD was relatively low, then Lexical VAD would have good Representation Alignment with Emojis. A random selection rate would be 25.0%, so any value above this would be notable. We will refer to the condition when $rep_A = rep_B$ as SELF-ALIGNMENT.

We use Shannon Entropy rather than inter-rater agreement because the latter assumes all options are equally likely to be chosen. However, in reality this is unlikely to be the case, and for some emotions or representations, one output may be consistently better than others.

4.1.1 Human Evaluation 1: Results and Discussion

Figure 2 shows the results of the selection rates across representation types and models, while Table 2 shows the mean entropy values. Overall, Words had the best Representation Alignment with Words, regardless of the model. This was unsurprising, as humans express emotions in Words most often and LLMs are trained on natural language data. Words had a 61.9% selection rate using GPT-4 and 57.5% using LLaMA-3, as well as Shannon Entropies of 0.32 and 0.42 respectively.

More surprisingly, Lexical VAD also has a high self-selection rate of 52.0% for GPT-4, with the second lowest entropy value at 0.61. While the agreement is worse than that of Words, the high selection rate is noteworthy. This occurs even though the participants and the LLMs are given different instructions and prompts. (Appendix B and C) This indicates both humans and LLMs may be drawing on similar ideas or memorized information when considering the emotion representation. On the other hand, LLaMA-3's output never matched people's expectations of Lexical VAD, regardless of what the model was prompted with, resulting in a higher entropy value.

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

Table 2: Shannon Entropy values showing the amount of variability in responses for each representation that was presented to participants. Bolded values show the most agreement, underlined are second most.

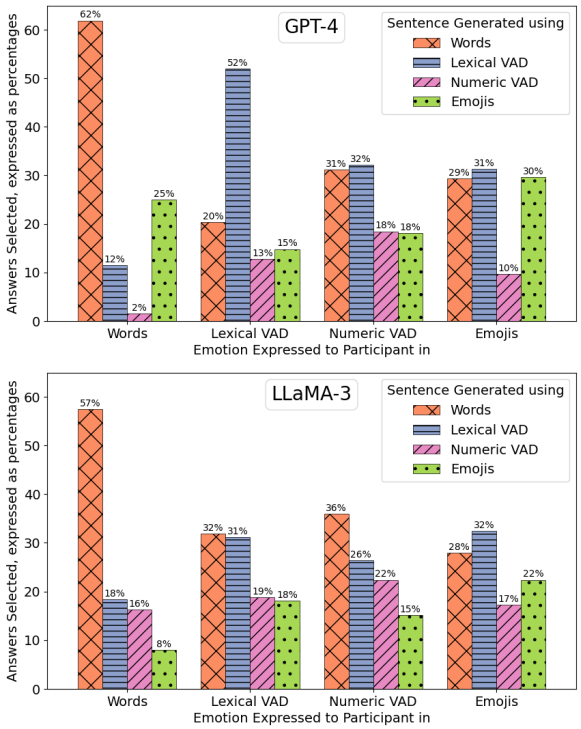

Figure 2: Percentage of times a sentence was selected. Each category on the x-axis corresponds to the condition the participant was in—what representation they saw. The colors delineate what representation was used for sentence generation. Results for GPT-4–generated sentences are on the top, LLaMA-3 on the bottom.

Numeric VAD, by contrast, had poor alignment, with the worst entropy values for both LLMs. We interpret this to mean that people could not consistently agree on an expected output for Numeric VAD emotions. One possibility is that participants struggled to conceptualize the numbers but were able to understand them more easily when using discrete words (leading to Lexical VAD doing better). This is prudential given some recent works offer users control over generative models using VAD (Tang et al., 2023; He et al., 2025).

Emojis, which have been shown to be subjective and ambiguous (Miller et al., 2017), did not

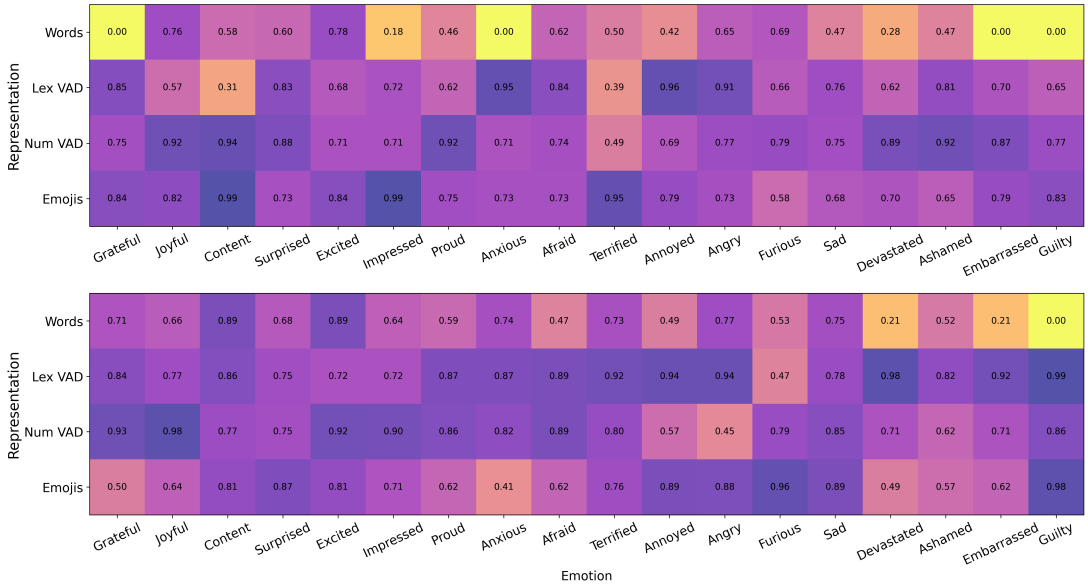

Figure 3: Heat maps for Shannon entropy of each emotion across representations. Lower (brighter) values are better, denoting more "agreement" between participants and the LLM. Top: GPT-4, bottom: LLaMA-3.

show particularly good alignment. However, Emojis did have some Representation Alignment with Lexical VAD for LLaMA-3 — 32.40% selection rate and .52 entropy. The trend for GPT-4 was less pronounced, with a selection rate of 31.38% and .67 entropy. Because there was a mismatch in representations it hints at that people's mental model of emojis may be similar to how LLaMA-3 works with Lexical VAD.

One potential explanation is that Lexical VAD for LLMs and Emojis for humans capture the same amount of information for imprecise emotions. In other words, LLMs can easily decompose Lexical VAD into emotions into components while Emojis are discrete symbols to text-based models like LLaMA-3. Meanwhile, humans can assign meaning to individual facial features in Emojis but may struggle to do the same with VAD scales. Further research will need to be done to verify this claim.

Ultimately, the highest Representation Alignment was Self-Alignment with Words and, for GPT-4, Lexical VAD Self-Alignment was a close second.

We also broke down Shannon Entropy values by emotion (see Figure 3). For individual emotions per representation, we found the results are fairly similar, with most values between 0.70 to 0.95 meaning poor agreement, with some outliers. The small gap in the models' agreement with participants for certain emotions point to potential differences in training. It is unclear why, for example, GPT-4 is much better at using Words to generate "grateful" sentences than LLaMA-3. To have emotions that people agree or disagree with for both models could mean a variety of things. For instance, is it that people agree more about whether a sentence is expressing "guilty", is there more "guilty" data both LLMs are trained, or does something inherent to the transformer architecture (such as attention) lend itself to picking up on "guilty" text better? A study on model architecture, training data, and conceptual alignment should be run to see if this trend continues.

4.2 Human Evaluation 2: Accuracy & Realism

The second set of questions asked the participant to consider how accurate or realistic a pair of an emotion and a sentence generated with that emotion is. The participants were asked to rate the three questions (shown in the example below) on a 5-point Likert slider, using the labels Not at all (1), Slightly (2), Moderately (3), Very (4), Extremely (5). For example:

> For the following questions, consider the emotion represented by these VAD values: **Very High Valence, Moderate Arousal, Low Dominance**
>
> And this sentence: **"This place has great Korean food; it always makes me so happy!"**

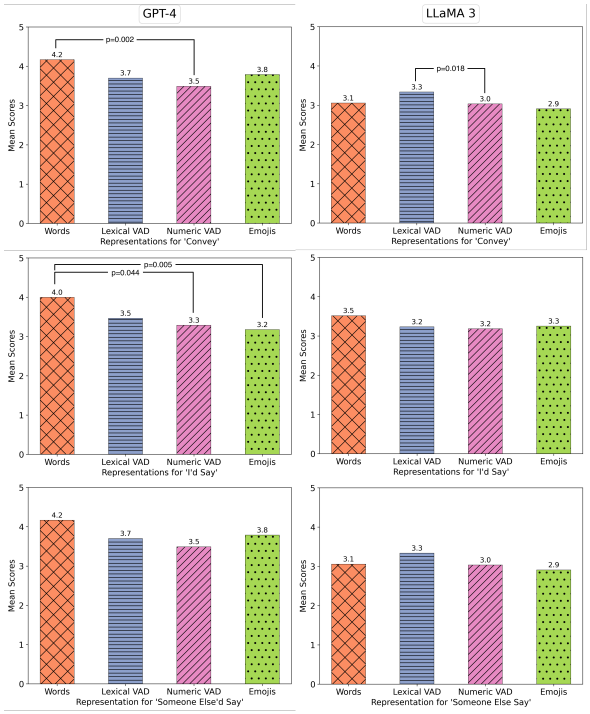

Figure 4: Left: GPT-4, Right: LLaMA-3. In order from Left to Right and Top to Bottom:
a, b. Histograms of the Mean Scores for 'Convey'
c, d. Histograms of the Mean Scores for 'You'd say'
e, f. Histograms of the Mean Scores for 'Someone Else'd Say'

How much does the sentence...
- Convey the emotion above?
- Sound like something that you would say?
- Sound like something that someone else would say?

We will refer to these three questions as "Convey", "You'd say", and "Someone Else'd say", respectively. The "Convey" question allows us to assess how well the generated sentence accurately expresses the emotion. The "You'd say" and "Someone Else'd say" questions are to measure how realistic and human-like the sentence is. The slider's default value for each question was Moderately (3). Participants were required to answer all questions.

Each user was given 6 questions of this type, randomly selected from the 36 total accuracy & realism questions (emotion-sentence pairings). Each question was answered by 3.7 people on average (1-8 participants each, median of 4), with 3 questions not answered and therefore not considered in the analysis.

4.2.1 Human Evaluation 2: Results and Discussion

The average Likert scores for "Convey", "You'd say", and "Someone Else'd say" across conditions (the emotion representations) and for both models can be found in Figure 4. We ran an ANOVA statistical significance test on the Likert ratings for the three questions for both GPT-4 and LLaMA-3.

The ANOVA showed that, for GPT-4, the emotion representation had a statistically significant effect on the rating for "Convey" and "I'd say", both $p < .01$. For LLaMA-3, ANOVA showed the emotion representation had a statistically significant effect on the rating for "Convey" and "Someone Else'd say", both $p < .05$. A pairwise t-test was run on the statistically significant results, which showed that Words was significantly better at "Convey" than Numeric VAD for GPT-4 ($p = 0.002$) and Lexical VAD was also significantly better at "Convey" than Numeric VAD for LLaMA-3 ($p = 0.018$) – further showing that Numeric VAD scores under-perform.

For "You'd say" questions, Words is significantly better than both Emojis ($p = 0.005$) and Numeric VAD ($p = 0.044$) when using GPT-4 to generate plausible-sounding sentences. This shows that Words would most likely be the preferred representation to capture an emotion appropriately. Additionally, LLaMA-3's generated sentences are overall perceived as slightly worse for conveying the emotion and sounding realistic. However, this is in line with the relative performances of each model.

We show mean scores for the "Convey" question broken down by emotion in Figure 5. The scores are relatively high across the board, with some outliers. For instance, GPT-4 using Words seemingly struggles most with sounding realistically "excited" or "proud", while using Numeric VAD makes GPT-4 struggle with emotions like "anxious" and "angry". LLaMA-3, on the other hand, sounds unrealistic when using "sad" or "excited" as Words.

Despite guardrails that make LLMs more amicable (Sharma et al., 2023; Zhou et al., 2024), we show that models can struggle to generate certain positive emotions under complex conditions, and that different emotions are captured best by different representations. Out of the representations we looked at, Words and Lexical VAD were the best for producing realistic sentences.

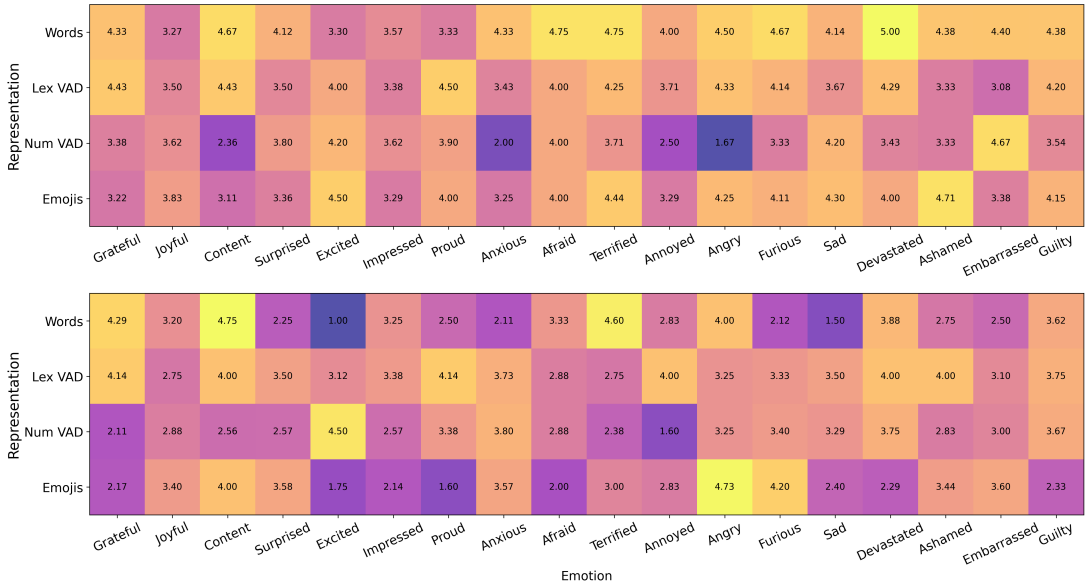

Figure 5: Mean "Convey" Scores for each emotion per representation. Higher (brighter) values are better. The top map shows results for GPT-4, while the bottom map is LLaMA-3.

5 Conclusion

We believe that research studying representations should measure the way they are perceived and represented to both humans and LLMs. This work introduced the evaluation paradigm of REPRESENTATION ALIGNMENT for determining if humans perceive representations in the same way that LLMs use them. Effectively evaluating how well human expectations for conceptual knowledge align with LLM output is a largely understudied problem, especially with downstream contexts in mind such as AAC. We adopted the problem of emotional sentence generation based on keywords and used Representation Alignment to study how various emotional representations align with participants' expectations. We hope that this experimental setup will be used to study other types of conceptual alignment as well, such as value alignment.

Using this paradigm, we showed that emotion-based Words provided the strongest alignment between an LLMs' understanding of an emotion & how it generates with it and humans' expectations of its use. Lexical VAD is a close second for both representation alignment and for accuracy and realism. In AAC applications, this can enable users to select their intended emotion or tone more precisely.

6 Limitations

Our study uses GPT-4-Turbo-2024-04-09 and LLaMA-3.3-70B with 8-bit weights due to compute constraints. It is possible that a more recent or bigger models such as GPT-4o or LLaMA-3-3.1-405B would give different results. Due to funding limitations, this study did not look at every state-of-the-art model available. However, we hope that this can serve as a proof-of-concept for other researchers who are interested in using our methodology.

This study was limited to English only. Other languages may give different outcomes. The participants were required to be native English speakers, which allowed us to control the quality of responses. However, it makes the results less generalizable in a global context. Our participants are also sampled from the general public and not necessarily AAC users. Assistive technologies need to be specialized to fit a person's needs, and therefore it is not enough to find overall Representation Alignment but also Representation Alignment for specific users. We believe that Representation Alignment will contribute to ease of use, but individual preferences and needs also matter.

Additionally, while we offer some training for interpreting VAD (whether lexical or numeric), it is still difficult to grasp without any prior knowledge and the heavy cognitive load needed to "calculate" these emotions may have affected the results.

Furthermore, while the emojis were selected based on what we believe was the best fit for each English word, there could be cultural discrepancies over what each emoji might mean. For example,

different online communities might use 😢 to mean embarrassed, worried, or relieved, depending on the context they see it used in and how their system displays it. More in-depth analyses of emoji use should gather general—like (Warriner et al., 2013) have done with Numeric VAD—and/or community-specific use to determine the exact meaning.

7 Ethical Considerations

LLMs can be used in many helpful cases, but they can also be used to impersonate individuals, or generate toxic or deceptive texts. We acknowledge that improving controlled text generation capabilities may also make it easier to generate the above, and that strong guardrails are necessary. In this work, we study how LLMs align with human emotions in terms of representations. This does not indicate that LLMs themselves are capable of possessing emotions, only that they are capable of recognizing and generating emotion when encoded in some textual form to the extent their model design and training data allows.

Having LLMs generate text for people can lead to questions of authorship. These issues are exacerbated when LLMs are introduce into AAC applications, potentially leading people to question the agency of AAC users and the validity of what they say. LLMs need to be carefully implemented into AAC applications with clear guidelines on how the AAC user should interface with the AI and provide potential scenarios where LLM use may not be useful or helpful. Additionally, many LLMs are too big to put onto tablets or phones and therefore require a network connection, which adds security risk.

Acknowledgments

We would like to thank all of our participants for their time and effort. We would also like to thank Foad Hamidi and others who have given us feedback throughout the study, and to the UMBC High Performance Computing Facility for helping us run our LLaMA experiments. This research is partially supported by a 2024 UMBC COEIT interdisciplinary proposal (CIP) Award.

References

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. Story Realization: Expanding Plot Events into Sentences. *AAAI Conference on Artificial Intelligence (AAAI)*, 34(5):7375—-7382.

Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. Nothing Comes Without Its World – Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):61–73.

Dallin J. Bailey, Francesca Herget, Derek Hansen, Forrest Burton, Grant Pitt, Tyson Harmon, and David Wingate. 2024. Generative AI applied to AAC for aphasia: a pilot study of Aphasia-GPT. *Aphasiology*, page 1–16.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xiang Chen and Xiaojun Wan. 2024. Evaluating, Understanding, and Improving Constrained Text Generation for Large Language Models.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*, pages 3734—-3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Mo El-Haj and Ryutaro Takanami. 2023. Unifying emotion analysis datasets using valence arousal dominance (VAD). In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 220–225, Vienna, Austria. NOVA CLUNL, Portugal.

Mauricio Fontana de Vargas and Karyn Moffatt. 2021. Automated generation of storytelling vocabulary from photographs for use in AAC. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1353–1364, Online. Association for Computational Linguistics.

Juliana Francis, Éva Székely, and Joakim Gustafson. 2024. ConnecTone: a modular AAC system prototype with contextual generative text prediction and style-adaptive conversational TTS. In *Interspeech*, page 1001–1002, Kos, Greece.

Luís Garcia, Luis de Oliveira, and David de Matos. 2014. Word and sentence prediction: Using the best of the two worlds to assist aac users. *Technology and Disability*, 26(2-3):79–91.

Muskan Garg. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. ACL.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs].

Yuting Guo and Jinho D. Choi. 2021. Enhancing Cognitive Models of Emotions with Representation Learning. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 141–148, Online. Association for Computational Linguistics.

Xingwei He and Victor O. K. Li. 2021. Show Me How To Revise: Improving Lexically Constrained Sentence Generation with XLNet. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12989–12997.

Yi He, Shengqi Dang, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. 2025. EmotiCrafter: Text-to-Emotional-Image Generation based on Valence-Arousal Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15218–15228.

Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume Volume 1: Long Papers, pages 1535–1546, Vancouver, Canada. ACL.

Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. An empirical study of LLaMA3 quantization: from LLMs to MLLMs. *Visual Intelligence*, 2(1):36.

C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *International AAAI Conference on Web and Social Media (ICWSM)*, 8(11):216–225.

Shaun K. Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. "At times avuncular and cantankerous, with the reflexes of a mongoose": Understanding Self-Expression through Augmentative and Alternative Communication Devices. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1166–1179, New York, NY, USA. Association for Computing Machinery.

Robert T. Kasper. 1989. A Flexible Interface for Linking Applications to Penman's Sentence Generator. In *Speech and Natural Language Workshop*, Philadelphia, Pennsylvania.

Linda K. Kaye and Christina R. Schweiger. 2023. Are emoji valid indicators of in-the-moment mood? *Computers in Human Behavior*, 148:107916.

Sopan Khosla. 2018. EmotionX-AR: CNN-DCNN autoencoder based emotion classifier. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 37–44, Melbourne, Australia. Association for Computational Linguistics.

Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *Journal of Artificial Intelligence Research*, 70:871–890.

Suzuna Kitayama and Tetsuya Hirotomi. 2024. An AAC Application for Generating Japanese Response Phrases Using GPT-4. In *Computers Helping People with Special Needs (ICCHP)*, volume 14751 of *LNCS*, page 144–152, Linz, Austria. Springer Nature Switzerland.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative Discriminator Guided Sequence Generation. In *Findings of the Association for Computational Linguistics: EMNLP*, page 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled Text Generation as Continuous Optimization with Multiple Constraints. In *Advances in Neural Information Processing Systems*, volume 34, pages 14542–14554. Curran Associates, Inc.

Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing Emotional Generation Capability of Large Language Models via Emotional Chain-of-Thought. ArXiv:2401.06836 [cs].

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024. Customizing Large Language Model Generation Style using Parameter-Efficient Finetuning. In *International Natural Language Generation Conference (INLG)*, pages 412–426, Tokyo, Japan. Association for Computational Linguistics.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):5293–5300.

Lara J. Martin and Malathy Nagalakshmi. 2024. Aging Up AAC: An Introspection on Augmentative and Alternative Communication Applications for Autistic Adults. ArXiv:2404.17730 [cs].

Albert Mehrabian. 1980. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Oelgeschlager, Gunn & Hain, Cambridge.

Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding Emoji Ambiguity in Context: The Role of Text in Emoji-Related Miscommunication. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):152–161.

Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI*, 10. Publisher: Frontiers.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358, Osaka, Japan. The COLING 2016 Organizing Committee.

Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. 2020. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 124–130, New York, NY, USA. Association for Computing Machinery.

Jinran Nie, Liner Yang, Yun Chen, Cunliang Kong, Junhui Zhu, and Erhong Yang. 2023. Lexical Complexity Controlled Sentence Generation for Language Learning. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 648–664, Harbin, China. Chinese Information Processing Society of China.

Hakim Norhashim and Jungpil Hahn. 2024. Measuring human-ai value alignment in large language models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1063–1073.

OpenAI. 2024. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

T S Polzin and A Waibel. 2000. Emotion-Sensitive Human-Computer Interfaces. In *ISCA Tutorial and Research Workshop (ITRW)*, pages 201–206.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Abdur Rasool, Muhammad Irfan Shahzad, Hafsa Aslam, Vincent Chan, and Muhammad Ali Arshad. 2025. Emotion-Aware Embedding Fusion in LLMs (Flan-T5, LLAMA 2, DeepSeek-R1, and ChatGPT 4) for Intelligent Response Generation. *AI*, 6(3):56.

Yarik Menchaca Resendiz and Roman Klinger. 2025. LLM-based Affective Text Generation Quality Based on Different Quantization Values. ArXiv:2501.19317 [cs].

Eryn Rigley, Adriane Chapman, Christine Evers, and Will McNeill. 2025. ME: Modelling Ethical Values for Value Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27608–27616.

Bidisha Samanta, Mohit Agarwal, and Niloy Ganguly. 2020. Fine-grained Sentiment Controlled Text Generation. ArXiv:2006.09891 [cs].

Yuichi Sasazawa, Terufumi Morishita, Hiroaki Ozaki, Osamu Imaichi, and Yasuhiro Sogawa. 2023. Controlling keywords and their positions in text generation. In *International Natural Language Generation Conference (INLG)*, pages 407–413, Prague, Czechia. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. ArXiv:2310.13548 [cs].

Junxiao Shen, Boyin Yang, John J Dudley, and Per Ola Kristensson. 2022. KWickChat: A Multi-Turn Dialogue System for AAC Using Context-Aware Sentence Generation by Bag-of-Keywords. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, pages 853–867, New York, NY, USA. Association for Computing Machinery.

Ishika Singh, Ahsan Barkati, Tushar Goswamy, and Ashutosh Modi. 2020. Adapting a Language Model for Controlled Affective Text Generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis. In *24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)*, pages 12–16, Dublin, Ireland.

Keith Trnka, Debra Yarrington, John McCaw, Kathleen F. McCoy, and Christopher Pennington. 2007. The Effects of Word Prediction on Communication Rate for AAC. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, volume Companion Volume, Short Papers, page 173–176, Rochester, New York. Association for Computational Linguistics.

Kiyotaka Uchimoto, Hitoshi Isahara, and Satoshi Sekine. 2002. Text generation from keywords. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, USA. Association for Computational Linguistics.

Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K. Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 830:1–830:14, New York, NY, USA. Association for Computing Machinery.

Lin Wang, Juntao Li, Rui Yan, and Dongyan Zhao. 2020. Plan-CVAE: A Planning-based Conditional Variational Autoencoder for Story Generation. In *China National Conference on Chinese Computational Linguistics (CCL)*, pages 892–902, Haikou, China. Chinese Information Processing Society of China.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavioral Research Methods*, 45(4):1191–1207.

Karl Wiegand and Rupal Patel. 2012. Non-Syntactic Word Prediction for AAC. In *Workshop on Speech and Language Processing for Assistive Technologies*, page 28–36, Montréal, Canada. Association for Computational Linguistics.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. Enhancing Empathetic Response Generation by Augmenting LLMs with Small-scale Empathetic Models. ArXiv:2402.11801 [cs].

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-And-Write: Towards Better Automatic Storytelling. *AAAI Conference on Artificial Intelligence (AAAI)*, 33(1):7378–7385.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. *ACM Comput. Surv.*, 56(3):64:1–64:37.

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, and Ge Yu. 2024. Affective Computing in the Era of Large Language Models: A Survey from the NLP Perspective. ArXiv:2408.04638 [cs].

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. In *Twelfth International Conference on Learning Representations*, Vienna, Austria.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

A Computing Information

LLaMA-3 was run on UMBC's high performance computing cluster called Ada, which at the time, had:

- 4 8x RTX 2080 Ti GPUs with 384 GB CPU memory & 11GB GPU memory each,
- 7 8x RTX 6000 GPUs with 384 GB CPU memory & 24GB GPU memory each, and
- 2 8x RTX 8000 GPUs with 768 GB CPU memory & 48GB GPU memory each.

Each node had 48 threads and two 24-core Intel Cascade Lake CPUs. The node that was assigned for a particular job was random.

B Prompts

System Prompt:

> "You are engaging in a conversation with a human. Respond to the following line of dialogue based on the given emotion and the following keywords. Just add connective words and do not add any new information to the output sentence. Do not use the word 'emotion' in the response and express the sentiment in a different way.

The last line is only for when prompting with Words.

B.1 Words

> Here are some examples:
>
> Emotion: Proud
> Keywords: 'running', 'marathon', 'first'
> Dialogue: Running my first marathon felt like such a huge accomplishment!
>
> Emotion: Sad
> Keywords: 'banana', 'plant', 'brown'
> Dialogue: It really sucks that my banana plant's leaf is turning brown
>
> Now, respond to the following. Remember, do not use the word {emo_} in the dialogue:
>
> Emotion: {emo_}
> Keywords: {kwds_}
> Dialogue:

B.2 Lexical VAD

> Valence refers to the intrinsic attractiveness or averseness of an event, object, or situation. In the context of emotions in text, valence represents the positivity or negativity of the emotion expressed. For example, words like "happy," "joyful," or "excited" have positive valence, whereas words like "sad," "angry," or "frustrated" have negative valence.
>
> It essentially measures the degree of pleasantness or unpleasantness of the emotion.
>
> Arousal indicates the level of alertness, excitement, or energy associated with an emotion. It ranges from high arousal (e.g., excitement, anger) to low arousal (e.g., calm, boredom). In text, high-arousal words might include "thrilled," "furious," or "ecstatic," while low-arousal words could be "relaxed," "content," or "lethargic."
>
> This dimension measures how stimulating or soothing the emotional state is.
>
> Dominance reflects the degree of control, influence, or power that one feels in a particular emotional state. High dominance implies feelings of control and empowerment, while low dominance suggests feelings of submissiveness or lack of control. In text, emotions like "confident," "powerful," or "authoritative" would have high dominance, whereas "helpless," "weak," or "submissive" would have low dominance.
>
> It gauges the extent to which an individual feels in control or overpowered by the emotion.
>
> Now, assume you are a normal human. Say a line of natural dialogue based on the given keywords. Just add connective words and do not add any new information to the output sentence.
>
> For example:

Emotion: Very High Valence, High Arousal, Very High Dominance
Keywords: 'running', 'marathon', 'first'
Dialogue: Running my first marathon felt like such a huge accomplishment!

Emotion: Very Low Valence, Low Arousal, Low Dominance
Keywords: 'banana', 'plant', 'brown'
Dialogue: It really sucks that my banana plant is turning brown

Emotion: Very High Valence, Very High Arousal, High Dominance
Keywords: "visit", "parents", "month"
Dialogue: I'm finally going to visit my parents next month!

Now, respond to the following:
Emotion: {v_}, {a_}, and {d_}.
Keywords: {kwds_}
Dialogue:

B.3 Numeric VAD

Valence refers to the intrinsic attractiveness or averseness of an event, object, or situation. In the context of emotions in text, valence represents the positivity or negativity of the emotion expressed. For example, words like "happy," "joyful," or "excited" have positive valence, whereas words like "sad," "angry," or "frustrated" have negative valence.

It essentially measures the degree of pleasantness or unpleasantness of the emotion.
Arousal indicates the level of alertness, excitement, or energy associated with an emotion. It ranges from high arousal (e.g., excitement, anger) to low arousal (e.g., calm, boredom). In text, high-arousal words might include "thrilled," "furious," or "ecstatic," while low-arousal words could be "relaxed," "content," or "lethargic."

This dimension measures how stimulating or soothing the emotional state is.

Dominance reflects the degree of control, influence, or power that one feels in a particular emotional state. High dominance implies feelings of control and empowerment, while low dominance suggests feelings of submissiveness or lack of control. In text, emotions like "confident," "powerful," or "authoritative" would have high dominance, whereas "helpless," "weak," or "submissive" would have low dominance.

It gauges the extent to which an individual feels in control or overpowered by the emotion.

Here's how each dimension can be defined on a scale from -5.0 to 5.0:

Valence:
-5.0: Extremely negative (e.g., intense sadness, extreme anger)
-2.5: Moderately negative (e.g., mild annoyance, slight disappointment)
0.0: Neutral (e.g., indifferent, no strong emotional reaction)
2.5: Moderately positive (e.g., mild pleasure, slight happiness)
5.0: Extremely positive (e.g., intense joy, deep love)

Arousal:
-5.0: Extremely low arousal (e.g., deep sleep, total relaxation)
-2.5: Moderately low arousal (e.g., relaxed, slightly tired)
0.0: Neutral arousal (e.g., alert but not excited, calm)
2.5: Moderately high arousal (e.g., interested, mildly excited)
5.0: Extremely high arousal (e.g., highly excited, very agitated)

Dominance: -5.0: Extremely low dominance (e.g., feeling completely powerless, totally submissive)
-2.5: Moderately low dominance (e.g., somewhat submissive, slightly dominated)
0.0: Neutral dominance (e.g., feeling neither in control nor dominated)

2.5: Moderately high dominance (e.g., feeling somewhat in control, slightly assertive)
5.0: Extremely high dominance (e.g., feeling very powerful, completely in control)

These scales provide a way to quantify and compare the emotional dimensions in a structured manner.

Now, assume you are a normal human. Say a line of natural dialogue based on the given keywords. Just add connective words and do not add any new information to the output sentence.

For example:

Emotion: Valence: 4.0, Arousal: 1.0, Dominance: 2.5
Keywords: 'running', 'marathon', 'first'
Dialogue: Running my first marathon felt like such a huge accomplishment!

Emotion: Valence: -4.0, Arousal: -2.5, Dominance: -4.0
Keywords: 'banana', 'plant', 'brown'
Dialogue: It really sucks that my banana plant is turning brown

Emotion: Valence: 2.5, Arousal: 4.0, Dominance: 1.0
Keywords: "visit", "parents", "month"
Dialogue: I'm finally going to visit my parents next month!

Now, respond to the following:
Emotion: {v_}, {a_}, and {d_}.
Keywords: {kwds_}
Dialogue:

B.4 Emojis

You are engaging in a conversation with a human. Respond to the following line of dialogue based on the given emotion and the following keywords.

Just add connective words and do not add any new information to the output sentence. The response should be exactly one line with nothing else other than the responding dialogue.

For example:

Emotion: 🤩
Keywords: 'running', 'marathon', 'first'
Dialogue: Running my first marathon felt like such a huge accomplishment!

Emotion: 😕
Keywords: 'banana', 'plant', 'brown'
Dialogue: It really sucks that my banana plant is turning brown

Emotion: 😃
Keywords: "visit", "parents", "month"
Dialogue: I'm finally going to visit my parents next month!

Now, respond to the following:
Emotion: {emo_}
Keywords: {kwds_}
Dialogue:

C VAD Training

In the following sections, we give the training instructions given to the participants verbatim. Figures 6 and 7 were shown to the participants at the time of training, but not when they were answering the survey questions.

C.1 Lexical VAD

For this study we will be using a popular model used for quantifying emotion called the Valence-Arousal-Dominance (VAD) model.

1. Valence — How pleasant you feel. A low valence would mean you are feeling negative/unpleasant whereas high valence would mean you are feeling positive or pleasant.
2. Arousal — How engaged or alert you feel. Low arousal would mean that you are more on the calmer or sleepier extreme, while high arousal would mean you are more active and energetic.
3. Dominance — How much control you have over what you feel. Low dominance implies no control and High dominance implies feeling very much in control of your emotion.

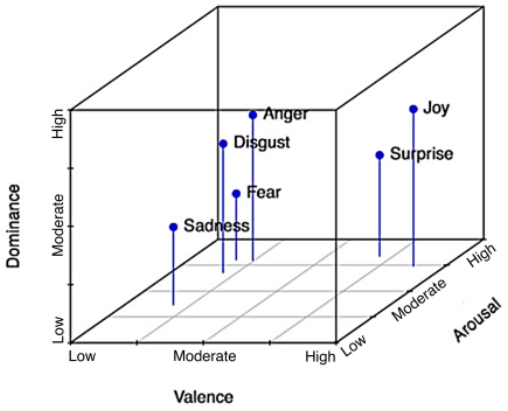

Figure 6: Lexical VAD visualization

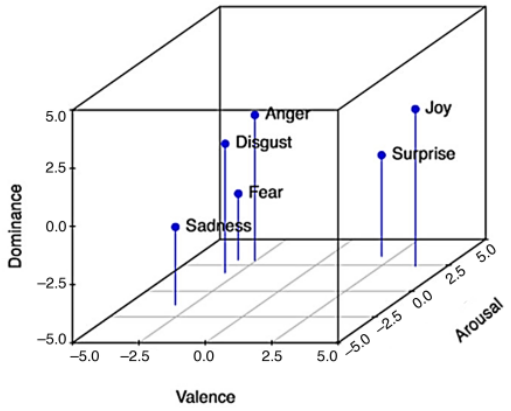

Figure 7: Numeric VAD Visualization

Please take a moment to study the figure [6], as it might be helpful for visualizing the above. It shows some common emotions we usually feel and how they map to the VAD model.

For example, consider the difference between Angry and Furious. Both of these would have low valence and moderate-to-high dominance. Being Angry has high arousal as it takes a lot of energy to feel so. Being Furious would take even more energy, as you might feel like you're about to burst. So, Angry would have High Arousal while Furious would have Very High Arousal.

Similarly, consider the difference between feeling Grateful and Joyous. Both of them are positive emotions. Grateful should have High Valence, as you are feeling pleased but not over the top, while Joyous will have Very High valence as you are really happy and elated.

Before you begin, you will go through a series of questions designed to help you understand the VAD model of emotion, followed by a practice question.

[We include a sample of the questions here:]

> What is the emotion that corresponds to VAD values High Valence, Very High Arousal and Moderate Dominance?
>
> Surprise
> Joy
> Anger
>
> Correct Answer: Surprise.
>
> High Valence indicates this is more of a positive emotion. Very High Arousal means there is a lot of energy behind this, while Moderate Dominance shows that you are not entirely in control. This could be either Joy or Surprise, but having higher arousal and lower dominance suggests Surprise is the answer.

C.2 Numeric VAD

For this study we will be using a popular model used for quantifying emotion called the Valence-Arousal-Dominance (VAD) model. In this model, the X, Y, and Z axes span only from -5 to 5 and can be defined as follows

1. Valence — How pleasant you feel on a range from -5 to 5. Here, -5 would mean you are feeling very negative/unpleasant whereas a 5 would mean you are feeling very positive or pleasant.

2. Arousal — How engaged or alert you feel on a range from -5 to 5. -5 would mean that you are more on the calmer or sleepier extreme while 5 would mean you are more active and energetic.

3. Dominance — How much control you have over what you feel on a range from -5 to 5. In this case, -5 implies no control and 5 implies feeling very much in control of your emotion.

Please view the figure [7] for a visual representation of these ranges.

Before you begin, you will go through a series of questions designed to help you understand the VAD model of emotion. In the first set of questions you will be provided the numerical values and need to choose the discrete emotion those VAD values correspond to. Then you will be given a practice question similar to the rest of the questions in the survey.

[We include a sample of the questions here:]

What is the discrete emotion that corresponds to VAD values -2.5 (Valence), 2.5 (Arousal), and 2.0 (Dominance)?

Sadness
Anger
Fear

Anger is the correct answer!

-2.5 Valence indicates this is a negative or unpleasant emotion. 2.5 arousal means it takes a lot of energy to feel this way. Therefore, it cannot be Sadness. 2.0 Dominance means you are somewhat in control of how you feel, so it is unlikely to be Fear either. So, Anger is the most appropriate option.

D Survey Questions

[Any text appearing within brackets like this in the following section is a note and did not appear in the survey.]

Informed Consent Information

Informed consent: You must be 18 years or older to participate in this study.

The purpose of this study is to see if large language models like ChatGPT describe emotions in the same way that people do. You are being asked to volunteer because you are a native English speaker.

You will be shown a series of 4 different sentences and need to determine if each sentence conveys a certain emotion. Note that the emotion may be displayed in an abstract way. You will be taught how to read this abstraction before answering the questions.

The survey may take about 15 minutes to complete.

You are welcome to withdraw or discontinue participation at anytime, but due to the volume of participants expected from crowdsourcing, we will not be paying participants for incomplete surveys. If you withdraw from the study or do not complete the survey, your data will be deleted.

Please take your time and do the best you can. There are no right or wrong answers, but we reserve the right to not pay if we determine that you are not following directions or taking the task seriously.

All data obtained will be anonymous. There is no way for us to find out who you are, and your data will not be shared with any other parties under any circumstance.

Any information learned and collected from this study in which you might be identified will remain confidential. The investigator will attempt to keep your personal information confidential. To help protect your confidentiality, your data will only be linked to a randomly-assigned ID. Any information required to pay you (i.e., username) will be kept in a spreadsheet on a secure server separate from the other data you provide.

Only the investigator and members of the research team will have access to these records. If information learned from this study is published, you will not be identified by name and all results will be reported in aggregate. By signing this form, however, you allow the research study investigator to make your records available to the University of Maryland, Baltimore County's Institutional Review Board (IRB) and regulatory agencies as required to do so by law.

Introduction to the questions

In the following survey, you will be asked questions based on understanding and recognizing emotions in text. Following the practice questions, there will be 16 questions in total.

The emotions will be described as a word.

For example: Angry, Happy, Annoyed

In the following survey, you will be asked questions based on understanding and recognizing emotions in text. Following the practice questions, there will be 16 survey questions in total.

The emotions will be described in terms of valence, arousal and dominance. For example: High Valence, High Arousal, Low Dominance.

In the next page, we will explain what these terms are and how they relate to emotions.

In the following survey, you will be asked questions based on understanding and recognizing emotions in text. Following the practice questions, there will be 16 survey questions in total.

The emotions will be described in terms of valence, arousal and dominance. For example: Valence: -2.0, Arousal: 3.0, Dominance: 4.0.

In the next page, we will explain what these terms are and how they relate to emotions.

In the following survey, you will be asked questions based on understanding and recognizing emotions in text. Following the practice questions, there will be 16 survey questions in total.

The emotions will be described using emojis.

For example: 😄, 😗, 🥵

Representational Alignment Instructions (same for all representations)

For each question below, you will be shown an emotion and a set of sentences. Given the specified emotion, pick the sentence that is the best fit.

Note: In some cases, one or more of the choices might be identical. If you feel that sentence is the best fit, feel free to pick any one. Also, the sentences are not meant to be ironic or sarcastic.

Accuracy and Realism Instructions (same for all representations)

For the next set of questions, you will be given a sentence and an emotion described in a word.

We will ask you to rate the sentence based on how well it conveys the given emotion, and how realistic it sounds (i.e. it sounds like something a person would say). Please rate how well the sentence reflects each statement.

[Bonus Question — we used this as a filter to gauge how much attention users gave to the survey. In a handful of cases, we removed answers to this question that seemed to be written in extremely poor English or written by a language model.]

Think of a movie, television show, or book that you watched or read recently that made you feel a strong emotion.

Please share the name of the movie, show, or book. Then tell us what that emotion was in plain English, and why did you feel that way?
(Your response should be at least 30 characters long.)

E Additional Analysis: Adherence to Keywords

Model	One	Two	Three	Acc
GPT-4, 1x	1.00	1.00	.936	.978
LLaMA-3, 1x	.908	.897	.781	.862
LLaMA-3, 3x	.969	.969	.850	.930
LLaMA-3, 10x	.981	.981	.853	.938

Table 3: Percentage of generated sentences that contain the respective number of keywords. *x indicates the # of times the model was sampled. Accuracy is percentage of the three keywords that are correct, averaged across all sentences.

Since we were constraining the model to generate using content-based keywords verbatim, we lemmatized each inputted keyword and compared them against the words in the generated sentence. Shown in Table 3, all sentences generated by GPT-4 had at least two keywords present, and 93.6% had all three. By contrast, 9.17% of LLaMA-3's generated sentences had no keywords present at all. Re-sampling improved results, reducing this percentage to just 1.94%. Similar to Chen and Wan

(2024), this overall trend shows that conditioning on emotions in addition to content keywords does not significantly impair the LLM's ability to copy the keywords when generating.