UMBC

**THESIS APPROVAL SHEET**

Title of Thesis: Connecting Language and Emotion in Large Language Models for Human-AI Collaboration

Name of Candidate: Shadab Hafiz Choudhury shadabc1@umbc.edu

Master of Science 2025

Graduate Program: Computer Science

Thesis and Abstract Approved:

DocuSigned by:
Lara Martin
2544E895FBC742A...

Lara Martin

laramar@umbc.edu

Assistant Professor

Computer Science and Electrical Engineering

4/30/2025 | 12:35 PM EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

# Shadab Hafiz Choudhury

shadabc1@umbc.edu
www.shadabchy.com
github.com/namerlight

## Research Interests

Accessibility, Human-AI Collaboration, Responsible NLP, Responsible Multimodal Models

## Education

**Aug '23 – May '25** **University of Maryland, Baltimore County**, *MS in Computer Science*
Thesis – Connecting Language and Emotion for Human-AI Collaboration

**Sept '16 – Dec '20** **North South University**, *BS in Computer Science and Engineering*
Thesis - NAVI: Navigational Assistance for the Visually Impaired using Computer Vision

## Selected Publications

**2025** **S. H. Choudhury**, A. Kumar and L. J. Martin, "GPT's Devastated and LLaMA's Content: Emotion Representation Alignment in LLMs for Keyword-based Generation," *arXiv preprint*, Mar. 2025

**2025** A. S. M. Mohsin and **S. H. Choudhury**, "Quantifying Nanoparticles and Monomer-Dimer Distribution in Optical Images using Deep Learning," *ACS Omega*, vol. 10, no. 1, pp. 862-870, Jan. 2025

**2024** A. S. M. Mohsin and **S. H. Choudhury**, "Label-free Quantification of Gold Nanoparticles at the Single-Cell level using a Multi-Column Convolutional Neural Network," *Analyst*, vol. 149, no. 8, pp. 2412–2419, Apr. 2024

**2021** I. J. Ananya, S. Suad, **S. H. Choudhury**, and M. A. Khan, "A Comparative Study on Approaches to Acoustic Scene Classification Using CNNs," *MICAI 2021, Lecture Notes in Computer Science,* vol. 13067, pp. 81–91, Oct. 2021

**2021** **S. H. Choudhury**, A. J. Aurin, T. A. Mitaly, and R. M. Rahman, "Predicting the Possibility of COVID-19 Infection using Fuzzy Logic System," *Int'l Journal of Intelligent Information and Database Systems*, vol. 14, no. 3, pp. 239–256, Jan. 2021

## Research Experience

**May '24 – Aug '24** **Graduate Research Assistant**, *University of Maryland, Baltimore County (PI: Dr. Lara J. Martin)*

Fine-tuned LLMs and smaller language models for emotion classification and emotional text generation.

Designed surveys and carried out user studies on emotional representations and how emotion generation capabilities of LLMs align with human expectations.

**Sept '21 – Aug '23** **Research Assistant**, *Brac University (PI: Dr. Abu S. M. Mohsin)*

Designed custom density mapping CNNs and used transfer learning and ensembles for novel nanoparticle detection and analysis problems, achieving 78-93% test accuracy across tasks.

Created real-time water quality and quantity, and emergency monitoring dashboards using React, ChartJS, Google Maps API and Firestore.

Enhanced water quality forecasting by 38% over prior work using ARIMA and Regression models; built Flask API to serve real-time predictions.

**Oct '20 – Feb '21** **Undergraduate Directed Research**, *North South University (PI: Dr. Md. Ashrafuzzaman Khan)*

Analysed audio signal processing methods for feature extraction via images and embeddings, then tested multiple neural network architectures for classifying environmental audio.

Achieved 93% accuracy from spectrograms and 81% accuracy with embedding models.

## Academic Service

**2025** **ACL 2025**, Secondary Reviewer

**2024** **Wordplay @ ACL 2024**, Secondary Reviewer

## Teaching Experience

**Aug '23 – May '25** **Graduate Teaching Assistant**, *University of Maryland, Baltimore County*

TA for CMSC 341 – Data Structures in C++ and Java. Proctored exams, held office hours, and graded projects and homework assignments.

**Feb '20 – Jun '20** **Marker**, *North South University*

Grader for 4 sections of ENG102 (Introduction to Composition) and ENG105 (Advanced Composition). Proctored and graded exams, quizzes and assignments.

## Industry Experience

**Feb '23 – Jun '23** **Python Developer**, *Dviz Technologies*

Developed prompt engineering pipeline for an LLM chatbot to categorize and recommend construction products

Achieved 90% accuracy for categorizing products from datasheets using GPT 3.5 embeddings and XGBoost.

**May '22 – Feb '23** **Data Scientist and Machine Learning Engineer**, *Neovotech*

Built customizable English and Swedish text-to-speech API using FastAPI for automated content creation.

Designed recommender system for matching outfits in database with in-wild images.

Designed fashionable outfit-generation system from text descriptions using StyleGAN.

**Feb '22 – May '22** **Data Scientist and Machine Learning Intern**, *Eucaps AB | Neovotech*

Built 8+ automated scrapers for data mining financial and stock news sites with Selenium.

Designed pipeline for neural machine translation and summarization of scraped data.

## Grants and Awards

**2024** HackUMBC 2024, Best First-Time Hackers, *Category Winner*

**2016 – 2020** Merit-based Scholarship, North South University *(50% Tuition Waiver)*

**2015, 2017** The Daily Star 0- and A-Level Awards for Academic Excellence

**2023 (Declined)** €1300 Travel Grant, Mediterranean Machine Learning Summer School

## Skills

**Programming** Python, JavaScript

**Technologies** Git, Docker, MySQL, MongoDB, ImageJ, Google Firebase, Amazon EC2, Slurm

**Libraries** PyTorch, OpenCV, Pandas, NumPy, Librosa, FastAPI, Matplotlib, SciKit-learn, Django, Flask, Selenium, RabbitMQ, Gradio, Scrapy, BeautifulSoup, ReactJS

## Presentations and Talks

**2022** Water Analytics Tools: Quality & Quantity Monitoring & Purification Using IoT, ML & Nanotechnology

**2020** Workshop on Python by NSU ACM Student Chapter R&D

## Volunteering

**2019 – 2020** **NSU ACM Student Chapter**, *Chapter Officer, Webmaster*

Organized and led research and development meetings. Assisted organizing workshops on Python and Motion Graphics.

Directed social media presence and outreach of the Chapter. Successfully promoted national-scale competitive events with over 600 participants.

**2017 – 2018** **NSU ACM Student Chapter**, *Publications Team Lead*

Ensured regular release of promotional materials for events. Coordinated with North South University's public relations.

Co-edited three annual student publications.

## Miscellaneous

**2021** Water Innovation Challenge Competition by Bangladesh 2030 Water Resources Group, *Final Round*

**2018** Wordsmiths 2018 by NSU Department of English and NSU Communications Club. *Second Place*

# ABSTRACT


Title of Dissertation: Connecting Language and Emotion in Large Language Models for Human-AI Collaboration

Shadab Hafiz Choudhury
Master of Science, 2025

Dissertation directed by: Dr. Lara Martin
Assistant Professor
Department of Computer Science and Electrical Engineering



Large Language Models demonstrate linguistic abilities on par with humans, able to generate short texts, stories, instructions, and even code that's often indistinguishable from what is created by humans. This allows humans to use large language models (LLMs) *collaboratively* — as communication aides or writing assistants.

However, humans cannot always assume an LLM will behave the same way another person would. This is particularly evident in subjective scenarios such as where emotion is involved. In this work, I explore to what depth do LLMs perceive and understand human emotions, and look at ways of describing an emotion to an LLM for collaborative work. First, I study the problem of *classifying emotions* and show that LLMs perform well on their own, and can also improve smaller models at the same task. Secondly, I focus on *generating emotions*, using the problem space of keyword-constrained generation and a human participant-study to see where human expectations and LLM outputs diverge and how we can minimize any such misalignment. Here, I find that using English Words and Lexical expressions Valence-Arousal-Dominance (VAD) scales lead to good alignment and generation quality, while Numeric dimensions of VAD or Emojis fare worse.

# Connecting Language and Emotion in Large Language Models for Human-AI Collaboration

by

**Shadab Hafiz Choudhury**

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Master of Science
2025

**Advisory Committee:**
Assistant Professor Lara Martin, Chair/Advisor
Associate Professor Frank Ferraro
Associate Professor Cynthia Matuszek

© Copyright by
Shadab Hafiz Choudhury
2025

# Acknowledgments

The last two years at UMBC, and the last few months working on this Thesis, were stressful. But they were also some of the most exhilarating and gratifying periods of my life. I could name dozens of people here in the acknowledgments, because every single person I met or interacted with throughout this time, as well as my previous mentors and colleagues, contributed to it in some minute way. But I would like to keep this short and sweet.

First up is my amazing advisor Lara Martin. Obviously, this thesis was only possible due to the guidance and mentorship I received, and my time at UMBC was so fruitful in large part thanks to all the time we spent conversing (and spilling tea). I don't think there's much more I could say here I haven't already expressed in person.

I also want to acknowledge my friends and labmates at UMBC — Naren Sivakumar, Naomi Tack, Luke Zimmermann, Sourajit Saha, Shaswati Saha, and the other members of the LARA Lab, as well as the CVG and IRAL labs. And of course, this also goes for the faculty at UMBC whom I frequently interacted or collaborated with — Tejas Gokhale, Foad Hamidi, Frank Ferraro, and Cynthia Matuszek.

Finally, my parents, my sister, and my friends — Ishrat, Faria, Sarah, Ashley — who all cheered me on despite the hundreds or thousands of miles between us.

Table of Contents

List of Tables

List of Figures

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| HCI | Human Computer Interaction |
| NN | Neural Network |
| LM | Language Model |
| LLM | Large Language Model |
| MLLM | Multimodal Large Language Model |
| VAD | Valence-Arousal-Dominance |

# Chapter 1: Introduction

## 1.1 Humans and Large Language Models

The last few years have seen an explosive growth of artificial intelligence (AI) used in human-facing applications — that is, applications where people directly interface with an AI model in a back-and-forth manner.

Large Language Models (LLMs) in particular were the single biggest driver of this change, and the main antecedent for this is simply their ability to both understand and generate text on par with a human. They can understand questions, orders, and even long, detailed instructions with different requirements and conditionals, and respond accordingly. The response are not just eloquent and realistic, but also informative, follows instructions, and can solve varied basic programming and logical tasks.

The above is not perfect, of course. LLMs may fail to correctly follow an instruction. They may generate text that is irrelevant, misleading, or factually incorrect. They may be simply incapable of understanding certain topics or solving certain problems due to how they function on a mechanical or algorithmic level.

This thesis is concerned with one specific facet of an LLM's capabilities — its ability to understand and generate *Emotions* in text. This is particularly important for Human-AI collaboration as *all* human interactions are coloured by emotions. Emotions can provide

context for social interactions, predicate specific reactions or responses in an individual, and in a Human-Human interaction be used to provoke a particular reaction in another person. A system that can do the aforementioned could be considered the holy grail of affective computing [1]. To that end, the broad, overarching questions asked: Can an LLM recognize and interact with emotions in text in the same way a human would? If not, that means there is a gap between how an LLM deals with emotions and how humans would. In that case, can we do anything to reduce this gap?

## 1.2 Research Questions

While the two aforementioned questions are important and underlie this whole thesis, they are also huge, broad-stroke questions that form the basis of entire sub-fields of affective computing and language modeling research. Rather than tackle everything all at once, I focus on two smaller research questions.

**RQ1**: To what extent can Large Language Models understand and recognize emotions in human dialogue?

**RQ2**: What is the best way to express an emotion to a Large Language Model during Human-AI collaboration?

First, **RQ1** is important in the sense it establishes a baseline for LLMs' capabilities. LLMs are fundamentally models for *understanding* text — their ability to generate text is a separate step that is preceded by how well it understand a textual input. Therefore,

before investigating generation or human-AI collaboration, I investigate how well can LLMs recognize emotions in human dialogue. This is a very straightforward classification task.

I approached this from two angles. The first angle is the straightforward approach, simply classifying a line of dialogue using an LLM. For the second, I dived deeper and instead asked the LLM to *analyze* a given line (from a training dataset) and describe the emotions expressed in detail. After that, I trained small language models using the descriptions alongside the original training data. This experiment showed us whether or not the description of the line aligned with the labeled emotion.

Chapter 3, describes the experimental process, results and discussion for **RQ1** in detail.

Once I noted that LLMs are capable of recognizing and describing detailed emotions from text, I moved onto generating emotional text. It is well known that LLMs are excellent at generating text [2, 3], so instead of evaluating the quality of generated emotional text, I focused on *communication*. That is, I investigated what are the best ways to communicate an emotion to an LLM so that it generates an appropriate sentence.

Humans tend to communicate emotions in text in two ways: describing the emotion in words ("I'm so angry!") or using an emoji ("😡"). However, both words and emojis can be ambiguous and the exact emotional tone of different words can be hard to measure [4]. To resolve this problem, Mehrabian developed the Valence-Arousal-Dominance (VAD) scale, which measures emotion on three numeric dimensions [5].

As all of these are valid methods for expressing emotions, I evaluated each of them

in turn through a user survey. I generated sentences using some keywords and an emotion expressed as either English Words, VAD described in words (Lexical VAD), VAD given as numbers (Numeric VAD), or Emojis. Then I asked participants to select the sentence that best matched a given emotion.

Chapter 4, describes the experimental process, results and discussion for **RQ2** in detail.

Over the years a lot of work has been done using language models for emotion classification, sentiment analysis, and related problems. Before describing the experiments, in Chapter 2, I first look at the history and background of language processing algorithms and early language models in affective computing, in the lead up to modern LLMs. I also discuss other similar works that study emotions in LLMs, and highlight the research gap this thesis aims to cover.

## 1.3 Contributions

This thesis focuses on a small but important aspect of text understanding and controlled text generation in LLMs. The contributions of this thesis are as follows. I show that -

1. Large Language Models are somewhat effective as emotion classifiers, achieving 76.9% accuracy on on two emotion classification datasets: SemEval-2025 and GoEmotions. They can classify both coarse emotions (5 classes), but not fine-grained emotions (28 classes).

2. LLMs can be used to describe the emotional content of a line of dialogue. This can

be used to extend training datasets and improve the capabilities of small language models. On a dataset with coarse emotion labels, there was an average improvement of 2.9%.

3. Using Words to describe an emotion in a prompt gives the best alignment between human expectations and an LLM's generated text. Lexical VAD is a close second, while Numeric VAD and Emojis have bad alignment.

4. For generating the most realistic and clearly emotional sentence, Words and Lexical VAD are the best options for expressing an emotion in a prompt.

The study covering **RQ2** was partially supported by a 2024 UMBC COEIT Interdisciplinary Proposal (CIP) Award. It has been published as a preprint at the time of this thesis' completion [6].

# Chapter 2: Background

A lot of previous work has been done at the intersection of affective computing and language modeling. Some early works in machine learning in fact involved analyzing speech and language, which were relatively easy to process. This includes datasets like IEMOCAP [7] and techniques like principal component analysis or support vector machines [8,9]. However, as machine learning has grown, so has the breadth of available datasets, benchmarks and methods for both understanding and generating emotion in language.

This chapter will discuss previous research that has been done on emotional text recognition and generation, as well as define any literature gaps that this thesis aims to cover.

## 2.1 Modeling Emotion in Language

A large number of datasets and benchmarks exist for modeling emotion in language. Fairy Tales was one of the earliest datasets, featuring short stories annotated with emotions. [10, 11]. More recent datasets include EmotionLines, CBET, Empathetic Dialogues and GoEmotions [12–15], all of which focus on human-human conversation records annotated with emotions per line.

Early work in the area such as [16] and [17] simply used a keyword-based approach, assigning an emotional component to certain words in a sentence, then classifying based off

those emotional features. While such methods are certainly valid even today — a modern language model would assign an certain weight to a token-level feature, which is conceptually the same thing, only self-supervised rather than handcrafted — they suffer from an inability to handle context (such a model would be incapable of distinguishing between 'tears' (of happiness) versus 'tears' (of sadness), for instance) and from the need to have every word labeled. And of course, many words can be ambiguous or completely neutral.

With sequential neural networks, methods using Recurrent NN's or Long Short Term Memory (LSTM) networks became popular [18], to varying degrees of success on various benchmarks [19–22]. This was also the case with more modern transformer-based models like BERT [23] and it's subsequents [24–27].

In the age of LLMs, emotion classification and understanding is *arguably* a solved problem. [28] shows that state of the art LLMs are fully capable of responding in an emotional or empathetic way, performing very well on evaluation metrics. However, they also state that alignment is one potential avenue of future work. [29] show that LLMs can be used for annotating emotion, essentially replacing human labelers, but their results still have a significant margin of error. Additionally, [30] shows that humans and LLMs are not always aligned and that there is a gap in emotional understanding.

## 2.2 Generating Emotions with Language Models

Emotional sentence generation has been widely studied as part of style-transfer or empathetic dialogue generation problems. A lot of early work on emotional sentence generation,

prior to neural networks, relied on rule-based systems [31]. With neural networks, the typical approach became to condition the output on specific emotional words [32, 33], and this did not change even when scaling up to transformers like GPT-2 [34, 35]. [36, 37] utilized the VAD space, adding an emotional vector to the internal representation of the text. [38] conditioned a variational autoencoder on Emojis instead.

Work on generating emotions with LLMs has also been done. A variety of methods like chain-of-thought, retrieval-augmented generation, prompt tuning, etc. have all been successfully used [39–44] have all been used to varying degrees of success. However, one thing all of these approaches have in common is that they are all based on words.

So, while smaller language models have experimented with using VAD and Emojis, research on using LLMs to generate emotion have not. This is one potential gap that I partially hope to address.

# Chapter 3: Recognizing Emotions

Recognizing emotions in text is not always an easy task. Even humans make mistakes all the time. People often say the wrong thing and annoy or anger the person they are talking to. Part of the reason for this is that we learn to recognize emotions in a multimodal manner — a combination of speech, tone, body language, and many other small cues.

In language and text, most of those cues do not exist. The only way to recognize an emotion is by looking at the words or phrasing to find contextual clues that can help detect emotion. As described in Chapter 2.1, this comes with its own challenges.

In this study, I investigated how Large Language Models classify an emotion in a sentence, and how well can they describe the emotion that's being expressed in that sentence.

## 3.1 Datasets

This study was primarily done on two datasets: SemEval-2025's Task 11's Multi-Label Emotion Detection dataset, hence called SemEval-11 [45], and the GoEmotions dataset [15].

The SemEval-11 dataset consisted of only 5 emotions: "anger", "fear", "sadness", "joy", and "surprise". These emotions are based off Ekman's emotion classes and are widely used in a majority of other emotion datasets.

The GoEmotions dataset consisted of 28 emotions, making it a more fine-grained classification task: "admiration", "amusement", "anger", "annoyance", "approval", "caring",

"confusion", "curiosity", "desire", "disappointment", "disapproval", "disgust", "embarrassment", "excitement", "fear", "gratitude", "grief", "joy", "love", "nervousness", "optimism", "pride", "realization", "relief", "remorse","sadness", "surprise", and "neutral".

To control the variable of dataset size, I used the full SemEval-11 training dataset, consisting of 2768 data points for training and validation, and 116 data points for testing. GoEmotions is a much larger dataset of 58k data points, from which I sampled 2768 and 116 data points randomly to use for training and validation, and testing respectively.

### 3.2 Experimental Design

Two experiments were carried out at this stage.

1. Classifying emotions using an LLM and Zero-Shot prompting.

2. Using an LLM to describe the emotion in the sentence, then fine-tuning a small Language model for classification using the description as additional data.

For both experiments, Gemma-2-3B, hence referred to as Gemma-2 [46], is used as the LLM of choice for several reasons. Firstly, a large number of evaluations and experiments already exist with larger LLMs like GPT-4, LLaMA, and other competing models. Comparatively, there are far fewer evaluations carried out on models with less than 4B parameters.

Secondly, models *smaller* than Gemma-2, including LLaMA-3.2-1B, LLaMA-3.2-3B, [47] and SmolLM-1.7B [48] failed to follow the instructions for both experiments.

### 3.2.1 Classification

Initially I intended to evaluate on Zero-Shot classification. However, all open-source LLMs tested failed to return a consistent or coherent answer in the zero-shot setting.

In the few shot setting, they perform much better. I selected two arbitrary samples from the training dataset as exemplars. The prompt is given below:

```
Classify the given sentence into one of the following emotions: [list
of emotions go here].
Name the emotion without any description or reasoning.
For Example:
Passage: "By far the coolest thing I've seen on this thread yet"
Emotion: "joy"
Passage: "You should dm her and say I'm sorry"
Emotion: "sadness"
Passage:{passage}
Emotion:
```

The `[list of emotions go here]` was replaced by the emotions in the SemEval-11 and GoEmotions datasets respectively, with each emotion in words separated by commas, at evaluation time.

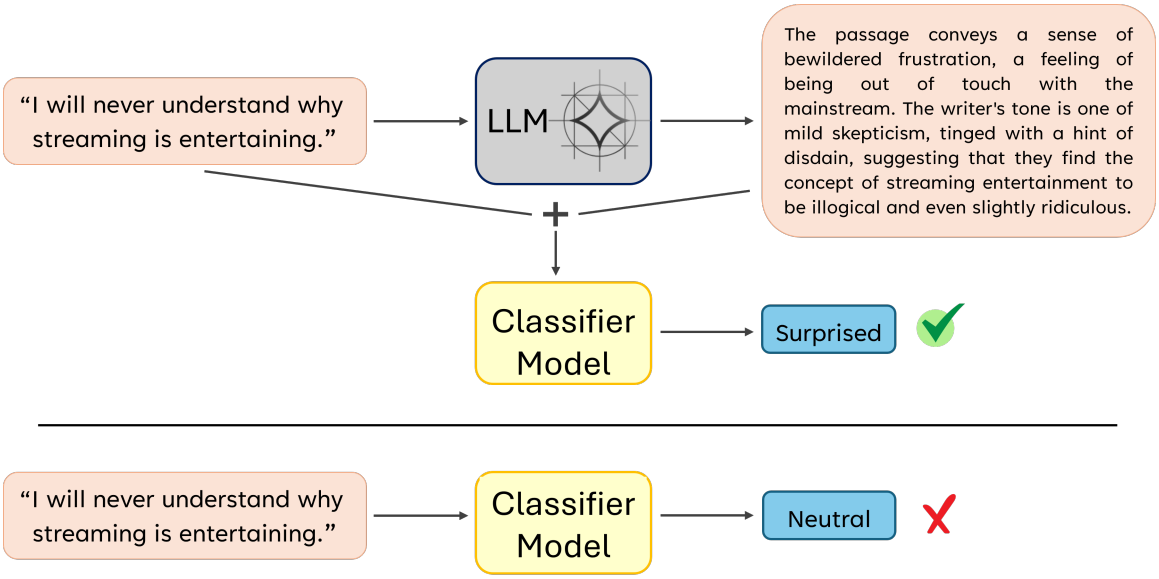


Figure 3.1: Summary of Experiment 2 for Recognizing Emotions. The top and the bottom results are compared across the full test dataset.

### 3.2.2 Describing Emotions

Figure 3.1 summarizes the overall process of Experiment 2.

Two approaches were used to compare the classifiers. The LLM was asked to describe the emotion using the following zero shot prompt:

```
Write a long paragraph describing the unique emotional tone style of
the following passage without referring to specifics about the topic.
Do not include anything after that paragraph. Any explanations should
be in the same paragraph as the description.
```

```
Passage: {passage}

Description:
```

The descriptions were generated for both the training and validation sets, as well as at evaluation time.

Four classifier models were tested: BERT [23], RoBERTa-Base, RoBERTa-Large [49], and ModernBERT [50]. They were evaluated on the mean F1 score across all labels.

The models were all fine-tuned using a standard set of parameters. No hyperparameter optimization was carried out. The training:validation ratio was 90:10. They were fine-tuned for 5 epochs, with a learning rate of $2 \times 10^{-5}$, and batch sizes of 4.

## 3.3 Results and Discussion

### 3.3.1 Classification Results

Table 3.1 shows the results of carrying out Few-Shot classification.

It appears that classifying SemEval-11 is a relatively straightforward task for both models. GPT-4 achieves 0.769% accuracy, which is comparable to a smaller Language model directly fine-tuned on this dataset (as shown in the following section, Section 3.3.2). Gemma-2 is much worse, but still somewhat acceptable. However, on the GoEmotions dataset, both models perform quite poorly.

One potential hypothesis is that the LLMs are not 'bounded' to any strict labels. So, in the SemEval dataset, when prompted to stick to a small selection of five emotional labels, both models tended to respond with one of those labels. However, in the GoEmotions

| Model | SemEval-11 | GoEmotions |
|---|---|---|
| Baseline | 0.200 | 0.036 |
| GPT-4 | 0.769 | 0.071 |
| Gemma-2 | 0.552 | 0.181 |

Table 3.1: Few-Shot Classification accuracy, computed as the rate of correct predictions (direct accuracy). I compare against the baseline random-guess results here.

dataset, with 28 different labels, attention was likely unevenly distributed across the labels. This may have caused many of the predictions to fall in the same 'group', but not precisely the same emotion. For instance, the LLMs may have been unable to distinguish between "disappointment", "disapproval", "disgust" in the text as effectively.

### 3.3.2 Describing Emotions Results

Tables 3.2 and 3.3 shows the results of Experiment 2 on the two respective datasets.

The results are somewhat interesting and contradictory here. On SemEval-11, there is a very clear improvement from using Descriptions as additional context, averaging a 2.9% improvement across the models. RoBERTa-large performs the best in general, indicating that this is a problem that is solved by a bigger model, but using descriptions allows ModernBERT to reach almost same level of accuracy with fewer parameters.

However, there is no improvement to be seen on the GoEmotions dataset. This actually follows on from the results of experiment 1: That LLMs are not able to distinguish

| Model | No Description | With Description | Improvement |
|---|---|---|---|
| BERT-base-uncased | 0.709 | 0.725 | 1.6% |
| RoBERTa-Base | 0.678 | 0.716 | 3.8% |
| RoBERTa-Large | 0.774 | 0.780 | 0.6% |
| ModernBERT | 0.716 | 0.771 | 5.5% |

Table 3.2: Test Results on SemEval-11 when using finetuned small LMs with and without descriptions, as well as the percentage improvement from using descriptions.

| Model | No Description | With Description | Improvement |
|---|---|---|---|
| BERT-base-uncased | 0.840 | 0.828 | -1.2% |
| RoBERTa-Base | 0.829 | 0.831 | 0.03% |
| RoBERTa-Large | 0.815 | 0.825 | 1.0% |
| ModernBERT | 0.838 | 0.827 | -0.09% |

Table 3.3: Test Results on GoEmotion when using finetuned small LMs with and without descriptions, as well as the percentage improvement from using descriptions

between all the emotions in the 28 classes, so the descriptions overlap across the different emotions and potentially even make classification harder for smaller LMs. In future work, it may be interesting to study exactly at which level of granularity do LLMs start failing.

To summarize:

- Both models perform very well at classifying five emotions in the SemEval-11 dataset, but much worse at the twenty-eight emotions in the GoEmotions dataset.
- Using an LLM to generate descriptions follows on from the previous result. It improves the ability of smaller models when there are only a few emotions, but is ineffectual when there are many emotions to choose from.

# Chapter 4: Generating Emotions

What is the best way to express an emotion to a Large Language Model during Human-AI collaboration? I investigated this from the perspective of *alignment* — that is, ensuring that Humans and LLMs are on the same page in regards to both the emotion that is to be expressed and how intense it should be.

To study alignment, I framed this problem as a keyword-based sentence generation problem. Keyword-based, alternatively called lexically constrained, sentence generation involves generating a sentence or text that contains a specific keyword or set of keywords. This is useful for scenarios where an user wants to maintain a high level of control over the generated sentence. Keyword-based sentence generation is especially useful for uses like assistive technologies where users have limited motor control or simply do not have the time to input full sentences [51, 52]. It also has uses in human-robot interaction [53], advertising and marketing [54], and so on.

I then added an additional constraint on top of the keyword-based generation: emotion. There are now two constraints on the final output sentence, an emotional one and a lexical one. While small language models exist for keyword-based sentence generation, using LLMs has some advantages for the aforementioned use cases. LLMs have higher quality text generation, have a wide range of information recall and reasoning abilities, and their outputs can be improved through techniques like retrieval-augmented generation. LLMs

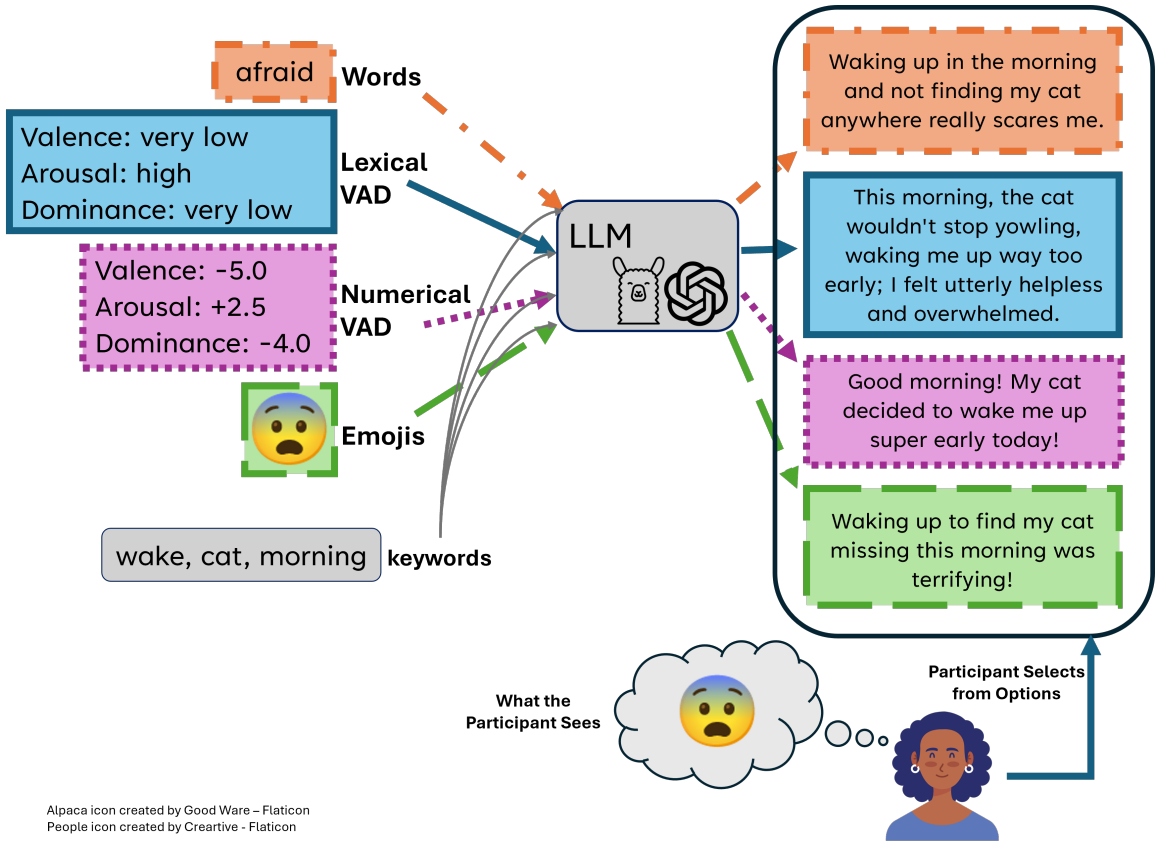


Figure 4.1: Representation Alignment study. Three keywords and an emotion from one of the four representations are used to generate a sentence. Participants are shown the emotion in only one of the representations and select the sentence they think best fits that emotion.

have also been shown to have near-perfect keyword-based generation abilities [55], almost always using given keywords in the sentence.

To generate a sentence, an LLM must be given a prompt that describes what the output should be like. The straightforward way to make the output 'emotional' is to simply mention the emotion in the prompt in words and say the output should express it. However, as described in Chapter 1.2, ambiguity is a limiting factor.

Consider these two words for describing a very similar emotion: "angry" and "furious". 'Angry' can be seen in a couple different ways. Some people become loud and aggressive when they are angry, while others have a 'cold' anger. This can vary across cultures and across people who have different levels of english fluency. On the other hand, the term 'furious' almost always refers to a more energetic outburst of anger. In cases like this, the utility of the VAD scale is highlighted.

In the Valence-Arousal-Dominance scale [5], an emotion is mapped on three axes described here:

**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.

**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.

**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.

These scales are a less subjective, more fine-grained way of describing emotions. They are primarily used in psychology but have also seen some use in affective computing.

In this chapter, I describe the different emotional representations, the process of generating emotional, keyword-based sentences using different representations of emotions, the design of the user survey, and finally discuss the results of the participant study.

## 4.1 Representing Emotions

I choose four different emotion representations to compare:

- **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**.

Words is self-explanatory. I choose a list of emotions and their most commonly used english term based off [15]. The results seen in Chapter 3 clearly show that dealing with 28 emotions is too difficult for LLMs and might also be for humans, so I narrow it down to 18 emotions, removing some emotions that were too difficult to distinguish in between. The final list of emotions was: Grateful, Joyful, Content, Surprised, Excited, Impressed, Proud, Anxious, Afraid, Terrified, Annoyed, Angry, Furious, Sad, Devastated, Ashamed, Embarrassed, and Guilty.

The goal was to cover all major 'groups' of emotions, and to cover a range within each group. For instance, Anxious, Afraid and Terrified are all similar emotions falling under the 'Fear' category, but in respective order are progressively more intense.

The VAD scale is typically a numeric scale where the values range from -1.00 to 1.00. Russel and Mehrabian's original work [5] only describes the VAD values for 7 emotions. [56] extended this to cover a wider range of emotions that correspond to the 18 emotions I selected. However, parsing decimal values like -0.46 or 0.03 is very unintuitive and may have posed a major cognitive load for non-technical users during the user survey.

Therefore, the values were converted to two different scales that are easier to comprehend. First, they were normalized to a wider range of -5.0 to +5.0, then rounded to the nearest 0.5. There was a loss of fine-grained information when converting to this discrete numeric representation, but each emotion is still distinguishable, so it was acceptable. This gives the Numeric VAD representation.

For Lexical VAD, instead of converting to a numeric value in the range of -5.0 to +5.0, the values were segmented into five ranges. From the lowest to the highest value, all values falling within each segment were mapped to the terms Very Low, Low, Moderate, High, or Very High respectively.

Finally, Emojis were selected based on the closest interpretation to the emotion. In some cases, the emoji was named after the emotion, which made it a simple choice. In other cases, I went with the closest matching popularly used emojis. In the survey, Emojis were embedded as unicode so that participants would see the set that they were used to seeing

| Group | 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 4.1: All 18 Emotions used in the Study and their Representations.

on their device.

Table 4.1 shows the full list of emotions used, including the category, the corresponding emoji, and the lexical and numeric VAD values. The category of the emotions were used when selecting which emotions to use but were not integrated into the generation nor shown to participants.

## 4.2 Sentence Generation

Two different LLMs were used for generating sentences: GPT-4-Turbo 2024-04-09 [3] and LLaMA-3-70B [47], referred to as GPT-4 and LLaMA-3 from here on respectively. The proprietary GPT-4 model was used through OpenAI's API, while the open-source LLaMA-3 model was fetched from [57] and run locally, but had to be quantized to 8-bit integer weights [58] in order to fit on the available hardware.

Other models considered at this stage included LLaMA-3-8B, LLaMA-3.1-8B, Gemma-2-9B, and Gemma-2-27B. Out of all the models tested, LLaMA-3-70B had the best output even after quantization.

I limited each input to just three content keywords. This struck a balance between giving LLMs sufficient context without almost writing a short sentence. The keywords were sets of arbitrarily-chosen, common everyday words like [`Place`, `Great`, `Korean`], [`Finals`, `Semester`, `Math`].

For each of the four emotion representations, 90 sentences were generated — five sets of keywords for all 18 emotions, for a total of 360 sentences per model.

For generating a sentence with emotions represented as Words or Emojis, I used plain few-shot prompting [59]. The exemplars were selected randomly, but included at least one positive emotion and one negative emotion. The keywords and sentence in the exemplars were the same, with the emotion simply being written in either English or as an emoji.

For prompting using either form of VAD, I used step-back chain-of-thought prompting [**?**]. First, I prompted the model to give an explanation of VAD, then convert it to scale from -5.0 to +5.0 if using Numeric VAD. Once it has a description of VAD as well as a numeric mapping, I use a prompt similar to the few-shot prompt used for Words and Emojis, once again replacing the emotion with either Lexical or Numeric VAD.

The prompts are given in full in Appendix A.

## 4.3 User Study Design

The user study was evaluated and approved by the UMBC IRB under Protocol 1380: "Evaluation of Language Generation Technologies".

For the user study, each participant was randomly assigned an emotion representation (one of Words, Lexical VAD, Numeric VAD or Emoji). They were then asked to answer 15 questions in total. There were two types of questions: alignment and realism.

### 4.3.1 Participants

A total of 200 participants were recruited using the crowdsourcing platform Prolific — 100 participants for each LLM being evaluated. The participants were paid at a rate of $14/hr. The average response time of the survey was approximately 15 minutes, so each participant

was paid $3.50. Each participant was only allowed to complete the survey once.

Participants were required to be 18+, fluent in English, and residing in the United States. Participants were shown a consent form at the start of the survey, and at the end of the survey they were asked to answer a bonus Question, which was used as a filter to gauge how much attention users gave to the survey. Users answered this question with poor English, or gave an irrelevant answer were excluded from the final study results. Users who completed the survey abnormally quickly (in 2-3 minutes or less) were also excluded.

Participants were instructed on how to parse the emotions and what to consider for the response. The exact instructions are given in Appendix B.

### 4.3.2 Alignment Questions

The alignment questions focus on the alignment between an user and an LLM. Here, I gave the user an emotion in the form of their assigned representation, followed by four sentences. These four sentences were generated by an LLM using a different representation each. The user was asked to select the one that best matched the given emotion. For instance:

Q1. **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!

In this case, choice 1 was generated by prompting the LLM with Words. Choice 2 was generated by prompting the LLM with Lexical VAD. Choice 3 was with Numeric VAD, and Choice 4 was with emojis. The display order to the user was randomized.

Each user was given 10 questions of this type. One thing to note at this time was the cognitive load on the participants as well as the time spent taking the survey. While participants who got Words or Emoji representations could understand the emotion easily, it was much harder for users in VAD representations to understand the emotion, as very few normal people are familiar with the system. 10 was a good compromise between having a large number of answers and having responders who were focused through the entire survey.

### 4.3.3 Realism Questions

Following the alignment question, participants were asked to answer 5 questions relating to the 'realism' of the generated sentences. In this case, realism simply describes how accurate a generated sentence is when it comes to expressing an emotion, as well as how natural it sounds.

Participants were asked to answer three 5-point Likert scale questions [60]. Participants were asked to rate the questions on a 5-point Likert (Not at all, Slightly, Moderately, Very, Extremely), where 1 corresponded to Not at all and 5 corresponded to Extremely. The example below shows the questions asked:

> 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!"**

How much does the sentence...

- Convey the emotion above?

- Sound like something that you would say?

- Sound like something that someone else would say?

From this point onwards, these questions will be referred to as "Convey", "You'd say", and "Someone Else'd say", respectively.

The "Convey" question allows us to directly assess whether the sentence generated by the LLM accurately expresses the emotion. The "You'd say" and "Someone Else'd say" questions allow us to see if the sentence is actually realistic and human-sounding.

## 4.4 Results and Discussion

After removing any invalid responses, there were 26, 25, 28 and 29 participants for Words, Lexical VAD, Numeric VAD and Emojis respectively for GPT-4, and 25 participants exactly for each of the four conditions for LLaMA-3. Due to this difference, the counts were normalized during evaluation and discussion.

### 4.4.1 Representation Alignment

Representation Alignment as defined as a combination of two factors:

1. If the participant was more likely to select the sentence that was generated using the

same representation they were given (*match rate*)

2. The average Shannon Entropy for sentences using that representation was lower than for other representations— representing better agreement across participants (*entropy*)

A high match rate and a low entropy would indicate good alignment, and vice versa means bad alignment. For example, if participants who were shown Lexical VAD emotions were more likely to select sentences generated using Lexical VAD (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. A random match rate would be 25.0%, so any value above this would be notable. Table 4.2 shows the entropies and Figure 4.2 shows the match rates of both GPT-4 and LLaMA-3. Each category on the x-axis corresponds to the condition the participant was in — what representation they saw. The colors delineate what emotion representation was used for sentence generation. Results for GPT-4–generated sentences are on the top, LLaMA-3 on the bottom.

Shannon Entropy was selected over inter-rater agreement measures like Krippendorff's Alpha because it expresses the gap between different representations better. Due to the large number of questions (90) and small number of raters per question (7), inter-rater agreement was very close to random across all four representations.

Overall, most participants agreed that the Words representation was the best fit when they were presented the emotion in Words. This was an expected result, as humans express emotions in Words most often—even more so than emojis, which can be subjective and

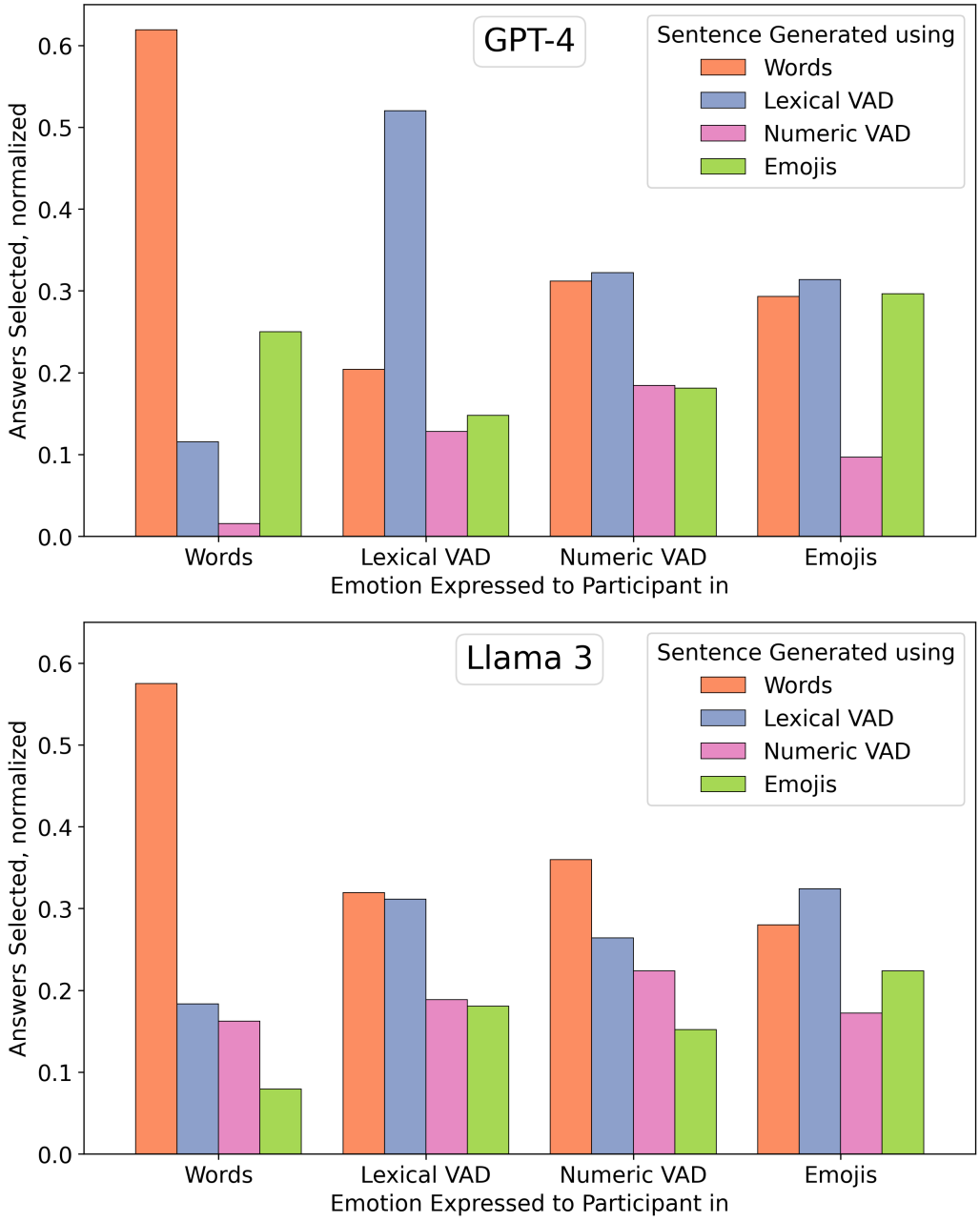


Figure 4.2: Match Rates for Emotion Representation. This is the percentage of times a sentence was selected per emotional representation

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

Table 4.2: Shannon Entropy values. Bolded values show the most agreement, underlined are second most.

ambiguous [61]—and LLMs are also trained on natural language data. Words had a 61.9% match rate using GPT-4 and 57.5% using LLaMA-3, as well as entropy values of 0.32 and 0.42 respectively.

Surprisingly, Lexical VAD also has a high match rate of 52.0% with GPT-4 and a lower but still notable 31.2% with LLaMA-3. The entropy values were 0.61 and 0.72 respectively. While the agreement is worse than that of Words, the high match rate is noteworthy. This occurs even though even though the participants and the LLMs are given different instructions (Appendix B) and prompts (Appendix A). This indicates both humans and LLMs may be drawing on similar ideas or memorized information when considering the emotion.

Numeric VAD had poor alignment. One possibility is that participants struggled to conceptualize the numbers, but were able to understand them more easily when expressed

in quantitative words (leading to Lexical VAD doing better). Numeric VAD has been used in other works to control the output of generative models [62, 63], so this result indicates that they may not be as good an idea for LLMs.

Finally, Emojis, despite being widely used, did not show particularly good alignment. While it has good entropy scores, the match rate only barely exceeds 25.0% with GPT-4 only, at 29.7%. In fact, when participants were shown Emojis, they tended to select the response generated by Lexical VAD regardless of the model.

While the trend is small, it was interesting as it happened across the models. 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, Lexical VAD breaks down emotions into components that LLMs can work with while Emojis are discrete symbols to text-based models, and humans can break Emojis down into individual facial features but have a harder time understanding VAD scales due to a higher cognitive load.

Figure 4.3 shows the entropy values broken down on a per-emotion basis. For individual emotions per representation, the results are fairly similar, with most values between 0.70 to 0.95 meaning poor agreement, with some outliers. For GPT-4 using Words, the emotions "grateful", "anxious", "embarrassed", and "guilty" had perfect agreement, contributing to the representation's overall low entropy. The emotion "guilty" had perfect average entropy for both LLaMA-3 and GPT-4.

To summarize:

- Words shows the highest alignment through match rate and mean entropy. Lexical

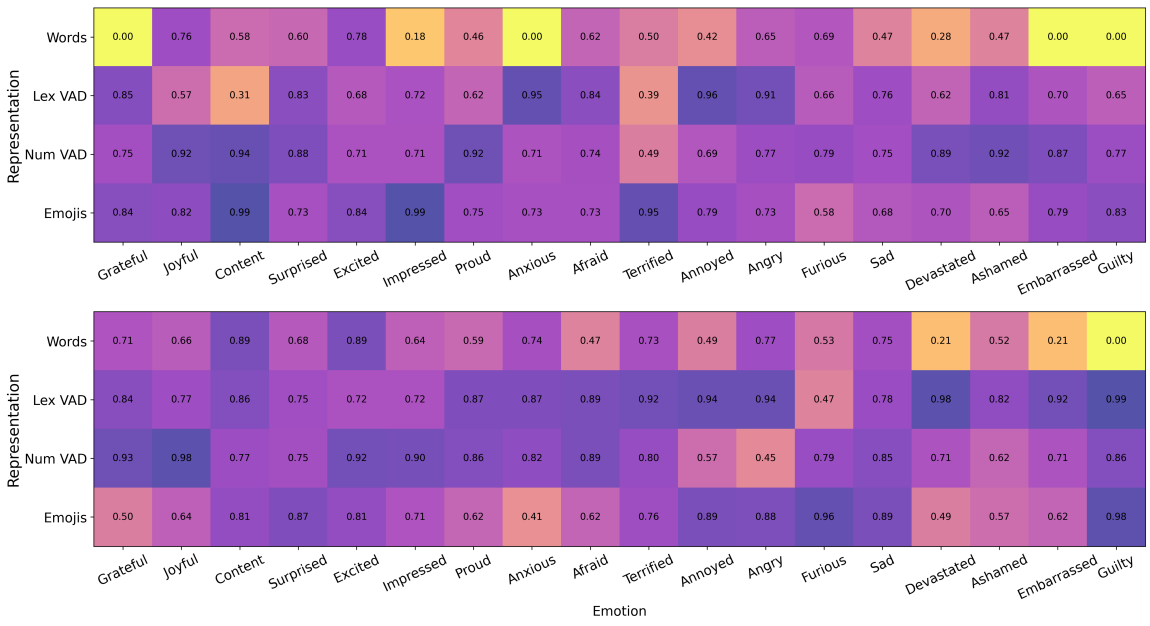


Figure 4.3: Shannon entropy of each emotion across representations. Lower (brighter) values are better, denoting more "agreement" between participants and the LLM. The top heat map is for GPT-4, while the bottom is for LLaMA-3.

VAD is a close second in terms of match rate.

- There is no clear trend in the alignment on a per-emotion basis, regardless of the LLM used.

### 4.4.2 Realism

The average Likert scores for these three questions across conditions and for both models are shown in Figure 4.4.

To analyze the significances of these results, an ANOVA statistical significance test was run on the Likert ratings for all three Realism questions for both models.

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 $p < .05$.

When a pairwise t-test was run on the statistically significant results, Words was found to be significantly better at conveying the emotion than Numeric VAD for GPT-4 ($p = 0.002$), while Lexical VAD was significantly better at "Convey" than Numeric VAD for LLaMA-3 ($p = 0.018$). These results further show that Numeric VAD scores underperform.

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 for accurate emotional expression.

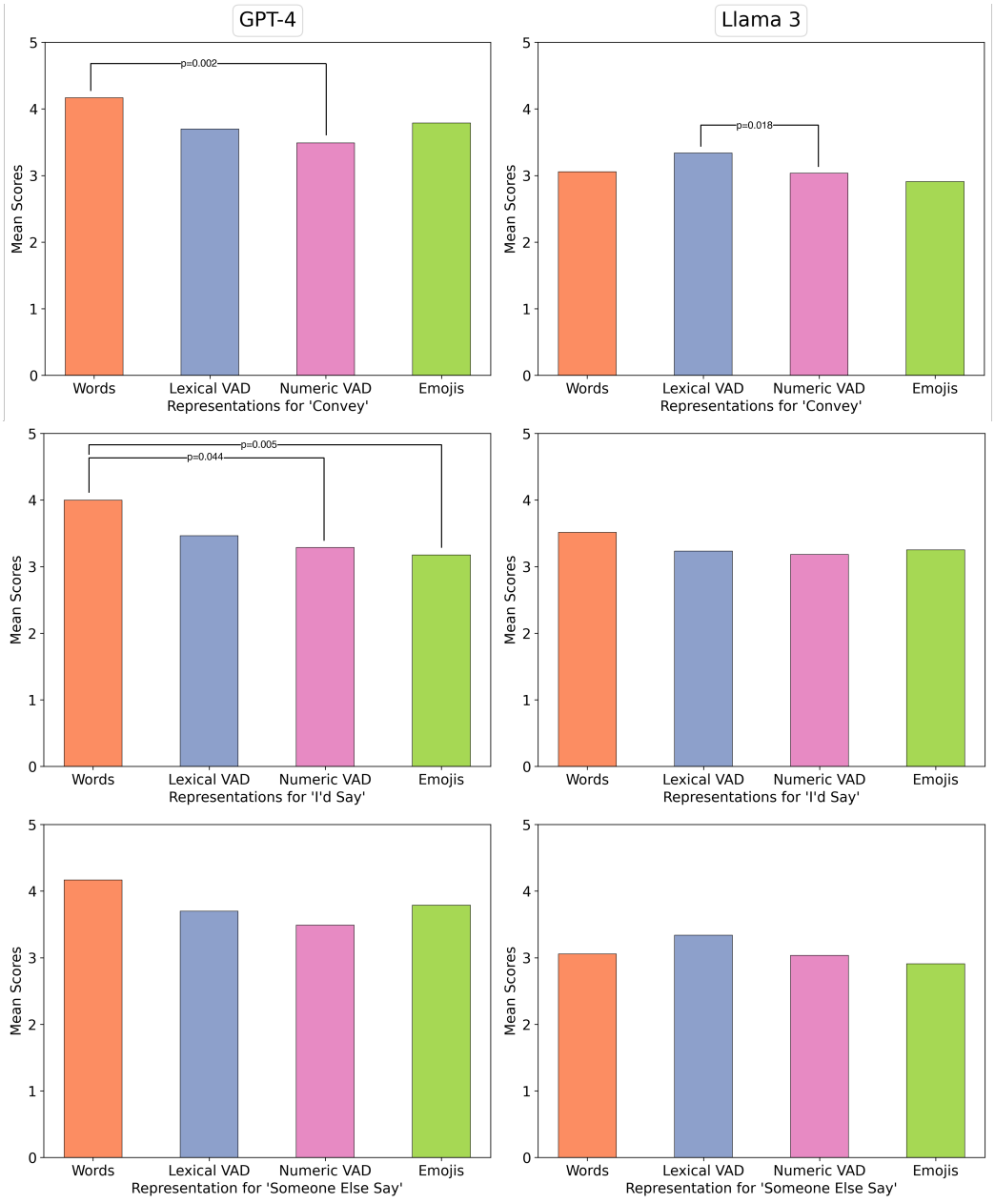


Figure 4.4: Left: GPT-4, Right: LLaMA-3. In order from Left to Right and Top to Bottom: a, b. Histograms of the Mean Scores for 'Convey'; c, d. Histograms of the Mean Scores for 'You'd say'; e, f. Histograms of the Mean Scores for 'Someone Else'd Say'.

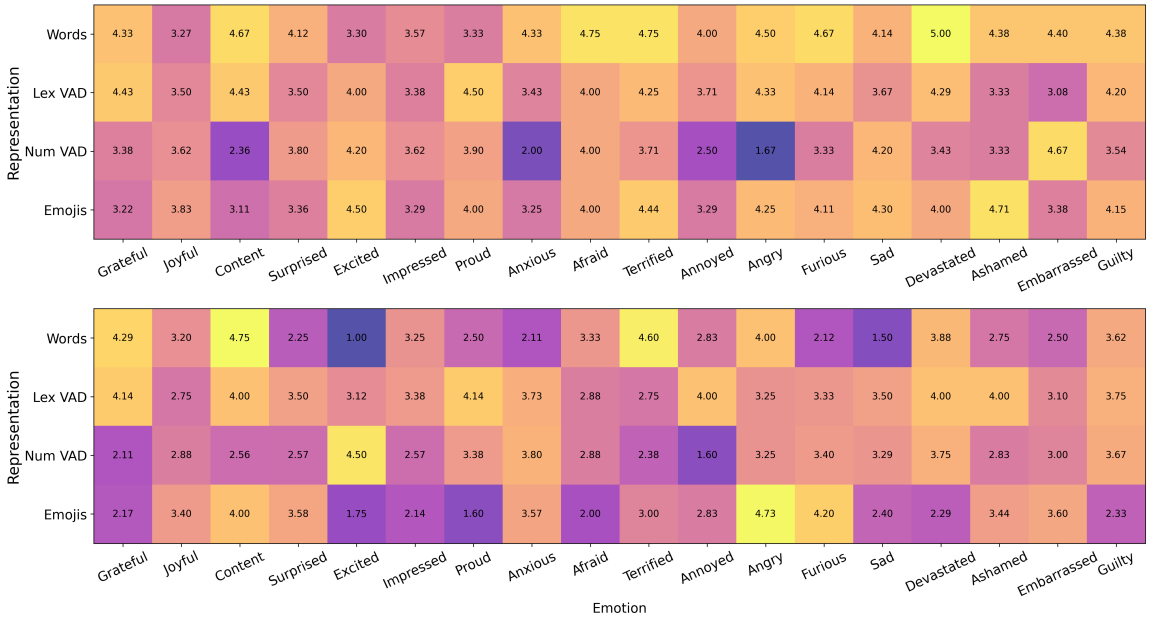


Figure 4.5: Mean "Convey" Scores for each emotion per representation. Higher (brigher) values are better. The top map shows results for GPT-4, while the bottom map is LLaMA-3.

Additionally, compared to GPT-4 LLaMA-3's generated sentences are considered slightly worse for conveying the emotion and sounding realistic across the board.

The mean scores for the "Convey" question are broken down by emotion in Figure 4.5. The scores are relatively high across the board, with some outliers. For instance, such as Words with GPT-4 seemingly struggling with sounding "excited" or "proud" in a realistic way, while Numeric VAD struggles with a lot more emotions like "anxious" and "angry". LLaMA-3, on the other hand, finds it very hard to sound "surprised" or "excited" when the emotion is described as a word, but not when either form of VAD is used.

This leads to another finding: that despite guardrails to make LLMs more amicable

and positive [64–66], they can still struggle to generate certain positive emotions under complex conditions.

To summarize:

- Words and Lexical VAD appear to be the best representations for generating realistic sentences across the board.
- Different emotions are expressed best by using different representations, but Words and Lexical VAD are slightly better on average once again.

# Chapter 5: Conclusion

LLMs are increasingly being used to generate essays and stories, parse and respond to emails, and to help humans communicate with each other. However, if LLMs are to be used as an intermediary between humans and in Human-AI collaboration, then they must be able to understand all aspects of human communication — including the emotional aspect. They must be able to understand an emotion as it is communicated to them, and generate an appropriate emotion as directed by a human.

In this thesis, I looked at two facets of the emotional understanding and expressivity of Large Language Models. First, their ability to understand emotions, evaluated through an emotion classification tasks on two datasets. Both these datasets consists of short lines of text in the form of an interpersonal conversation and are labeled with up to 28 different emotions. Next, their ability to generate emotions, evaluated through an human user study of a keyword-based sentence generation experiment.

In Chapter 3, the study on emotion classification indicates that LLMs are quite good at both simply recognizing and labeling the emotion, as well as describing it in detail and justifying the answer of a classification result — as long as there's a limited number of emotions. Once there are too many emotions, LLMs become ineffectual. We also showed that using an LLM to generate descriptions of emotions led to better accuracy in small language models, but once again only when there were a few emotions.

Following up, in Chapter 4, I studied how LLMs respond to different ways of representing emotions. For this study, four representations were used: Words, Lexical VAD, Numeric VAD, and Emojis. People typically expect Words and Emojis to perform the best, as they are the most prevalent in human interaction and so would be prevalent in the training data of LLMs. However, it turns out that Lexical VAD performs very well, close to that of Words and much better than Emojis.

Overall, this thesis shows that LLMs are capable of recognizing emotions in the same way a human might, as well as generating them, and this bodes well for their use in Human-AI collaboration. However, LLMs can struggle to recognize very fine grained emotions as they are described in words. One way to overcome this is to use alternative scales like VAD where emotions are less subjective. Thus, during generation, it is best to use either English words or Valence-Arousal-Dominance values given in words to describe the emotion to the LLM.

## 5.1 Limitations

There are some limitations to the work described in this thesis. Firstly, the results may not generalize perfectly to every single LLM, as I tested a relatively narrow range of LLMs throughout the two main studies. For recognizing emotions, the results may be worse on other datasets, or using other LLMs. However, as a very small LLM was used to generate descriptions, I expect other LLMs to actually perform *better*.

Another factor is that LLM training data is not well known, and it is possible that

the models may have already been trained on the data they were tested on. This type of train-test leakage is a major issue with LLMs in general.

In the second study, the keyword based prompt framing may have affected the results, as well as the use of a quantized model. Additionally, due to funding limitations I only investigated two models; carrying out further user studies on other models may lead to a more robust conclusion. Another limitation is that the participants were all English speakers in the United States, and people from other countries or cultures may have responded differently.

## 5.2 Future Work

I showed that using LLMs can be used to enhance the training data of smaller language models. This is worth exploring in other tasks as well that are limited by a lack of training data beyond emotion classification.

Researchers and developers using LLMs for downstream controlled text generation tasks should consider using Lexical VAD if they need a greater degree of control and less ambiguity than english Words.

Finally it may be prudent to actively research how to improve both classification and controlled generation using VAD. This could be done in many ways — including texts where lexical VAD is used to describe emotions in the pretraining data, or in the post-training stage such as by instruction-tuning the LLMs on lexical VAD data.

# Appendix A: Prompt Engineering

Listing all the prompts used in Chapter 4.

The system prompt was:

```
"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. We explicitly forbid it from treating the Emotion as a keyword.

## A.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:
```

## A.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:

## A.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:

## A.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:
```

# Appendix B: Survey Questions

In this section, all the instructions given in the survey are listed.

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

## B.1 VAD Training

As participants were very likely to be completely unfamiliar with the VAD model, a short training session was included in the survey that explained how the VAD model (whether Lexical or Numeric) works.

### B.1.1 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

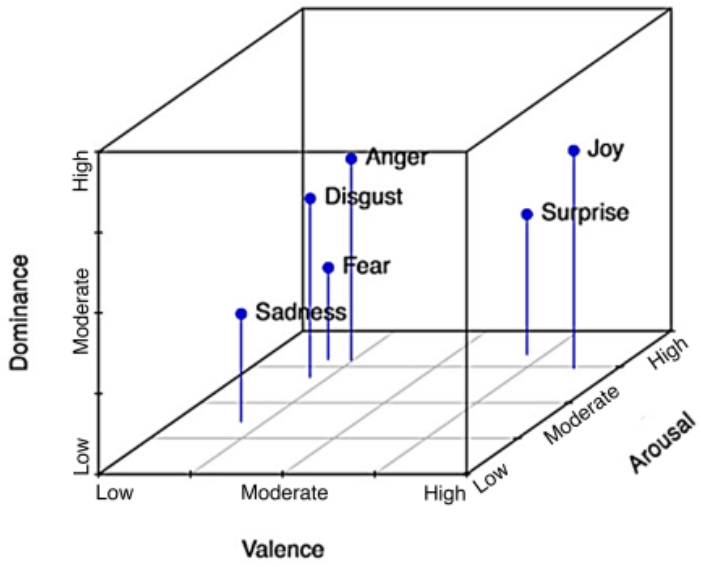


Figure B.1: Lexical VAD Visualization

implies no control and High dominance implies feeling very much in control of your emotion.

Please take a moment to study figure B.1, 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.

### B.1.2 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 B.2 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.

---

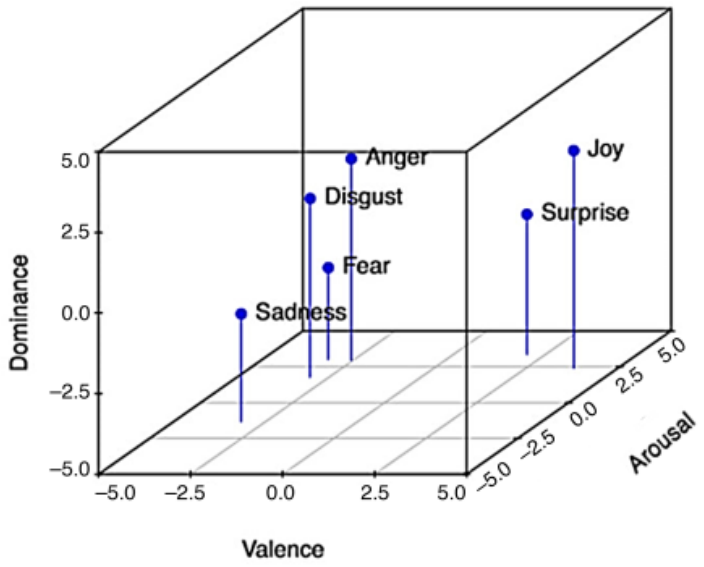


Figure B.2: Numeric VAD Visualization

[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.

## B.2 Survey Questions

The following sections shows the consent form as well as the basic instructions for the main survey questions sections.

**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. (B.1)

---

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. (B.1)

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.)

# Bibliography

[1] Rosalind W. Picard. *Affective computing*. MIT press, 2000.

[2] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. October 2020.

[3] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine

McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].

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

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

[6] Shadab Choudhury, Asha Kumar, and Lara J. Martin. GPT's Devastated and LLaMA's Content: Emotion Representation Alignment in LLMs for Keyword-based Generation, March 2025. arXiv:2503.11881 [cs].

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.

[8] Yang Li and Yunxin Zhao. Recognizing emotions in speech using short-term and long-term features. In *5th International Conference on Spoken Language Processing (ICSLP 1998)*, pages paper 0379–0. ISCA, November 1998.

[9] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Emotion Classification Using Web Blog Corpora. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, pages 275–278, November 2007.

[10] Cecilia Ovesdotter Alm and Richard Sproat. Emotional Sequencing and Development in Fairy Tales. In Jianhua Tao, Tieniu Tan, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 668–674, Berlin, Heidelberg, 2005. Springer.

[11] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff, editors, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[12] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[13] Ameneh Gholipour Shahraki. Emotion Mining from Text, 2015.

[14] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics.

[16] Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo. Detecting Implicit Expressions of Sentiment in Text Based on Commonsense Knowledge. In Alexandra Balahur, Ester Boldrini, Andres Montoyo, and Patricio Martinez-Barco, editors, *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 53–60, Portland, Oregon, June 2011. Association for Computational Linguistics.

[17] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 1556–1560, New York, NY, USA, March 2008. Association for Computing Machinery.

[18] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances, May 2019. arXiv:1905.02947 [cs] version: 1.

[19] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, 2017. arXiv:1708.00524 [stat].

[20] Panpan Li, Jun Li, Feiqiang Sun, and Peng Wang. Short Text Emotion Analysis Based on Recurrent Neural Network. In *Proceedings of the 6th International Conference on Information Engineering*, ICIE '17, pages 1–5, New York, NY, USA, August 2017. Association for Computing Machinery.

[21] Sayyed M. Zahiri and Jinho D. Choi. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. *CoRR*, January 2017.

[22] Wenxiang Jiao, Michael R. Lyu, and Irwin King. Exploiting Unsupervised Data for Emotion Recognition in Conversations, October 2020. arXiv:2010.01908 [cs] version: 2.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[24] Linkai Luo and Yue Wang. EmotionX-HSU: Adopting Pre-trained BERT for Emotion Classification, July 2019. arXiv:1907.09669 [cs].

[25] Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534, March 2023.

[26] Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121, December 2020. ISSN: 2576-8964.

[27] Mohammad Rostami, Digbalay Bose, Shrikanth Narayanan, and Aram Galstyan. Domain Adaptation for Sentiment Analysis Using Robust Internal Representations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for*

*Computational Linguistics: EMNLP 2023*, pages 11484–11498, Singapore, December 2023. Association for Computational Linguistics.

[28] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is ChatGPT Equipped with Emotional Dialogue Capabilities?, April 2023. arXiv:2304.09582 [cs].

[29] Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496, Barcelona Spain, August 2024. ACM.

[30] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the Emotional Intelligence of Large Language Models, July 2024. arXiv:2402.12071 [cs].

[31] T S Polzin and A Waibel. Emotion-Sensitive Human-Computer Interfaces. In *ISCA tutorial and research workshop (ITRW) on Speech and Emotion*. ISCA, 2000.

[32] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Affect-LM: A Neural Language Model for Customizable Affective Text Generation, April 2017. arXiv:1704.06851 [cs].

[33] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. Generating Responses with a Specific Emotion in Dialog. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy, July 2019. Association for Computational Linguistics.

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners.

[35] Ishika Singh, Ahsan Barkati, Tushar Goswamy, and Ashutosh Modi. Adapting a Language Model for Controlled Affective Text Generation, November 2020. arXiv:2011.04000 [cs].

[36] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-Driven Dialog Generation, April 2019. arXiv:1904.02793 [cs].

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

[38] Xianda Zhou and William Yang Wang. MojiTalk: Generating Emotional Responses at Scale, May 2018. arXiv:1711.04090 [cs].

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

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

[41] Abdur Rasool, Muhammad Irfan Shahzad, Hafsa Aslam, and Vincent Chan. Emotion-Aware Response Generation Using Affect-Enriched Embeddings with LLMs, October 2024. arXiv:2410.01306 [cs].

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

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

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

[45] Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, March 2025. arXiv:2503.07269 [cs].

[46] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David

Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, October 2024. arXiv:2408.00118 [cs].

[47] 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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hai-

ley Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan

Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked,

Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. arXiv:2407.21783 [cs].

[48] Loubna Ben Allal, Anton Lozhkov, and Elie Bakouch. SmolLM - blazingly fast and remarkably powerful, February 2025.

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. arXiv:1907.11692 [cs].

[50] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, December 2024. arXiv:2412.13663 [cs].

[51] Junxiao Shen, Boyin Yang, John J Dudley, and Per Ola Kristensson. 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, March 2022. Association for Computing Machinery.

[52] Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K. Kane. "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 1–14, New York, NY, USA, April 2023. Association for Computing Machinery.

[53] Lucrezia Grassi, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. Enhancing LLM-Based Human-Robot Interaction with Nuances for Diversity Awareness. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 2287–2294, August 2024. ISSN: 1944-9437.

[54] Hao Zhou, Minlie Huang, Yishun Mao, Changlei Zhu, Peng Shu, and Xiaoyan Zhu. Domain-Constrained Advertising Keyword Generation. In *The World Wide Web Conference*, WWW '19, pages 2448–2459, New York, NY, USA, May 2019. Association for Computing Machinery.

[55] Xiang Chen and Xiaojun Wan. Evaluating, Understanding, and Improving Constrained Text Generation for Large Language Models, March 2024. arXiv:2310.16343 [cs].

[56] Yuting Guo and Jinho D. Choi. Enhancing Cognitive Models of Emotions with Representation Learning. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 141–148, Online, June 2021. Association for Computational Linguistics.

[57] meta-llama. meta-llama/Llama-3.3-70B-Instruct · Hugging Face, December 2024.

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

[59] 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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[60] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55, 1932.

[61] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 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, May 2017. Number: 1.

[62] Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis, June 2023. arXiv:2306.00648 [cs, eess].

[63] Yi He, Shengqi Dang, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. EmotiCrafter: Text-to-Emotional-Image Generation based on Valence-Arousal Model, January 2025. arXiv:2501.05710 [cs] version: 1.

[64] 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. Towards Understanding Sycophancy in Language Models, October 2023. arXiv:2310.13548 [cs].

[65] Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue,

editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia, May 2024. ELRA and ICCL.

[66] Henry Papadatos and Rachel Freedman. Your LLM Judge may be biased. March 2024.

ProQuest Number: 31999463

INFORMATION TO ALL USERS
The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.


Distributed by
ProQuest LLC a part of Clarivate ( 2025).
Copyright of the Dissertation is held by the Author unless otherwise noted.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

ProQuest LLC
789 East Eisenhower Parkway
Ann Arbor, MI 48108 USA