Fluent but Unfeeling: The Emotional Blind Spots of Language Models

Bangzhao Shu[1]*, Isha Joshi[1]*, Melissa Karnaze[2], Anh C. Pham[3], Ishita Kakkar[3], Sindhu Kothe[2], Arpine Hovasapian[4], Mai ElSherief[1]

[1]Northeastern University
[2]UC San Diego
[3]University of Massachusetts Amherst
[4]Independent Researcher
{shu.b, joshi.ishaa, m.elsherif}@northeastern.edu,
{mkarnaze, skothe}@ucsd.edu,
{acpham, ikakkar}@umass.edu

Abstract

The versatility of Large Language Models (LLMs) in natural language understanding has made them increasingly popular in mental health research. While many studies explore LLMs' capabilities in emotion recognition, a critical gap remains in evaluating whether LLMs align with human emotions at a fine-grained level. Existing research typically focuses on classifying emotions into predefined, limited categories, overlooking more nuanced expressions. To address this gap, we introduce EXPRESS, a benchmark dataset curated from Reddit communities featuring 251 fine-grained, self-disclosed emotion labels. Our comprehensive evaluation framework examines predicted emotion terms and decomposes them into eight basic emotions using established emotion theories, enabling a fine-grained comparison. Systematic testing of prevalent LLMs under various prompt settings reveals that accurately predicting emotions that align with human self-disclosed emotions remains challenging. Qualitative analysis further shows that while certain LLMs generate emotion terms consistent with established emotion theories and definitions, they sometimes fail to capture contextual cues as effectively as human self-disclosures. These findings highlight the limitations of LLMs in fine-grained emotion alignment and offer insights for future research aimed at enhancing their contextual understanding.

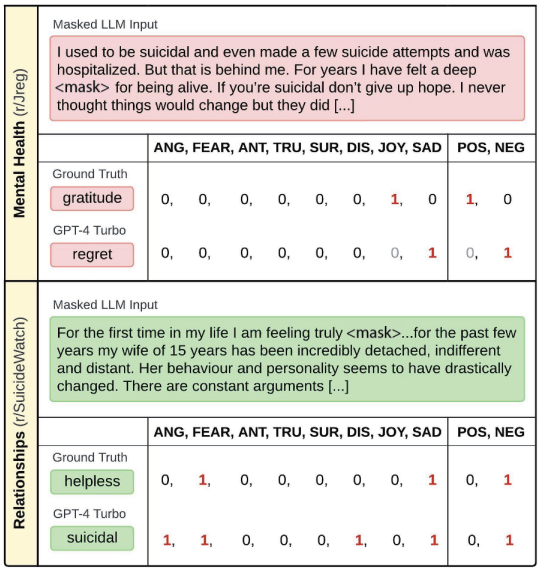

Figure 1: An example of EXPRESS and our emotion recognition evaluation framework is presented. Language models are prompted to predict masked emotions in the text. Both the predicted and self-disclosed emotion words are then decomposed into eight basic emotion dimensions and two sentiment dimensions. In these examples, GPT-4 Turbo fails to accurately capture emotions in human self-disclosures.

Introduction

LLMs have been trained on vast amounts of written human language (e.g., from the internet), some of which contains descriptions of emotional experiences and emotional discourse (Achiam et al. 2023; Brown et al. 2020). LLMs have been designed to interface with users with some knowledge of human emotion (Wang et al. 2023). From an evolutionary standpoint, universal human emotions, like fear, joy, and disgust, developed to solve unique sets of challenges faced by ancestral humans, and LLMs can learn about emotions through training.

Specifically, LLMs can be assessed on competencies in *emotional intelligence*, a construct that encompasses how people can respond strategically to and leverage their emotional experiences in productive ways, regardless of how adaptive or ill-suited to a modern-day situation an emotional reaction might initially be (Brackett, Rivers, and Salovey 2011; Salovey and Mayer 1990). While LLMs cannot possess complete emotional intelligence, their stochastic outputs can be evaluated in two of the four domains of emotional intelligence as posited by Salovey and Mayer: accurately perceiving what emotion is being expressed by a human via their written expressions of emotional disclosures; and demonstrating accurate analysis of what a human is likely experiencing given contextual cues (such as a person's situation, appraisals of that situation, and/or corresponding bodily responses). The extent to which frontier foundation models can perform such tasks is an empirical question. As shown

*These authors contributed equally to this work.

in Figure 1, the LLM's predicted emotions fail to align with self-disclosed emotions when asked to provide completions to various human experience prompts.

Accurate emotion recognition has the potential to substantially enhance a broad spectrum of natural language processing (NLP) tasks. By integrating fine-grained affective understanding, dialogue systems can become more emotionally aware, thereby improving their capacity to comprehend and generate human-like emotions (Liu et al. 2021b). Moreover, emotion recognition facilitates quantitative analyses of social dynamics, such as political discourse, customer service, and public opinion mining. Finally, it also enables critical NLP-driven applications, including automated depression screening (Jurafsky and Martin 2024).

Substantial research efforts have been made to study the emotion recognition capabilities of LLMs. These studies usually focus on constructing various emotion recognition benchmarks from different data sources, domains, and emotion theories (Strapparava and Mihalcea 2007; Chatterjee et al. 2019). However, current benchmarks face three main **limitations**. First, little attention has been given to the self-disclosure of emotional experiences, with existing research typically relying on crowdsourcing or expert annotations to label emotions in text, which might lead to unreliable evaluations (Singh, Caragea, and Li 2024; Sabour et al. 2024). Second, these studies are often limited by a short, predefined list of emotions, restricting their ability to capture fine-grained nuances in emotional experiences (Feng et al. 2022). Third, current benchmarks primarily focus on short sentences, contexts, or dialogues, neglecting the self-disclosure of emotions in longer contexts (Demszky et al. 2020).

To address this gap, we construct a new benchmark, EXPRESS (**EX**periences and **PR**ocessed **E**motions in **S**elf-disclosure **S**tories), which consists of 33,679 human experiences and their associated self-disclosed emotions. The assessment framework is illustrated in Figure 1. In this framework, we mask emotion words in human self-disclosure texts and prompt the LMs to predict the masked words. To move beyond predefined emotion classes and evaluate models' emotion recognition capabilities at a finer granularity, we decompose the emotions into 10 dimensions of basic emotions and sentiments based on Plutchik's Wheel (Plutchik 1980; Mohammad and Turney 2013). Using EXPRESS, we benchmark the emotion recognition performance of 14 language models.

Our results reveal that the emotion recognition capabilities of LMs correlate with model size, model family, model architecture, and prompting strategies. While the accuracy of predicting emotions at the lexical level is low, some LMs demonstrate the ability to capture the basic emotions underlying compound emotions, even when the predicted emotion word does not exactly match the label. Furthermore, we find that LLMs are capable of in-context learning for emotion recognition when provided with examples. We also find that, under the best settings, LLMs are able to generate emotions consistent with theoretical definitions, but they are sometimes less contextually "aware" than the emotions self-disclosed by humans. These findings are particularly valuable for the development of emotion-aware AI systems, especially in future mental health applications (Hua et al. 2025; Ji et al. 2022; Adhikary et al. 2024).

This paper makes the following contributions:

- We present EXPRESS, a novel benchmark designed for emotion recognition that serves as a resource for evaluating models' emotion recognition capabilities and facilitating potential emotion alignment.
- We propose a multi-faceted emotion recognition evaluation framework that encompasses multiple metrics: lexical accuracy, decomposed emotion vector accuracy, and contextual accuracy.
- Our framework incorporates various prompting strategies, including few-shot learning and Chain of Thought prompting, to evaluate models' emotional reasoning and in-context learning capabilities. We release our code and dataset on GitHub: https://github.com/Computing-for-Social-Good-CSG/express-emotion-recognition.git

Related Work

Emotion Taxonomy

While emotion theorists have nuanced definitions of emotions that may differ in their descriptions of emotional features, most agree that emotions are functional in preparing us to respond to perceived or real environmental stimuli (Gross and Feldman Barrett 2011). Theories of emotion define discrete emotional experiences or reactions to events (real or imagined) as representations that may include correlated physiological responses, appraisal processes, subjective feelings, and action tendencies (Schiller et al. 2024).

Many theories of emotion are studied in the field of affective science, but computational models of emotion in NLP have primarily been based on two families of theories. The first views emotions as fixed atomic units, often referred to as basic emotions, while the second conceptualizes emotion as existing within a multidimensional space (Jurafsky and Martin 2024). Within the family of fixed atomic unit theories, one prominent theory proposes six basic emotions: surprise, happiness, anger, fear, disgust, and sadness (Ekman et al. 1999). Another theory, known as the Plutchik Wheel of Emotion, posits that emotions consist of eight basic emotions in four opposing pairs: joy-sadness, anger-fear, trust-disgust, and anticipation-surprise (Plutchik 1980).

Recent advances in psychology have introduced new conceptual and methodological approaches to capturing the more complex "semantic space" of emotion, which aligns with the second family of theories (Cowen et al. 2019). These models typically use two dimensions—valence and arousal—to describe emotions (Russell 1980). Some models further include a third dimension, dominance, to provide additional nuance in describing emotions (Fontaine et al. 2007).

Emotion Recognition Benchmarks

Substantial datasets for emotion recognition exist, each with varying emotion label domains. Table 1 summarizes key characteristics of existing datasets, including size, number of emotion labels, average context length, and annotation methods.

Work/Dataset	Source	Domain	Size	No. Emotions	Avg Context Length	Topic Diversity	Annotator
GoEmotions (Demszky et al. 2020)	Reddit	General Reddit Comments	**58,009**	27	13	Unknown	Crowdsourced
EmoTrigger (Singh, Caragea, and Li 2024)	CancerNet, Twitter, Reddit	Health-related, Natural disasters, General Reddit posts	900	8	22	5	Expert
EmoWOZ (Feng et al. 2022)	Amazon MTurk	Human-human + human-machine conversations	11,000	7	187 / 13	7	Expert
SemEval Task #14 (Strapparava and Mihalcea 2007)	NYT, CNN, BBC, Google News	News Headlines	1,250	6	7	13	Trained
EmoContext (Chatterjee et al. 2019)	Twitter	Conversational agent interactions	38,424	4	23	Unknown	Trained
ISEAR (Scherer and Wallbott 1994)	Survey	Personal narratives on major emotional events	7,665	7	23	31	**Self-reported**
DailyDialogue (Li et al. 2017)	Various Websites	Everyday multi-turn dialogues	13,118	6	115 / 15	10	Expert
EmoBank (Buechel and Hahn 2017)	MASC + SemEval-2007	News headlines, blogs, fiction, etc.	10,548	9	15	33	Crowdsourced
CrowdFlower (Van Pelt and Sorokin 2012)	Twitter	General Twitter tweets	39,740	13	36	9	Crowdsourced
EmoInt (Mohammad and Bravo-Marquez 2017)	Twitter	General Twitter tweets	7097	4	16	46	Crowdsourced
DENS (Liu, Osama, and De Andrade 2019)	Gutenberg, Wattpad	Long-form English narratives	9,710	8	86	7	Crowdsourced + Expert
Emotion-Stimulus (Ghazi, Inkpen, and Szpakowicz 2015)	FrameNet	Blogs	820	9	18	28	Expert
Tales Emotion (Alm and Sproat 2005)	Grimm Fairy Tales	Blogs	15,302	8	21	10	Trained
XED (Öhman et al. 2018)	OPUS	Movie subtitles	33,548	8	9	Unknown	Crowdsourced
EXPRESS (ours)	Reddit	Reddit Posts	33,679	**251**	**259**	**49**	**Self-disclosed**

Table 1: Comparison of emotion datasets in terms of size, domain, number of emotions, average text length, topic diversity, and annotation method.

Emotion label coverage is often limited. Most datasets include only a small set of predefined emotion categories. For example, Emotion-Stimulus (Ghazi, Inkpen, and Szpakowicz 2015), Tales Emotions (Alm and Sproat 2005), and SemEval Task 14 (Strapparava and Mihalcea 2007) rely on Ekman's six basic emotions (Ekman et al. 1999), sometimes supplemented with a few additional labels. XED (Öhman et al. 2018) and EmoTrigger (Singh, Caragea, and Li 2024) use Plutchik's eight primary emotions. These limited sets of emotion labels, focused on a coarse level of classification, restrict the ability to study fine-grained emotional nuances.

Context lengths tend to be short. Many datasets are built on platforms such as Twitter or Reddit comments, or focus on dialogue utterances and news headlines. As a result, the average context length for emotion recognition typically ranges from 10 to 36 words, with CrowdFlower reaching the upper bound at 36 words (Van Pelt and Sorokin 2012).

Dataset sizes vary. While small datasets like Emotion-Stimulus contain under 1,000 examples, larger resources such as EmoContext (Chatterjee et al. 2019), DailyDialogue (Li et al. 2017), and GoEmotions (Demszky et al. 2020) include tens of thousands of samples.

Annotations are typically crowdsourced or expert-labeled. Many benchmarks rely on either expert labeling or crowdsourcing, where annotators infer emotions from external text. ISEAR (Scherer and Wallbott 1994) is an exception that uses self-reported emotion events, but its ecological validity is limited by its collection method: participants were asked to describe experiences for a fixed list of seven predefined emotions, constraining natural emotional expression and overlooking subtle or more complex feelings. In addition, relying on external annotations limits the granularity of emotion labels.

Prior research provides a foundation for assessing the emotion recognition capabilities of LLMs, though it faces the limitations mentioned above. Our work addresses these issues by scaling up context length to an average of 259 words and the range of emotion labels to 251. We use self-disclosed emotions as ground truth labels without external annotation because they are considered ecologically valid. As naturally occurring disclosures, they allow individuals to freely and authentically share their internal experiences, including emotional reactions to past events, without being constrained by predefined categories or researcher-led methods (Pennebaker and Beall 1986; Frattaroli 2006; Davitz 2013). Moreover, self-report remains a cornerstone of methods for empirically investigating subjectively felt emotional experiences (Mauss and Robinson 2009), further supporting the use of self-disclosed emotions as ground truth. Additionally, our framework allows models to generate context-based, non-predefined emotions, ensuring sufficient variation in emotional nuances. Together, these improvements establish our dataset as an ecologically valid and fine-grained benchmark for evaluating emotion recognition capabilities in language models.

EXPRESS: A Comprehensive Benchmark for Emotion Recognition

Selection of an Emotion Lexicon. The Berkeley Well-Being Institute synthesized a complete list of 271 emotions (Davis 2024) based on multiple emotion theories: Discrete, Circumplex (Russell 1980), Plutchik's Wheel (Plutchik 1980), and other emotion theories. Instead of selecting a single emotion theory, we used the Berkeley Well-Being list of emotions because it combines multiple theories and represents the largest available emotion lexicon. A primer on the emotion theories used in this paper is provided in the Appendix.

Collecting self-disclosed experiences and emotions. Our primary objective is to assess whether LMs can predict emotions based on real-life nuanced experiences. To achieve this, we created prompts that embed a self-disclosed emotion.

Experience Contextualized Prompts	Self-disclosed Emotion
Feeling <mask> after getting a role in a movie. [...], now am scared cause the director doesn't know that I got no acting experiences or skills (am very bad at memorizing my lines) if he finds out he won't hire me for the movie. I don't know what to do.	afraid
How do I do this... I feel <mask>. Battling so many health issues right now, mostly gi related. [...] I am incredibly incredibly <mask> and going through 1 to 5 juul pods a day... I feel like there's no way out of this. Any advice? I can't imagine my depression and fatigue getting worse...	panicked, depressed

Table 2: Examples of naturally occurring emotionally-centered prompts in EXPRESS. Self-disclosed emotions (ground truth) are replaced by <mask> token.

To ensure we evaluate LMs in scenarios that mirror actual language usage, we construct our prompts from natural contexts that we retrieve from Reddit, rather than crowdsourcing prompts.

Because of its pseudonymity, Reddit is a popular platform for discussing real-life experiences. Reddit served as the primary data source due to its support for longer posts (up to 40,000 characters) which enabled the collection of rich and nuanced human experiences and the corresponding evoked emotions. The Reddit API Praw (Boe 2021) [1] was utilized to collect posts from all subreddits containing at least one emotion from the Berkeley Well-Being list.

Emotion Masking. To mask the self-disclosed emotions in the collected posts, we designed a comprehensive regular expression protocol, as not all emotion keywords in a post are self-disclosed by the author. For example, the author might use emotion keywords to describe external events or other people's feelings rather than their own. Our protocol includes three main patterns: 'I feel + emotion', 'I am + emotion', and 'no-pronoun + feeling + emotion'. To make the algorithm robust to variations in natural language phrasing, we designed a series of rules, with details provided in Appendix.

We included the pattern 'feel' because prior work indicates that humans use the word 'feeling' interchangeably with 'emotion', even though feelings and emotions are not the same (Davis 2024). Feelings encompass both emotional experiences (e.g., feeling sad) and physical sensations (e.g., feeling hungry). This distinction justifies our use of the following pattern-matching formats: (*I + feel/am + emotion*) and (*no-pronoun + feeling + emotion*).

Due to the context window size limitations of some language models and the fact that some posts are extremely lengthy, we segmented the posts into chunks of 512 tokens. During the segmentation process, we ensured that the context surrounding the target masked emotions was maximized. If multiple masked emotions existed, they were grouped based on their relative positions in the text. Table A.1 in Appendix outlines the algorithm used to perform post segmentation.

The resulting dataset (EXPRESS) comprises 33,697 posts with an average word count of 259 per post. EXPRESS posts originate from 6,930 unique subreddits and span the time period from June 2009 - April 2024. More details of the dataset are provided in the Appendix (Table A.2, Table A.3, Table A.4). Across the dataset, a total of 52,632 emotion words were identified covering 251 (92.62%) out of 271 Berkeley Well-Being emotions. We create prompts from EXPRESS by replacing the original emotion with a <mask> token. Table 2 depicts examples from our dataset.

[1]BSD 2-Clause License: licensed under a permissive license allowing redistribution and modification with the retention of copyright and disclaimer notices

Evaluating Emotion Recognition Capabilities of LLMs

Using EXPRESS, we measured the emotion recognition capabilities of 14 prevalent language models including four masked language models, three Seq2Seq language models, and seven causal language models.

Model Details

Using our dataset, we evaluated several variants of open-source and closed-source language models widely used in current research. We included four prevalent masked language models, as they are specifically designed for masked language modeling tasks. These models are RoBERTa-base (Liu et al. 2019), Longformer (Beltagy, Peters, and Cohan 2020), Mental-RoBERTa (Ji et al. 2022), and Mental-Longformer (Ji et al. 2023), the latter two of which have been further pretrained on mental health-related corpora. For Seq2Seq language models, we included three models from the Flan-T5 family (large, XL, and XXL) (Chung et al. 2024). For causal language models, we focused on instruction-tuned models of varying sizes, as they are fine-tuned to follow instructions. These include Llama3.1-Instruct (8B, 70B) (Grattafiori et al. 2024), Gemma2-Instruct (2B, 9B, 27B) (Team et al. 2024), GPT-3.5-turbo, and GPT-4o (OpenAI et al. 2024). The temperature was set to 0.0 for all experiments to minimize the effects of randomness.

Experimental Setup

We evaluated the performance of the four masked language models by directly filling in the masked emotions. For the remaining 10 models, we prompted them to predict the masked emotions based on the context. We designed the experiments with four different settings: zero-shot, few-shot with 4 random examples, few-shot with 4 nearest examples, and Chain-of-Thought (CoT) prompting (Wei et al. 2022).

The zero-shot setting served as the basic test of the emotion recognition ability of LLMs based on self-disclosed emotional experiences. Models were directly instructed to predict the <mask> word with an emotion based on the context. The prompt template used in this setting is detailed in Appendix (Table A.9).

To investigate whether LLMs can enhance their emotion detection ability by learning from examples, we included two

few-shot settings. In both settings, we used four examples, as prior work has shown that using a larger number of exemplars does not significantly improve model performance (Min et al. 2022). Additionally, we ensured that the number of <mask> tokens in the exemplars matched that of the target post to avoid confounding effects. The first few-shot setting used four examples randomly selected from EXPRESS, while the second used the Bert-base-uncased model (Devlin et al. 2019) to compute sentence embeddings and applied Euclidean distance to find the four nearest examples to the Reddit posts (Liu et al. 2021a). By comparing these two settings, we aimed to explore whether providing similar experiences in the examples could further enhance the models' ability.

Studies have shown that CoT prompting improves performance across a range of arithmetic, commonsense, and symbolic reasoning tasks (Wei et al. 2022). However, some studies have also suggested that CoT prompting may not enhance performance in socially sensitive domains, such as addressing harmful questions (Shaikh et al. 2023). In this work, we included the CoT prompting setting to examine whether it could further improve the models' ability to predict emotions.

Measuring Accuracy of Emotion Recognition

To evaluate LMs' emotion recognition capabilities on a fundamental level, beyond calculating lexicon accuracy, we adopted the approach of the NRC Emotion Lexicon (EmoLex) (Mohammad and Turney 2013). EmoLex is a widely used resource that analyzes 14,182 unigrams and associates these unigrams, through crowdsourcing, with eight basic emotions—anger, anticipation, disgust, fear, joy, sadness, surprise, and trust—as well as with positive and negative sentiment. The associations are represented as binary scores (0 or 1), indicating whether a word is linked to a particular emotion or sentiment. We leveraged EmoLex to construct vector representations for each predicted and actual emotion. These vectors, as shown in Figure 1, are 10-dimensional: 8 dimensions correspond to the basic emotions, and 2 represent positive and negative sentiment. By converting all words into these 10-dimensional vectors, we evaluated the model-predicted emotions against the self-disclosed emotions on a basic emotion and sentiment level.

During this process, some model-generated emotions were not included in the NRC Emotion Lexicon. To address this, we replicated EmoLex's crowdsourcing task on Amazon Mechanical Turk (AMT) to generate vector representations for these additional emotions. Further details are provided in Appendix (Figure A.2).

We evaluated the results using three metrics: (1) Lexical Accuracy (Acc_L), defined as the percentage of exact lexical matches (N_{lm}) from all masks (N): $Acc_L = N_{lm}/N$; (2) Vector Accuracy (Acc_V), defined as the percentage of exact vector matches (N_{vm}) between the 10-dimensional basic emotion vectors for the predicted and actual emotions across all masks: $Acc_V = N_{vm}/N$; and (3) Average Vector F-1 score ($F1_V$), which balances precision and recall to evaluate the model's ability to predict each dimension of the 10-dimensional emotion vector. More details on the F1-score calculation for vectors are provided in the Appendix section titled Evaluation Metrics.

We included Acc_V as a metric because predicted and actual emotions may not align lexically due to the diversity of emotion vocabulary but could still match at the basic emotion level. For example, 'angry' and 'furious' share the same emotion vector but are two different labels lexically. A higher Acc_V indicated greater alignment between the actual and predicted emotions. Similarly, $F1_V$ was included to assess how closely the predicted emotions approximated the self-disclosed ones, even when there was no exact match across all dimensions.

Results

Here, we present our findings on the emotion recognition capabilities of LLMs evaluated on the EXPRESS dataset.

Fine-Grained Emotion Alignment is Challenging for LMs. Models demonstrated significant variability in their ability to predict emotions from emotional experiences in the zero-shot setting, as shown in Table 3. Acc_L ranged from 0.051 to 0.318, while Acc_V, slightly higher, ranged from 0.097 to 0.388. $F1_V$ ranged from 0.434 to 0.711, compared to a baseline of randomly generated vectors at 0.322. Overall, the results show that it is challenging to predict human self-disclosed emotions from emotional experiences. The substantial number of emotion words, along with the similarity and overlapping nature of some emotion terms, may contribute to the low Acc_L. However, Acc_V, which evaluates the decomposed vectors of emotion terms, remains relatively low, increasing by only about 0.06 on average. This indicates that for most models, language models fail to align with human self-disclosed emotions on at least one basic emotion or sentiment dimension in the majority of predictions.

Some models, such as Flan-T5-large, XL, and Gemma2-2B, have Acc_V around 0.1, meaning they can only accurately predict around 10% of the emotions. Their $F1_V$ ranges from 0.4 to 0.5, which, although higher than randomly generated vectors, indicates their limited ability to correctly predict emotions. On the other hand, some models demonstrate relatively better emotion recognition ability, including the four masked language models, Llama-3.1-70B, Gemma-2-27B, and GPT-4o. These models achieve Acc_V over 0.3 and $F1_V$ exceeding 0.6, indicating a certain degree of alignment with self-disclosed emotions or, at the very least, some aspects of them.

Model Family and Size Matter. Although the tested language models do not achieve either Acc_L or Acc_V higher than 0.4 under zero-shot settings, the emotion recognition performance varies significantly between models, as shown in Figure 2. Three factors appear to have the most significant impact on performance: **Model Architecture**, **Model Family**, and **Model Size**.

The four masked language models are all ranked among the top seven models in both accuracy metrics, despite having far fewer parameters (RoBERTa with 125M and Longformer with 149M). This may be due to their specialization in mask-filling tasks: they are specifically designed and trained to

Model Architecture	Model		Zero-shot			Few-shot (random)			Few-shot (nearest)			CoT		
	Name	Variant	Acc_L	Acc_V	$F1_V$	Acc_L	Acc_V	$F1_V$	Acc_L	Acc_V	$F1_V$	Acc_L	Acc_V	$F1_V$
Masked LMs	RoBERTa	base	**0.318**	0.369	0.658	-	-	-	-	-	-	-	-	-
		mental	0.313	0.366	0.654	-	-	-	-	-	-	-	-	-
	Longformer	base	0.309	0.358	0.645	-	-	-	-	-	-	-	-	-
		mental	0.277	0.330	0.633	-	-	-	-	-	-	-	-	-
Seq2Seq	Flan-T5	large	0.051	0.097	0.434	-	-	-	-	-	-	-	-	-
		xl	0.063	0.111	0.471	-	-	-	-	-	-	-	-	-
		xxl	0.102	0.167	0.532	-	-	-	-	-	-	-	-	-
Causal LMs	Llama-3.1	8B	0.149	0.222	0.591	0.175	0.248	0.619	0.195	0.265	0.624	0.153	0.228	0.586
		70B	0.264	0.338	0.675	0.265	0.343	0.677	0.279	0.356	0.683	0.198	0.272	0.650
	Gemma-2	2B	0.078	0.147	0.530	0.105	0.174	0.550	0.122	0.193	0.557	0.052	0.110	0.514
		9B	0.223	0.293	0.639	0.268	0.338	0.673	0.278	0.347	0.677	0.152	0.200	0.634
		27B	0.268	0.341	0.667	0.290	0.358	0.675	0.298	0.366	0.678	0.154	0.211	0.636
	GPT	3.5-turbo	0.217	0.285	0.629	0.247	0.313	0.647	0.250	0.315	0.647	0.176	0.244	0.602
		4o	0.313	**0.388**	**0.711**	0.350	0.422	0.732	**0.364**	**0.436**	**0.738**	0.310	0.383	0.723
	Average (Causal)		0.210	0.272	0.647	0.242	0.314	0.653	0.255	0.325	0.658	0.171	0.235	0.621
Baselines	Random		-	0.001	0.322	-	-	-	-	-	-	-	-	-
	All-zeros		-	0.037	0.000	-	-	-	-	-	-	-	-	-

Table 3: Performance comparison of language models across different evaluation settings. Metrics include lexical accuracy (Acc_L), vector accuracy (Acc_V), and average vector F1 ($F1_V$) across zero-shot, few-shot (random and nearest), and CoT settings.

fill masks and do not need to interpret complex prompts as Seq2seq or causal language models do. In Seq2seq and causal language models, the model family plays a crucial role. For example, with similar model sizes, Flan-T5-xxl, Llama3.1-8B, and Gemma-2-9B exhibit vastly different Acc_L, Acc_V, and $F1_V$ scores. Flan-T5-xxl and Gemma2-9B have differences of 0.156, 0.174, and 0.135 in Acc_L, Acc_V, and $F1_V$, respectively. GPT-3.5-turbo, despite having 175B parameters, performs worse than smaller models such as Llama-3.1-70B and Gemma-2-27B.

Another key factor influencing performance is model size. Within all model families—Flan-T5, Llama-3.1, Gemma-2, and GPTs—the ability to predict emotions improves consistently as the number of parameters increases, without exception. Interestingly, the best-performing causal language model, GPT-4o, with 1.75T parameters, performs similarly to the three masked language models. This highlights the significant gap in mask-filling ability between masked language models and causal language models in emotion recognition tasks. The Wilcoxon Signed-Rank tests were conducted, and the results are shown in Table A.11.

Chain of Thought Doesn't Help Models Predict Emotions. Studies show that CoT prompting enables models to reason step by step to arrive at the final answer and has been proven to improve performance on various NLP tasks (Wei et al. 2022). Given that most emotional experience texts in our dataset are long (259 words per post on average), leading models to reason step by step might be a potential way to achieve higher performance. To test this, we adapted CoT prompting by instructing models to "Think step by step" (Kojima et al. 2022). The prompts we used are shown in the Appendix (Table A.9).

However, as Table 3 shows, for all seven tested models across three model families and various sizes, performance consistently worsened, with average decreases of 0.039, 0.037, and 0.026 for Acc_L, Acc_V, and $F1_V$, respectively. The only models unaffected were GPT-4o and Llama-3.1-8B. Notably, for the three Gemma-2 models, regardless of size, the models' ability to respond in the correct format as instructed deteriorated significantly, dropping from an average of 99.9% to 76%. The Wilcoxon Signed-Rank tests were conducted and the results show that all models, except for these two, perform statistically significantly worse under CoT settings ($p < 0.001$). For smaller models, such as Gemma-2-9B, CoT prompting sometimes caused them to deviate from the original instruction and fail to respond with emotion words. For larger models, while CoT prompting did not reduce the rate of valid responses, overall performance declined. This aligns with prior findings (Shaikh et al. 2023), suggesting that CoT may underperform in socially situated tasks. It also echoes recent work (Chochlakis et al. 2025), which found that CoT fails to improve outcomes in complex, context-sensitive, and subjective tasks, particularly for larger models, which tend to rely heavily on their built-in prior knowledge, potentially leading them to overlook or disregard the specific context provided in the prompt.

Error Analysis. To better understand the details of the prediction errors, we conducted an error analysis on the best-performing causal language models: Llama-3.1-70B-Instruct, Gemma-2-27B-It, and GPT-4o. We observed a significant difference between the distribution of emotion words used in human self-disclosures and those predicted by the models. Humans tend to use emotion words such as happy, scared, sad, tired, and embarrassed, whereas the models frequently overused emotion words such as anxious, grateful, overwhelmed, ashamed, frustrated, and relieved, as shown in Figure A.3 in the Appendix.

Among the most common mispredictions, errors frequently occur when the model predicts a similar but distinct emotion, or one with a different intensity. For example, models often predict grateful, a deeper and more enduring emotional expression, when the true label is thankful. They also tend to overuse words like frustrated and anxious to represent a wide range of emotional experiences, whereas humans express

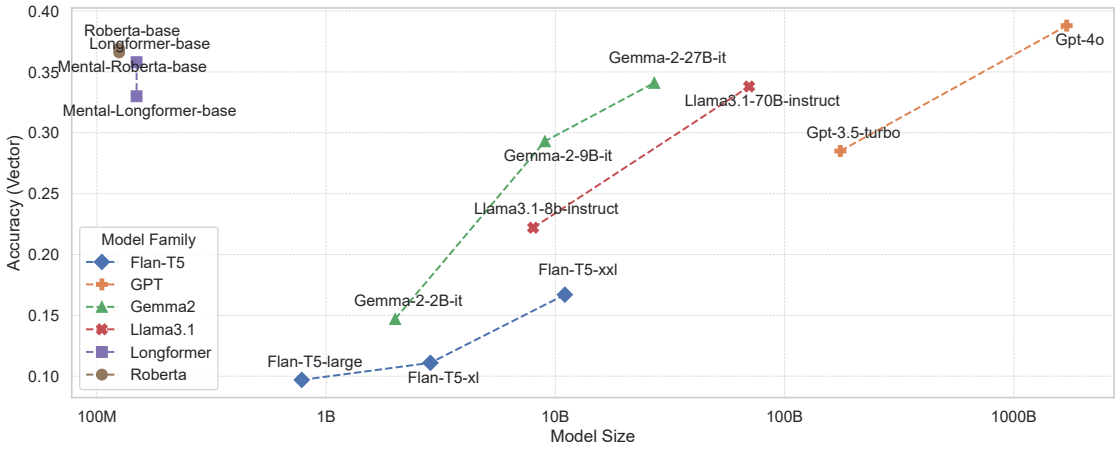

Figure 2: A comparison of model size, family, and emotion detection vector accuracy. The results show that model performance is significantly influenced by the model family and generally improves with increasing model size across four causal language model families.

these emotions more diversely and subtly in these cases using words like disheartened, demoralized, irritated, annoyed, restless, panicked, agitated, and stressed. Additionally, models often reduce emotional intensity by predicting angry instead of furious, or afraid instead of terrified. Table A.6 in the Appendix presents the normalized mispredictions commonly made by these three models.

Impact of Dataset Segmentation on Model Performance When constructing EXPRESS, we segmented long posts into segments of 512 tokens to ensure a fair and consistent evaluation across LLMs with different context length. However, this segmentation may limit the ability of LLMs with extended context windows, like GPT-4o to fully utilize their contextual reasoning capabilities, potentially underestimating their performance. To assess this trade-off, we conducted an additional analysis comparing model performance on segmented versus full posts. We randomly sampled 1,000 posts from the EXPRESS and evaluated GPT-4o, our best-performing model with a 128k-token context window, on both settings.

As Table 4 shows, despite a substantial increase in post length (approximately 1,000 additional words), performance remained effectively unchanged across all metrics. This suggests that a 512-token context window provides sufficient context for models to make comparable decisions on this task. Post segmentation not only ensures a fair and consistent evaluation across LLMs with different context lengths, but also does not disadvantage models with larger context capacities in this setting.

Setting	Avg Post Length (words)	AccL	AccV	F1V
Segmented Post	353	0.348	0.404	0.717
Full Post	1353	0.346	0.406	0.715

Table 4: Comparison of GPT-4o's performance on segmented vs. full posts.

Are Models Good Learners of Emotions?

Our results in the previous section indicate that accurately predicting self-disclosed emotions based on emotional experiences remains a challenge for language models, both at the lexical level and the basic emotion vector level. Even the best-performing models, including the four masked language models, Llama-3.1-70B-Instruct, Gemma-2-27B-Instruct, and GPT-4o, incorrectly predict at least one basic emotion dimension in the emotion vector for nearly 65% of instances.

However, most language models are neither designed nor trained specifically for emotional intelligence tasks, such as emotion recognition. As a result, they may lack the implicit capacity to predict emotions in a human-like manner. Therefore, it is crucial to examine whether LLMs can learn and improve their emotion detection capabilities. Demonstrating this potential would highlight their utility in assisting mental health-related tasks, particularly when specifically designed and trained for such applications. To evaluate this, we established two few-shot experimental settings.

The two few-shot settings use different strategies to select examples. The first setting is designed to examine whether LLMs can learn from random examples of emotions, serving as a baseline. The second few-shot setting is designed to assess whether LLMs learn better when provided with examples of similar emotional experiences. To achieve this, we use the BERT-base-uncased model (Devlin et al. 2019) to compute sentence embeddings and apply cosine similarity to identify the four nearest examples to the test query as few-shot examples.

Result As shown in Table 3 and Figure 3, all tested LLMs demonstrate improved emotion recognition ability when exposed to random examples. The average improvements in Acc_L, Acc_V, and $F1_V$ are 0.032, 0.042, and 0.006, respectively. Similar to, but even better than, the first few-shot setting, the few-shot setting with the nearest distanced examples further boosts the models' performance on emotion recognition. This setting outperforms the first setting for all models,

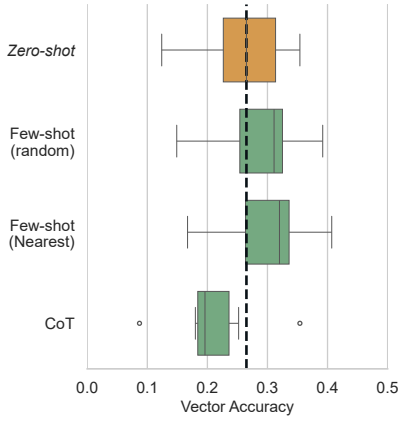

Figure 3: A comparison of zero-shot results and the other three settings. Models learn from examples and perform better compared to the zero-shot setting.

with average improvements in Acc_L, Acc_V, and $F1_V$ of 0.045, 0.053, and 0.011, respectively, compared to the zero-shot setting. This indicates that models can learn better when exposed to examples with similar emotional experiences. The improved results of the two few-shot settings demonstrate the ability of models to learn from provided examples, indicating their potential to become more emotionally aware in future training and adapt to mental health applications.

Human Evaluation of Predicted and Actual Emotions

To contextualize our empirical evaluation of LMs' emotion recognition capabilities across the previous experiments, we conducted a qualitative analysis of a subset of the predicted (by LLMs) and actual emotions. Our first research objective was to *assess which emotion (predicted/actual) would be picked by a domain emotion expert as being more plausible, or consistent with academic theory and empirical research about emotion experience*.

Our second research question was: *"How might a domain expert in emotion research determine that a certain emotion fits the natural contexts in the posts (e.g., select an LLM-generated vs. ground truth emotion as being accurate)?"* Essentially, we aim to compute a form of **contextual accuracy** to assess whether the predicted or ground truth emotion aligns with the local context of the text. Hence, we perform a qualitative analysis on a small subset of our dataset to compute this contextual accuracy, utilizing human evaluation by three emotion experts, whose backgrounds are detailed in the Human Evaluation section of the Appendix.

We framed this task as a Turing test (Turing 2009), aiming to see if an LLM can mimic human emotional intelligence by attempting to deceive the emotion experts into selecting the LLM's output over the self-disclosed emotion. To maximize the challenge, we utilize the best-performing model's outputs (GPT-4o with few-shot-nearest setting) for this task.

Setup

Our sample includes 213 EXPRESS posts, the associated self-disclosed emotions, and the corresponding predictions from the best-performing model. We selected the posts where the emotions predicted by the LLM differ from the self-disclosed emotions at the vector level. We randomized the order of presenting predicted or actual emotions first, and asked the coders to read each post and select one of the following options without knowing which options were predicted or actual: (1) self-disclosed emotion (SD), (2) LLM-generated emotion (LLM), (3) both emotions equally fit (BOTH), and (4) neither emotion fits (NEITHER). For clarity, it is important to note that selecting one emotion does not necessarily imply that the other option is unsuitable; rather, it indicates that the chosen emotion is preferred by the coders.

In the next phase, one coder went through all samples and provided rationales for their selections. Open coding (Khandkar 2009) was then conducted on the rationales to understand what the common themes were to justify the selections made, as presented in Table A.7 in Appendix. The codes were then presented to the other coders, who were also free to add additional codes; however, no additional codes were introduced.

Results

Two coders selected LLM over SD more frequently, while the third coder selected SD more often. We used a majority vote to aggregate the results from the three coders. Overall, SD was selected 89 times (40.0%), whereas LLM was selected 97 times (43.7%). The coders also considered BOTH emotions plausible 5 times (4.7%), and there were 22 instances (11.6%) where all three chose different options. However, given the subtle distinctions between emotions, agreement among the coders was relatively low, with only 36.2% of the samples receiving unanimous selection and a Fleiss' Kappa score of 0.21 (fair agreement).

We aggregated the codes from the three coders and reported the frequency of codes in Table A.7 in the Appendix. The most common reason for coders selecting the SD emotions was that the selection better matched other terms providing context for the affective experience (Code 3, N=101). The second most common reason was that the selection was judged as better fitting the degree of specificity in the affective experience being described (Code 2, N=39). For the LLM emotions, the most common reasons were similar (Code 2, N=69; Code 3, N=58, and Code 5, N=57). However, when choosing SD, Code 3 was significantly more frequent (36.4%), indicating that coders believed SD emotions were more contextually appropriate in many cases. In contrast, when choosing LLM, Code 2 and Code 5 were significantly more frequent (23% and 19%), suggesting that coders preferred LLM-generated emotions due to their higher specificity in describing the affective experience and better alignment with emotion definitions and theories. Examples of code selections and the corresponding rationales are presented in Table A.8.

Our results unexpectedly show that experts slightly preferred the LLM's predicted emotions over the self-disclosed emotions. The low agreement between coders further reveals

that both SD and LLM-generated emotions are plausible, differing primarily in nuanced ways. This finding highlights the inherent complexity and subtlety of our task, where human understanding and expression of emotions are highly subjective, shaped by an individual's unique experiences, personality, emotion granularity, and perspective. This pattern of findings, across the three coders, suggests that while our best-performing LLM is able to generate emotions consistent with theoretical definitions and convincing as being appropriate, there are also instances where LLM may miss important contextual cues in the excerpts that may seem intuitive to human coders. However, it is important to note that these findings are specific to the best-performing model under optimal settings, which may represent the upper bound of LLM performance. Further investigation is needed to explore whether, in such cases where a human coder can articulate how an important clue in an excerpt is "missed" by the LLM, the LLM is making "errors" (such as by not accounting for certain information) or registering the contextual information but making predictions based on different information within the excerpt, potentially relying on "internal world modeling" specific to the LLM.

Conclusion

Higher emotional understanding and prediction abilities in LLMs are crucial for empathetic interactions (Mayer and Geher 1996), as users often prefer models that align with their beliefs (Kirk et al. 2024). Misjudging or failing to respond empathetically in dialogue can lead to user discomfort (Ball and Breese 2001). However, existing benchmarks are limited by unreliable labels, a narrow range of emotion categories, and short emotional contexts.

To address these gaps, we constructed EXPRESS, an emotion recognition dataset created by masking self-disclosed emotions in Reddit posts. Our systematic evaluation revealed that LLMs still face challenges in aligning with human emotional expressions. Performance varied across model architectures and families, with consistent improvements as model size increased. Notably, while model performance varied among model families, masked language models performed comparably to some larger causal language models, such as GPT-4o. Given their significantly smaller model sizes, they offer a cost-effective alternative with similar performance.

We also tested whether CoT prompting improves LLM performance. Our findings, consistent with prior work (Shaikh et al. 2023; Chochlakis et al. 2025), show that CoT degrades performance in subjective tasks, possibly due to models relying too heavily on prior knowledge instead of contextual cues. Few-shot prompting, however, showed promise, indicating that with targeted design and exposure, LLMs can improve their emotion recognition performance even without explicit emotion-task training.

For the qualitative analysis, future research could explore how the affective terms predicted by LLMs differ from self-disclosed emotions and how expert observations might be used to fine-tune models for improved emotion recognition. The analysis also shows that LLM-generated emotions were preferred by experts half of the time, suggesting that GPT-4o, under the best settings, has the ability to generate reasonable emotions that fit the context. However, the error analysis and qualitative analysis reveal that while LLM emotions are sometimes more specific than SD emotions and consistent with emotion theories, they can also be overly general, frequently predicting common emotions such as 'anxious' or 'frustrated.' Moreover, LLMs are sometimes less effective at capturing contextual cues than SD emotions. These findings, along with the low alignment with SD emotions, highlight the importance of self-disclosed emotions in fine-grained emotion recognition tasks, which not only serve as a benchmark for evaluation but also provide valuable training material to improve model alignment.

Our benchmark and evaluation framework offer a systematic way to assess LLMs' capacity to understand and predict fine-grained, self-disclosed emotions. While our setup is intentionally controlled, it provides foundational insights into how models handle nuanced emotional expressions, a prerequisite for deployment in sensitive, real-world applications such as mental health support tools or social media moderation. As these applications demand accurate emotion recognition, models that underperform in our benchmark may require further careful examination before being reliably applied in such contexts.

Limitations

Diversity in emotion expression. While our research provides valuable insights into emotion recognition, it is primarily focused on neurotypical ways of expressing emotions. This limitation highlights the need for further research to explore and understand how emotional expressions may differ in neurodivergent populations, as differences in emotional expression are well-documented (Trevisan et al. 2017). Expanding the scope of future studies to include a more diverse range of emotional expressions will help create more inclusive models and improve the accuracy and applicability of emotion recognition systems across different populations (Mazefsky, Pelphrey, and Dahl 2012).

Bias in User Demographics. Reddit's user base is not fully representative of the general population. Studies have shown that Reddit users are predominantly male, young adults, with a strong representation from North America and Europe (Barthel et al. 2016; Singer et al. 2014). This demographic bias may influence the types of emotions expressed and the language used on the platform. Thus, fine-tuning models on EXPRESS may not generalize well to other populations, leading to potential biases in the emotion recognition model when applied to more diverse or global datasets. Future work could broaden the demographic coverage by incorporating data from other platforms, such as Quora, or region-specific platforms like Zhihu (Chinese) in other languages. Future work can also oversample specific subreddits to target different demographic groups, such as r/askwomen for women and r/over60 for older adults.

Limitations to Human Evaluation Approach. There are few, if any, real-world scenarios where a person would be tasked with predicting the emotion term that an emotion-experiencer would express via a written vignette. Rather, real-world scenarios involve some discussion and clarifying questions for an individual to learn about what a person might

be feeling or have experienced in the past (e.g., emotional disclosure to a friend, or a therapist). A task similar to the one in this study is a subscale of an emotional intelligence measure in which participants determine what emotion a person is feeling based on a scenario described. However, this task uses standardized vignettes developed by researchers, rather than disclosed in colloquial terms by everyday people. As suggested by the instances where both the self-disclosed and LLM terms were considered by the expert as plausible colloquial descriptions of affect, there may be cases where a term cannot be predicted by a human (at least without more context provided). Furthermore, future work could include further refinement through iterative codebook development and discussions to improve agreement among coders.

Ethics Statement

Emotion detection ability in LLMs could bring significant benefits to mental health applications. It would allow for the automation of mental health services, directing individuals to appropriate and personalized resources. This approach could enhance the accessibility of mental health support, particularly for vulnerable populations, such as ethnic minorities, where seeking help is often more stigmatized compared to majority groups (Stade et al. 2024; Habicht et al. 2024). However, there are potential risks as well. With increased emotional intelligence, large language models might become more persuasive, increasing the potential to manipulate vulnerable populations. Additionally, LLMs tend to be sycophantic (Sharma et al. 2024), and their emotional intelligence may lead them to better align with a human's opinions and sentiments. This personalization carries risks, as it can reinforce the user's existing beliefs (Kirk et al. 2024). Consequently, LLMs may avoid suggesting mental health resources that fall outside the user's comfort zone, instead adhering to assumptions based on the user's prompts and selectively presenting information that reflects the user's biases and beliefs.

The annotations used to obtain the basic emotion vectors for Plutchik's eight basic emotions were crowd-sourced, with workers receiving $0.10 per assignment, ensuring that compensation complied with the minimum wage requirements in the authors' location. Each HIT allowed sufficient time of 3 minutes for completion, aligned with the number of questions included. To ensure quality, only workers with a HIT approval rate of 95% or higher, at least 5,000 approved HITs, and who passed a task-specific qualification test were allowed to perform the annotations. We also recruited one domain expert on Upwork for qualitative analysis, compensating the expert at $30 per hour, with the entire evaluation process taking 11 hours. To maintain data anonymity, we discarded post IDs and account names before feeding the posts into the LLMs. Our dataset is self-annotated, strictly extracting emotions explicitly expressed by the post author.

References

Achiam, J.; Adler, S.; Agarwal, S.; Aleman, F. L.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adhikary, P. K.; Srivastava, A.; Kumar, S.; Singh, S. M.; Manuja, P.; et al. 2024. Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: Benchmark Study. *JMIR Mental Health*, 11: e57306.

Alm, C. O.; and Sproat, R. 2005. Perceptions of Emotions in Expressive Storytelling. In *INTERSPEECH*, volume 2005, 533–536.

Ball, G.; and Breese, J. 2001. *Emotion and personality in a conversational agent*, 189–219. Cambridge, MA, USA: MIT Press. ISBN 0262032783.

Barthel, M.; Stocking, G.; Holcomb, J.; and Mitchell, A. 2016. Reddit news users more likely to be male, young and digital in their news preferences. *Pew Research Center*, 25.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150.

Boe, B. 2021. PRAW: The Python Reddit API Wrapper. praw.readthedocs.io/en/latest/.

Brackett, M. A.; Rivers, S. E.; and Salovey, P. 2011. Emotional intelligence: Implications for personal, social, academic, and workplace success. *Social and personality psychology compass*, 5(1): 88–103.

Brown, T.; Mann, B.; Ryder, P., Nick Dhariwal; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Buechel, S.; and Hahn, U. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 578–585.

Chatterjee, A.; Narahari, K. N.; Joshi, M.; and Agrawal, P. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, 39–48.

Chochlakis, G.; Pandiyan, N. M.; Lerman, K.; and Narayanan, S. 2025. Larger Language Models Don't Care How You Think: Why Chain-of-Thought Prompting Fails in Subjective Tasks. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Chung, H. W.; Hou, L.; Longpre, J., Shayne Wei; et al. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Chung, M.; and Harris, C. R. 2018. Jealousy as a Specific Emotion: The Dynamic Functional Model. *Emotion Review*, 10(4): 272–287.

Cowen, A.; Sauter, D.; Tracy, J. L.; and Keltner, D. 2019. Mapping the Passions: Toward a High-Dimensional Taxonomy of Emotional Experience and Expression. *Psychological Science in the Public Interest*, 20(1): 69–90. PMID: 31313637.

Davis, T. 2024. List of Emotions: 271 Emotion Words (+ PDF) — berkeleywellbeing.com. https://www.berkeleywellbeing.com/list-of-emotions.html. [Accessed 06-06-2024].

Davitz, J. 2013. *The Language of Emotion*. Academic Press. ISBN 9781483261713.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Ekman, P.; Sorenson, E. R.; and Friesen, W. V. 1969. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875): 86–88.

Ekman, P.; et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60): 16.

Feng, S.; Lubis, N.; Geishauser, C.; Lin, H.-c.; Heck, M.; van Niekerk, C.; and Gasic, M. 2022. EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4096–4113.

Fontaine, J. R.; Scherer, K. R.; Roesch, E. B.; and Ellsworth, P. C. 2007. The World of Emotions is not Two-Dimensional. *Psychological Science*, 18(12): 1050–1057. PMID: 18031411.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Frattaroli, J. 2006. Experimental Disclosure and Its Moderators: A Meta-Analysis. *Psychological Bulletin*, 132(6): 823.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Ghazi, D.; Inkpen, D.; and Szpakowicz, S. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, 152–165. Springer.

Grattafiori, A.; Dubey, A.; Jauhri, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gross, J. J.; and Feldman Barrett, L. 2011. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion review*, 3(1): 8–16.

Habicht, J.; Viswanathan, S.; Carrington, B.; Hauser, T. U.; Harper, R.; and Rollwage, M. 2024. Closing the accessibility gap to mental health treatment with a personalized self-referral Chatbot. *Nature medicine*, 30(2): 595–602.

Harmon-Jones, E.; Price, T. F.; and Gable, P. A. 2012. The Influence of Affective States on Cognitive Broadening/Narrowing: Considering the Importance of Motivational Intensity. *Social and Personality Psychology Compass*, 6(4): 314–327.

Hua, Y.; Na, H.; Li, Z.; Liu, F.; Fang, X.; Clifton, D.; and Torous, J. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1): 230.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7184–7190.

Ji, S.; Zhang, T.; Yang, K.; Ananiadou, S.; Cambria, E.; and Tiedemann, J. 2023. Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health. *arXiv preprint arXiv:2304.10447*.

Jurafsky, D.; and Martin, J. H. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 20, 2024.

Khandkar, S. H. 2009. Open coding. *University of Calgary*, 23(2009): 2009.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 1–10.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA. ISBN 9781713871088.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995.

Liu, C.; Osama, M.; and De Andrade, A. 2019. DENS: A Dataset for Multi-class Emotion Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6293–6298.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021a. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021b. Towards Emotional Support Dialog Systems. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Mauss, I. B.; and Robinson, M. D. 2009. Measures of Emotion: A Review. *Cognition and Emotion*, 23(2): 209–237.

Mayer, J. D.; and Geher, G. 1996. Emotional intelligence and the identification of emotion. *Intelligence*, 22(2): 89–113.

Mazefsky, C. A.; Pelphrey, K. A.; and Dahl, R. E. 2012. The need for a broader approach to emotion regulation research in autism. *Child development perspectives*, 6(1): 92–97.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 174–184.

Mohammad, S.; and Bravo-Marquez, F. 2017. Emotion Intensities in Tweets. In Ide, N.; Herbelot, A.; and Màrquez, L., eds., *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 65–77.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3): 436–465.

Öhman, E.; Kajava, K.; Tiedemann, J.; and Honkela, T. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 24–30.

OpenAI; Achiam, J.; Adler, S.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pennebaker, J. W.; and Beall, S. K. 1986. Confronting a Traumatic Event: Toward an Understanding of Inhibition and Disease. *Journal of Abnormal Psychology*, 95(3): 274.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, 3–33. Elsevier.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Sabour, S.; Liu, S.; Zhang, Z.; Liu, J.; Zhou, J.; Sunaryo, A.; Lee, T.; Mihalcea, R.; and Huang, M. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5986–6004.

Salovey, P.; and Mayer, J. D. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3): 185–211.

Scherer, K. R.; and Wallbott, H. G. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2): 310.

Schiller, D.; Alessandra, N.; Alia-Klein, N.; Dolcos, F.; et al. 2024. The human affectome. *Neuroscience & Biobehavioral Reviews*, 158: 105450.

Shaikh, O.; Zhang, H.; Held, W.; Bernstein, M.; and Yang, D. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4454–4470. Toronto, Canada: Association for Computational Linguistics.

Sharma, M.; Tong, M.; Bowman, S. R.; et al. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.

Siemer, M.; Mauss, I.; and Gross, J. J. 2007. Same Situation–Different Emotions: How Appraisals Shape Our Emotions. *Emotion (Washington, D.C.)*, 7(3): 592–600.

Singer, P.; Flöck, F.; Meinhart, C.; Zeitfogel, E.; and Strohmaier, M. 2014. Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd international conference on world wide web*, 517–522.

Singh, S.; Caragea, C.; and Li, J. J. 2024. Language Models (Mostly) Do Not Consider Emotion Triggers When Predicting Emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 603–614.

Stade, E. C.; Stirman, S. W.; Ungar, L. H.; Boland, C. L.; Schwartz, H. A.; Yaden, D. B.; Sedoc, J.; DeRubeis, R. J.; Willer, R.; and Eichstaedt, J. C. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1): 12.

Strapparava, C.; and Mihalcea, R. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, 70–74.

Tangney, J. P.; Miller, R. S.; Flicker, L.; and Barlow, D. H. 1996. Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, 70: 1256–1269.

Team, G.; Riviere, M.; Pathak, S.; et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Trevisan, D. A.; Roberts, N.; Lin, C.; and Birmingham, E. 2017. How do adults and teens with self-declared Autism Spectrum Disorder experience eye contact? A qualitative analysis of first-hand accounts. *PloS one*, 12(11): e0188446.

Turing, A. M. 2009. *Computing machinery and intelligence*. Springer.

Van Pelt, C.; and Sorokin, A. 2012. Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 765–766.

Wang, X.; Li, X.; Yin, Z.; Wu, Y.; and Liu, J. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17: 18344909231213958.

Wei, J.; Wang, X.; Schuurmans, D.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

AAAI ICWSM Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our study evaluates LLMs' ability to detect human emotions using publicly available datasets. We aim to align AI more closely with human needs and values while taking care to avoid causing harm. By using ethically sourced data and focusing on enhancing empathy in AI, this research respects privacy norms, avoids perpetuating unfair profiling, and seeks to benefit society inclusively.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we accurately include the contributions and scope in the abstract and introduction
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we address the appropriateness of the proposed methodological approach in both the Introduction and the section 'EXPRESS: A Comprehensive Benchmark for Emotion Recognition'.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the section 'EXPRESS: A Comprehensive Benchmark for Emotion Recognition'. We describe the characteristics of our data such as the number of data points, word count, data timeline, etc. We also include our prompt formats in the Appendix
   (e) Did you describe the limitations of your work? Yes, see the Limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the Ethics Statement
   (g) Did you discuss any potential misuse of your work? Yes, see the Ethics Statement
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we provide detailed documentation about the dataset, including the entire data processing pipeline, as described in the section 'EXPRESS: A Comprehensive Benchmark for Emotion Recognition.' Additionally, we include comprehensive details about the experimental setup to ensure reproducibility, such as setting the model temperature to 0. These measures are intended to promote transparency, reproducibility, and accountability.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we provide the URL to an anonymized repository containing all our code and instructions to reproduce the results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we include the training details of topic modeling in Appendix, and state the inference parameters (temperature=0) in experiment setup.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, we set the temperature to 0 so that the results are determined without any random effects.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because we did not use any external storage or compute resources
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the section 'Evaluating Emotion Recognition Capabilities of LLMs'
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we only perform classification in our main experiment in this paper. We discuss the potential harm of misclassification in this task in various sections, including the Introduction and Conclusion.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see the Section "EXPRESS" where we cite the API of data collection. And we cited all the LLMs we examined, as well as methods.
   (b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethics Statement

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we discussed this issue, we didn't collect any identifiable information into our dataset and related content is also in Ethics Statement

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, we discussed it in the Limitation

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, it is included in our dataset repo.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, see the Appendix.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see the Ethics Statement.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we follow best practices proposed by Chancellor, Baumer, and De Choudhury (2019) to work with deidentified data, to use secure machines for our analyses with only authorized access to the paper's authors granted through the principle of least privilege, and to avoid sharing personally identifiable data in any form.

Appendix

Theories of Emotion

There are several popular theories of emotion. One class of theories views emotions as being discrete, otherwise known as basic emotions. Plutchik's wheel of emotion (Plutchik 1980) consists of 8 discrete emotions: *joy, sadness, anger, fear, trust, disgust, anticipation,* and *surprise*. Another similar theory is Ekman's basic emotions (Ekman, Sorenson, and Friesen 1969), consisting of 6 basic emotions: *happiness, anger, fear, sadness, disgust,* and *surprise*. On the other hand, some theories view emotions in a multi-dimensional space, such as Russel's Circumplex model of affect (Russell 1980). The VAD model places emotions among dimensions of valence, arousal, and dominance. Valence, arousal, and dominance are dimensions used to describe emotions: valence indicates the positivity or negativity of emotion, arousal reflects the intensity of emotional activation, and dominance measures the degree of control or influence an emotion exerts over an individual (Russell 1980).

Many works use these theories as a basis to create lexicons for emotion. Popular lexicons include the NRC Word-Emotion Association Lexicon, also known as EmoLex (Mohammad and Turney 2013), and the NRC Valence, Arousal, and Dominance lexicon (Mohammad 2018).

Emotion Masking Algorithm

We included the pattern 'feel' because prior work indicated that humans use the word feeling' interchangeably with emotion, even though feelings and emotions are not the same (Davis 2024). Feelings encompass both emotional experiences (e.g., feeling sad) and physical sensations (e.g., feeling hungry). This distinction justifies our pattern-matching format (*I* + *feel/am* + *emotion*) or (*no-pronoun* + *feeling* + *emotion*). Up to three words can be added between '*I*' and '*feel/am*', and between '*feel/am*' and the emotion, allowing the patterns to capture cases where extra words such as adverbs are present. For example, the word 'angry' in the phrase 'I have felt extremely angry' will be masked.

In addition to these basic patterns, several guidelines were introduced to ensure the accuracy of masking self-disclosed emotion words:

1. We avoided masking the word if there is a pronoun, noun, or verb between 'feel' and the emotion word, to exclude phrases like 'I feel he was sad.'
2. We avoided masking the word if there is an interrogative word between 'feel' or 'am' and the emotion word, to exclude phrases like 'I feel how happy he is.'
3. For the 'I am' pattern, we ensured that the emotion word is an adjective.
4. Finally, we performed a manual review to filter out posts that did not satisfy the conditions but were not detected by the protocol.

Post Segmentation

We segment the posts into chunks of 512 tokens using RoBERTa tokenizer due to input context length constraints of RoBERTa and Longformer. Table A.1 outlines the algorithm used to perform post segmentation.

Evaluation Metrics

The F1-score for a vector $\mathbf{v}_i$ is computed as:

$$F1(\mathbf{v}_i) = \frac{2 \cdot \text{Precision}(\mathbf{v}_i) \cdot \text{Recall}(\mathbf{v}_i)}{\text{Precision}(\mathbf{v}_i) + \text{Recall}(\mathbf{v}_i)}$$

where:

$$\text{Precision}(\mathbf{v}_i) = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}(\mathbf{v}_i) = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

Here, TP_i, FP_i, and FN_i represent the true positives, false positives, and false negatives for the prediction of the ten dimensions in vector $\mathbf{v}_i$, respectively.

The final F1-score across all vectors is obtained by averaging the vector-level F1-scores:

$$\text{Final F1} = \frac{1}{n} \sum_{i=1}^{n} F1(\mathbf{v}_i)$$

where n is the total number of vectors.

Post Segmentation Algorithm
Step 1: Initialize grouping of [MASK] tokens **Step 1.1:** Traverse through the tokens of the post one by one. **Step 1.2:** Whenever a [MASK] token is encountered, start a new group if it's the first one, or add it to the current group if it's the first [MASK] token found.
Step 2: Group nearby [MASK] tokens **Step 2.1:** Check the distance between the current [MASK] token and the last [MASK] token in the current group. **Step 2.2:** If the distance is less than or equal to 235 tokens, add this [MASK] token to the current group. **Step 2.3:** Else, finalize the current group and start a new group with this [MASK] token.
Step 3: Compute centroid for each [MASK] group For each group of [MASK] tokens, calculate the mean position by averaging the positions of all [MASK] tokens in that group.
Step 4: Create segments around centroids For each centroid, create a segment by selecting tokens around this central position. Take *n/2* tokens to the left and *n/2* tokens to the right of the centroid to create a segment of *n* tokens in total (510 tokens).
Step 5: Clip the segments Ensure that each segment makes sense contextually by clipping the segment to sentence boundaries. Clip the segment slightly to align with the nearest sentence-ending characters.
Step 6: Return the segments Once all [MASK] token groups have been segmented and clipped, return the list of these segments as the output.

Table A.1: Details the algorithm for clipping the dataset into 510-token segments

	mean	median	min	max
Words per Post	259 ($\pm$132.26)	281	20	492
Emotion words per Post	1.56 ($\pm$0.87)	1	1	5

Table A.2: Statistics of the Reddit posts in our dataset

Text	**Label**
My family's absolutely wonderful traditional Japanese New Year feast. Soy sauce *everywhere*, but this year I felt <mask> to just be with everyone, instead of sad I can't eat. They even made me separate rice balls with plain seaweed - sometimes family gatherings are stressful but this was great!	grateful
Today I decided to post this bare face photo to Snapchat instead of using a filter and actually preferred how I looked without it I dont have perfect skin, but I officially feel more <mask> without any makeup than I do with.	confident
Show this page on Pinterest the other day, and was instantly eager to recreate it on my journal, i hope i did justice. Also, on the right side is my playlist that i made last night with few of my favourite mood lifters and i feel <mask> about it. Love to hear your feedbacks! :).	happy
Decided to draw while high last night... then proceeded to write 8 pages trying to prove a 4th spatial dimension. I feel <mask> and like this picture represents a dark part of my emotions.	enlightened
My mom has been super into canning lately and was excited for summer fruits and veggies, but she broke her wrist last week and had to have surgery. I feel so <mask> for her. Any ideas about how to support her canning passion while she recovers?	sad
The rest of the gallery erupted in cheers and applause as the judge handed down the death sentence and I too felt a wave of <mask> that this monster would face justice. I recognized the fairness of the court, but hounded by an agonizing regret I also wondered where I went wrong and longed bitterly for a do-over for the little boy my son had once been.	relief
I feel so <mask> when there's no one beside me, as long as someone is there, a friend or a boyfriend nor a group of friends. I feel good and <mask>. The minute am alone it starts to feel lonely, like I have no one in the world.	lonely, cheerful
I just finished a two-year solo project and I feel <mask> and depressed at the same time. I'm playing with my app now ... it's alive! A ton of research, scrounging, all night coding around family life. Learned a ton, aged a lot but it was something I had to do after all those years of working on someone else's project. I'm so <mask> but I feel this weird sense of depression at the same time. A sort of loss. Since I'm a solo keyboard warrior, I guess you guys are the ones with whom I sharing this weird feeling(s).	elated, happy
Lately I feel immense <mask> for my life and I wish I knew who to thank. Lately I feel overwhelmingly <mask> for the good things that have come into my life. I even feel <mask> for some of the bad things that have happened to me in the last couple years, because without those I wouldn't be a person ready to accept the good things that are happening lately. I don't believe in God but I quite often throughout my life feel that there is some energy looking out for me. Someone who has my best interest at heart and has the wisdom and ability to nudge me in the directions I need to go. Not always the directions I want to go, but always the ones I need. I wish I knew who that was, so I could thank them.	gratitude, grateful, grateful

Table A.3: Examples from EXPRESS

Topic Modeling

We use BERTopic to perform topic modeling on our dataset. BERTopic leverages document embeddings, reducing their dimensionality before clustering them (Grootendorst 2022). To optimize our model's performance, we conducted hyper-parameter tuning on the HDBSCAN algorithm.[2] The highest

[2]We experiment with two hyperparameters:
min_sample = [2, 3, 5, 10, 15, 20, 25, 30, 50, 100]
min_cluster_size = [20, 25, 30, 50, 100, 150, 200, 220, 240, 250, 260, 280, 300, 330, 350, 380, 400, 420, 450, 460, 500, 520, 540,

Emotion	Count	Emotion	Count	Emotion	Count
happy	2999	guilty	2778	scared	2637
sad	2212	tired	1853	afraid	1624
anxious	1405	comfortable	1115	embarrassed	1113
angry	1107	worried	1100	depressed	1078
lonely	1055	grateful	991	confused	954
excited	924	nervous	894	confident	857
upset	826	overwhelmed	798	uncomfortable	776
proud	668	numb	645	ashamed	619
hopeless	596	surprised	522	uneasy	501
guilt	471	helpless	462	frustrated	458
interested	456	hurt	454	humiliated	424
thankful	414	disgusted	394	disappointed	386
miserable	370	mad	363	terrified	344
hopeful	329	love	326	shocked	319
stuck	313	isolated	289	jealous	283
restless	277	relief	275	loved	254
calm	248	fear	248	shame	243
optimistic	235	weak	234	unsure	232
anger	225	relieved	217	joy	213
stressed	209	bored	201	anxiety	192

Table A.4: Top 60 most frequent emotion lexicons in the EXPRESS dataset.

coherence score (C_v) achieved was 0.494, with a min_sample value of 3 and a min_cluster_size of 100, as illustrated in Figure A.1. This configuration resulted in the identification of 49 distinct topics, which are detailed in Table A.5, including the top four representative words and qualitative labels assigned to each topic.

We conducted similar topic modeling on other datasets if they did not report topic diversity and presented the results in Table 1.

Human Evaluation

Coder Background. All three of our coders are domain experts with Ph.D. degrees in related fields:

- **Coder One:** A domain expert with a Ph.D. in Psychological Science, specializing in Affective Science, with over 10 years of research experience using a combination of quantitative surveys, qualitative interviews, and experimental designs to understand emotional experience and regulation, as well as lay beliefs and academic theories about emotion.
- **Coder Two:** A domain expert with a Ph.D. in Psychology and Social Behavior, with multiple peer-reviewed publications related to emotion.
- **Coder Three:** A domain expert with a Ph.D. in Counseling Psychology. They have extensive experience working with diverse adult populations in various clinical settings, including VA Medical Centers. Their therapeutic approach emphasizes two main roles: providing empathetic support for clients' immediate concerns and fostering deeper self-awareness of internal and interpersonal patterns that may contribute to distress or dissatisfaction. They also have experience teaching as an adjunct instructor at major universities.

550, 560, 600, 620, 640, 650, 660, 670, 700]

Prompt Design

We manually tested multiple prompt variations, using accuracy score as a metric to assess performance improvements. Multiple prompt variations were tested, with accuracy score serving as the primary metric for evaluation. Our final prompt, which achieved the highest accuracy on our sample dataset, is detailed in Table A.9.

A critical consideration in our design was the definition of "emotion". We observed that providing an explicit definition of "emotion" often constrained the model's output, limiting it to a narrow set of predefined emotion terms. This conflicted with our objective of allowing the model to generate contextually appropriate emotion terms. To address this, we adopted a more flexible approach that avoids explicitly imposing a strict definition of emotions. This approach enables the model to capture nuanced emotional expressions without being restricted to predefined categories.

Amazon Mechanical Turk Annotation

Figures A.2 include screenshots of the user interface created for workers to provide annotations for the 10 dimension emotion vector.

Statistical Modeling

We conducted a Wilcoxon Signed-Rank Test to compare the LLMs' performance across the zero-shot, CoT, and two few-shot settings. This statistical test was applied to determine whether LLMs in the few-shot setting performed significantly better than those in the zero-shot setting, as well as whether the CoT setting performed significantly worse than the zero-shot setting. The LLMs, along with their corresponding p-values and Cohen's d, are presented in Table A.10.

In most cases, the few-shot settings outperform the zero-shot setting, while CoT settings perform worse than zero-shot. Only two settings do not show a significant difference from zero-shot: Llama-3.1-8B-instruct under the CoT setting and Llama-3.1-70B-instruct under the few-shot (random) setting. Additionally, the few-shot (nearest) setting consistently yields higher Cohen's d scores across all conditions.

To compare of performance of different models under zero-shot setting, we also conducted Wilcoxon Signed-Rank Tests across models. The results are presented in Table A.11.

Topic ID and Name	Qualitative label
-1_serenejsan_disgustawe__	no topic
0_like_friend_dont_know	relationships
1_back_could_eye_one	sex and pregnancy *(sex)*
2_job_work_year_like	new career or career transition *(career)*
3_like_day_trip_time	drugs and mental illness *(drugs)*
4_game_player_play_character	gaming
5_woman_men_trans_gay	sexuality and identity
6_weight_lb_mile_race	health and weight loss
7_god_church_jesus_christian	christianity
8_makeup_comfortable_confident_dress	self-confidence and body image *(self-confidence)*
9_hair_skin_picking_acne	personal appearance
10_anxiety_anxious_panic_attack	anxiety
11_husband_wife_ha_told	marriage
12_grief_died_mom_life	death
13_dog_cat_vet_puppy	pets
14_meditation_experience_like_life	meditation
15_film_movie_character_season	movies
16_team_optimistic_player_league	sports
17_anger_angry_like_emotion	anger
18_worthy_art_posting_drawing	art
19_pain_symptom_doctor_day	medical symptoms and issues *(medical symptoms)*
20_song_album_music_uzi	music
21_sad_cheer_today_depressed	sadness and loneliness *(loneliness)*
22_patient_nurse_surgery_hospital	surgery and medical procedures *(medical procedures)*
23_happy_today_joy_happiness	happiness
24_excited_excitement_tender_enjoy	excitement
25_people_like_dont_think	social isolation
26_school_class_teacher_student	school
27_drinking_drink_sober_alcohol	alcoholism
28_embarrassed_humiliated_like_today	embarrassment
29_grateful_gratitude_today_thankful	gratitude
30_fear_scared_tara_douma	fear
31_guilty_guilt_like_dont	guilt
32_porn_sex_sexual_lust	sexual and porn addiction *(sexual addiction)*
33_vegan_meat_food_eat	veganism
34_wedding_sister_family_friend	weddings
35_confident_today_lb_confidence	confidence due to weight loss *(confidence)*
36_sleep_bed_wake_night	sleep
37_dream_like_nightmare_woke	dreams and nightmares
38_pride_shame_proud_sense	pride and shame
39_ocd_thought_like_intrusive	OCD
40_book_ron_read_character	books
41_ako_house_lang_hector	housing
42_insulted_offended_insult_coach	insult
43_birthday_gift_today_friend	birthdays
44_affection_affectionate_love_hug	love
45_lust_passion_lustful_passionate	lust
46_manager_office_customer_work	workplace dynamics
47_happiness_happy_joy_life	happiness and depression
48_agony_despair_fazgoo_anguish	pain

Table A.5: The top 4 words for each topic generated through topic modeling and their respective qualitative labels.

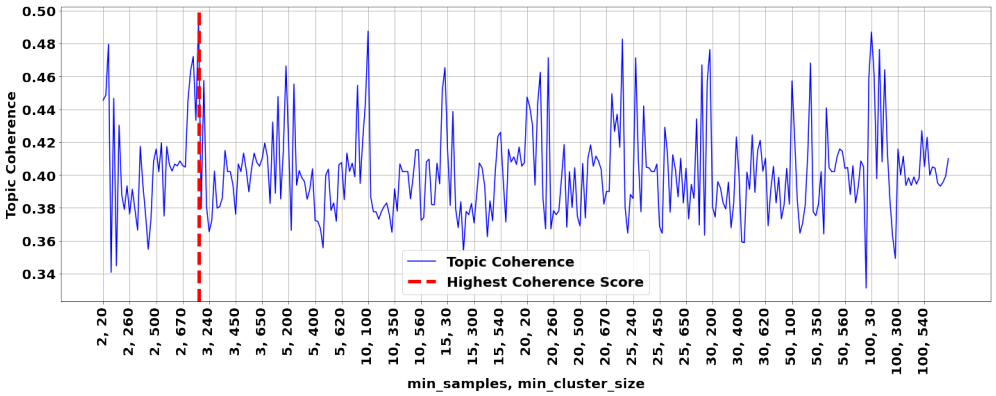

Figure A.1: The topic coherence graph illustrates the coherence scores for all the hyperparameter combinations tested. The highest coherence score achieved is 0.494. The x-axis represents the different hyperparameter combination pairs.

Figure A.2: Amazon Mechanical Turk user interface for EmoLex annotations

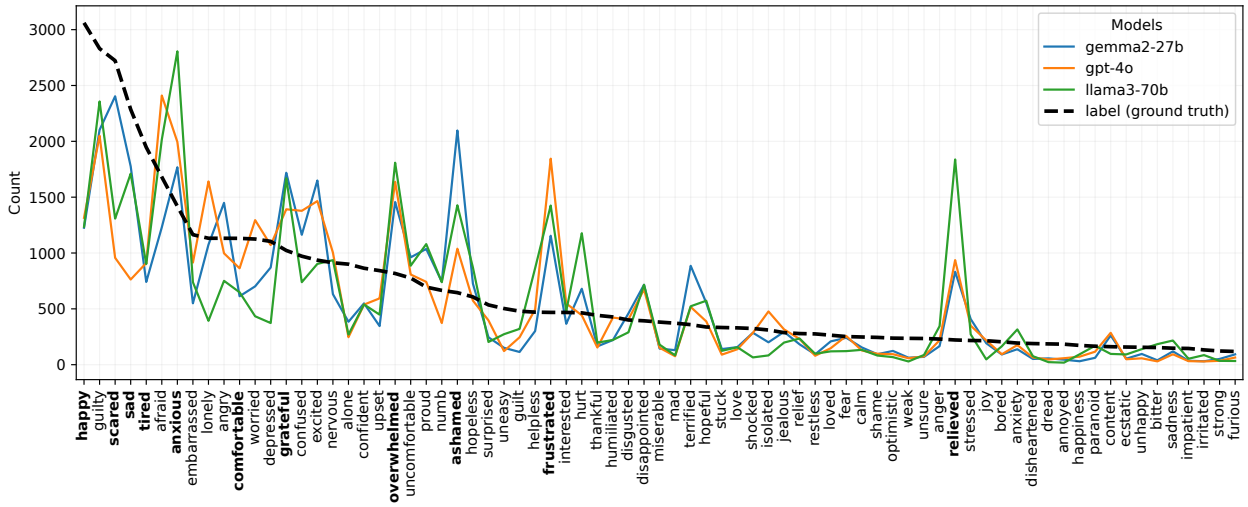

Figure A.3: Top 75 emotion counts: human labels vs. model predictions (GPT-4o, Llama 3.1-70B Instruct, Gemma 2-27B). Emotions are sorted by label frequency; larger differences are highlighted in bold.

Label	Prediction	Frequency
thankful	grateful	0.60
disheartened	frustrated	0.33
demoralized	frustrated	0.32
eager	excited	0.31
furious	angry	0.29
pleased	proud	0.29
optimistic	hopeful	0.27
terrified	afraid	0.27
restless	anxious	0.25
scared	afraid	0.24
irritated	frustrated	0.24
panicked	anxious	0.24
agitated	anxious	0.23
powerless	helpless	0.23
thrilled	excited	0.23
stress	anxiety	0.23
annoyed	frustrated	0.23
disoriented	overwhelmed	0.23
impatient	frustrated	0.22
passionate	excited	0.21

Table A.6: Top normalized mispredictions (vector level) by top three causal language models.

Category	Qualitative Code Description	Example
Applies to SD and LLM	Code 1 ($N_{SD} = 24, N_{LLM} = 19$): Valence (positive/-pleasant or negative/unpleasant) matched the context of the situation and/or the affective reaction described (N=18)	My sister just suffered an ectopic pregnancy. I feel guilty for being thankful that it happened before her state banned abortion. Yesterday my pregnant sister was rushed to the ER with excruciating pain. She was 7 weeks pregnant. Once they realized she had an ectopic pregnancy, the doctors gave her the medicine to stop it. Luckily, my sister is fine, nothing ruptured. I am <mask> but also livid - what if this happened four weeks later? What if the doctors didn't give her the proper care, so my sister had to endure the pain and potentially put her life at risk? What if it causes her to never have children again? All because of the Supreme Courts f*cked up decision to overturn Roe v Wade. *label: sad, predicted: relieved*
	Code 2 ($N_{SD} = 39, N_{LLM} = 69$): Terms have similar valence (positive/pleasant or negative/unpleasant) but the selection was judged as better fitting the degree of specificity in the affective experience being described (more general vs. more specific)	Am I a bad friend for feeling <mask> that my friend is replacing me? for the longest time, ive been really close to one of my friends. we always gossip together, go everywhere together, and do so many things together. unfortunately shes struggled with making friends for a really long time but recently shes been doing really well and making new friends and ive felt really proud of her. a friend of mine was hanging out with the both of us and the friend of mine had brought another friend. the friend of the friend has instantly clicked with my close friend and they were almost inseparable afterwards. i feel like i shouldnt feel sad that shes getting closer to other people but she doesnt really talk to me when the other girl is there. *label: upset, predicted: jealous*
	Code 3 ($N_{SD} = 101, N_{LLM} = 58$): Selection better matches other term(s) providing context of the the affective experience	Feel hatred for family sometimes. Like, when I say I'm <mask> or depressed. They tell me "then change it" like there's some f*cking switch you can flip and it all gets better. I'm working my body and brain into dust just to make rent. It ain't like the movies folks, the people closest to you cut the deepest *label: unhappy, predicted: frustrated*
	Code 4 ($N_{SD} = 20, N_{LLM} = 37$): Degree of intensity (greater or lesser) of the selection is better matched to or reflective of typical responses to the scenario described	I feel so <mask>. I need advice please. I was in school I I've gotten a weird feeling like I was about to throw up, well I did not but there was this loud sound of like choking or something (?) I just heard people giggle in the back what do i do now? Please give me some advice i feel horrible right now *label: humiliated, predicted: embarrassed*
	Code 5 ($N_{SD} = 25, N_{LLM} = 57$): Selection better fits a definition of the emotion described in the scenario, as proposed by a peer-reviewed published theory/definition of the emotion	I feel riddled with <mask> about whether hrt will be illegal within the near future, how can I deal with this? The far right has been launching an war on trans people and gender affirming care providers and I'm afraid that within a few years or less than a year, it will no longer be legal for me to access hrt. If the right doesn't ban hrt via legislature, they will at least try to intimidate doctors away from prescribing it via violent threats and intimidate pharmacies away from distributing it. And I don't want to go back to the way I was before, hrt has done so much for my happiness and confidence. But now it's going to be taken away from me again. And it causes me a ton of anxiety and sadness not knowing if I'll be forced to detransition someday soon. *label: fear, predicted: anxiety*
	Code 6 ($N_{SD} = 16, N_{LLM} = 16$): Selection reflects whether the event/scenario described has already occurred (post-goal emotion, e.g., happiness/joy or down) or is likely to occur (pre-goal emotion, e.g., excitement or nervous)	Feeling so <mask>. I (27F) posted in here a few weeks ago about the proposal that wasnt. Ive already been clear with my (30M) boyfriend on my timeline and hes on board. I just feel so taken for granted knowing that he has nothing planned for the foreseeable future. (Our weekends are all pretty much booked and spoken for from now until the end of the holiday season). Trying to shake this feeling of sadness and irritability whenever I see him. *label: melancholy, predicted: disappointed*
	Code 7 ($N_{SD} = 9, N_{LLM} = 11$): Alternative does not fit as well with evidence-based normative descriptions of affective experience in similar scenarios	I feel <mask> to admit that i like the last of us part 2! i enjoyed the game and the story not at all bad! as youtubers and people made it seem. (according to me) and i enjoyed the gameplay and it felt good. the only thing weird was the ¿!dual protaganist!¡ but i had no problems with it! and the game looks so amazing im on a ps4 and its 60 fps too!i really enjoyed the game as much as i did the first one. *label: scared, predicted: embarrassed*
	Code 8 ($N_{SD} = 18, N_{LLM} = 13$): Alternative is not as consistent with colloquial descriptions of similar scenarios (according to the coder)	My parents refuse me medical attention for my depression because they believe God will heal me even though i attempted suicide. Help me make sense of it all. hello all. Its not going good with me. My religious parents refuse me medical treatment for my depression and instead pray and force me to fast once a week. I tried to commit suicide last year November because i am messed up in life, in love with a black girl and they found out. My dad rambled racist stuff and i felt <mask> and lost. she is the only thing i have so i overdosed. I have had mental issues for a year now and things compounded and i found my way in hospital. My parents promised to get me helo. They didn't and change the topic. I am really suffering and God evem left me. I am 15 and will become an atheist because of this. I already tore a Bible in anger. Why would God allow this to happen and harden my parents heart and not even care. My girlfriends family are waay better to be around and they all non religious. They not bigotic and treat me well. all efforts to help me are shut down by my parents. they tell me God is testing me and i must be strong and not weak. help. I think i will leave religion if rhis continues. *label: hopeless, predicted: lost*
	Code 9 ($N_{SD} = 25, N_{LLM} = 20$): Alternative does not logically follow as well as the alternative from the described situation, based on coder's lay theory of normative affective experience in similar situations	Feeling <mask> after quiting weed. So I quit smoking weed at night 2 and a half weeks ago. I feel better and I'm not having any problems right now. But I feel so bored. I have stuff to do clean and cook, bake etc. I'm a stay at home mom and my husband works alot. I really want to play the Sims 3 on my laptop but I don't play it anymore because I used to be addicted to playing it. I'm figured I couldn't control my usage because of episodes before I was unmedicated and not diagnosed, but I'm stable now and think I can regulate how much I play but I've tried playing GTA on my ps4 and I still feel so <mask> idk if games would do anything for me. I don't feel like watching tv at all. I'm just here being bored all day. What do yall think what should I do? Has anyone been through this after stopping smoking weed? *label: bored, bored, predicted: anxious, disinterested*
Neither	Code 10 ($N = 4$): No information provided on the object/-subject about which affect is being expressed (Unclear what the person is feeling affect about)	Feeling <mask>. You guys are gonna have to deal with my attempts as I try to figure out face paint *label: melancholy, predicted: nervous*
Both	Code 11 ($N = 57$): Not enough information about the situation to determine whether on affect term was a better fit to the situation; rather, both terms could be applied such that the expression of affect would be equally or similarly believable	Just put my pc to sleep and felt very happy for some reason. Decided to take a picture. For absolutely no reason I felt very <mask> for having a computer like this. Its not the best out there but its mine. Does that happen to anyone else? *label: thankful, predicted: grateful*

Table A.7: Descriptions of the qualitative codes developed by the emotion expert to categorize instances from the dataset. The underlined text indicates the option selected by the expert.

Code	Selection and Reasoning	Example
9	Selection of hopeless (SD) instead of confused (LLM): based on the description of the situation indicating that the person appraises the situation as being bad, rather than unclear or confusing.	I feel <mask> and don't know what to do. To lay out a long story short. I have been in a relationship for over six years... Within two months of my transferring, she cheated on me with another girl. She continued to talk to this girl despite me not wanting her to... This continued and she got more abusive again.
9	Selection of frustrated (LLM) instead of numb (SD): based on the coder's lay theory that feeling the other emotions indicated in the excerpt would not constitute "numbness", or lack of ability to feel.	I'm <mask> because I got tired of seeing my mom and dads disappointing faces in everything I tried so hard at, just to fail. My parents have paid for multiple classes so I could actually be good at something but I never show any promise and It always ends up worse than when I started trying. I have one friend and I feel like hes my opposite... I'm starting to hate everything I once enjoyed because the thought of him being better than me at everything creeps into my mind. At this point I feel like there is no hope for me to be happy, just numb.
9	Selection of grateful (LLM) instead of happy (SD): based on the assumption that the person is sharing an account of less-than-ideal situation, rather than a situation to be happy for.	Unfortunately I'm still low contact with them despite wanting to be no-contact (long story short, I got caught packing up my stuff and it turned into this huge f****** ordeal to the point where my online friend turned roommate who works a job involved with law enforcement had to drive 600 miles to rescue me and lowkey intimidated my family into letting me go), but GOD I will take the mental equivalent of weekly probation calls over the shit I was dealing with before. Hopefully I can fully ditch their asses someday. I'm just so <mask> to be able to exist, indulge myself in my hobbies and go to places without immediately being interrogated about what I'm doing or where I'm going or being.
2	Selection of upset (SD) instead of jealous (LLM); based on lack of specificity of the alternative, that would be required to meet criteria for the definition of jealousy as a discrete emotion (Chung and Harris 2018).	am i bad friend for feeling <mask> that my friend is replacing me?...i feel like i shouldnt feel sad that shes getting closer to other people but she doesnt really talk to me when the other girl is there.
4	Selection of ashamed (LLM) instead of embarrassment (SD): given the use of the descriptor of feeling as being "so intense", and based on findings that shame is relatively intense in terms of both negative valence, and physiological arousal, relative to embarrassment (Tangney et al. 1996).	...After the break up the feeling I felt was so intense and awful. Id just cry myself to sleep in silence because I knew no one understood how miserable I was and am. I never talk about it because Im <mask> I still care...
5	Selection of angry (SD) instead of frustrated (LLM): based important criteria met for the definition of anger as involving attribution of the other person as being at blame (rather than the self) for negative situation, in addition to goal blockage (Siemer, Mauss, and Gross 2007).	...I messaged her again yesterday, she responded playing the blame game, wouldnt call me, said my family and I publicly posting that she fled was embarrassing her, were messed up for trying to hurt and embarrass her, etc. Basically blaming us for why she wouldnt reach out. I was <mask> and in tears. Classic behavior of someone struggling with addiction who wants to take no responsibility... This has been going on for years. I lost my adoptive father to alcohol and drugs. I sat with him as he was dying. I cannot watch my sister die. I cannot do this again.
6	Selection of scared (SD) instead of guilty (LLM): based on the description of the goal (maintaining the status quo in family in the context of the father's death) as being in future, rather than being in the past (attained, or not attained) (Harmon-Jones, Price, and Gable 2012).	I feel like I want to die (not really), sort of, because I wanted so bad to be there for my Dad, who I adored, and I just couldn't do it... I loved him so much, and I came back to the hospital after he passed, and held him forever, sobbing, telling him how much I missed him. He was so still, but I wasn't scared then for some reason.? I think I was more <mask> that I was braver than my siblings, and wanted to stay, but didn't' want to upset the stupid family dynamic, when everyone was like "well I'm not staying, I can't take this, " and I felt compelled to leave as well, some stupid, weak show of solidarity. I hate myself...

Table A.8: Examples of selection and the reasoning for the selection of emotion terms, including corresponding codes and related theories for reference.

Zero-shot Prompt:
You are an assistant tasked with predicting emotion words masked as ¡mask¿ in a given self-disclosure text from social media. Predict the {number of labels} ¡mask¿ token*s* based on the context.
Provide your answer in the format [example_format]. The length of the list must be {number of labels}. Only include words describing emotions, and provide no extra text or reasoning.

Text: {post}
Answer:

Chain-of-Thought (CoT) Prompt:
You are an assistant tasked with predicting emotion words masked as ¡mask¿ in a given self-disclosure text from social media. Predict the {number of labels} ¡mask¿ token*s* based on the context.
Think step by step to arrive at the final answer. In your response, first provide the reasoning, then provide your answer in the format: [example_format]. The length of the list must be {number of labels}. Only include words describing emotions in the list.

Text: {post}
Answer:

Few-shot Prompt:
You are an assistant tasked with predicting emotion words masked as ¡mask¿ in a given self-disclosure text from social media. Predict the {number of labels} ¡mask¿ token*s* based on the context.
Provide your answer in the format [example_format]. The length of the list must be {number of labels}. Only include words describing emotions, and provide no extra text or reasoning.

Examples:
Text: { example_text }
Answer: { example_answer }
Text: { example_text }
Answer: { example_answer }
Text: { example_text }
Answer: { example_answer }
Text: { example_text }
Answer: { example_answer }

Now, use this pattern for the given text.
Text: {post}
Answer:

Table A.9: Prompt templates for Zero-shot, Chain-of-Thought (CoT), and Few-shot setups. {number of labels} and [example_format] are adjustable based on the number of masks in the post.

Model Name	Setting	p-value	cohen's d
Llama-3.1-8B-instruct	Few-shot (random)	7.08e-11	0.027
	Few-shot (nearest)	7.32e-66	0.077
	CoT	1.0	-0.013
Llama-3.1-70B-instruct	Few-shot (random)	1.0	-0.34
	Few-shot (nearest)	2.70e-06	0.015
	CoT	0.0	0.161
Gemma-2-2B-it	Few-shot (random)	1.39e-08	0.026
	Few-shot (nearest)	2.12e-68	0.086
	CoT	1.86e-246	0.120
Gemma-2-9B-it	Few-shot (random)	5.91e-54	0.061
	Few-shot (nearest)	1.82e-89	0.082
	CoT	0.0	0.204
Gemma-2-27B-it	Few-shot (random)	0.000125	0.014
	Few-shot (nearest)	1.79e-13	0.028
	CoT	0.0	0.314
GPT-3.5-turbo	Few-shot (random)	7.99e-22	0.035
	Few-shot (nearest)	1.11e-37	0.048
	CoT	4.79e-234	0.116
GPT-4o	Few-shot (random)	1.66e-53	0.058
	Few-shot (nearest)	6.25e-111	0.086
	CoT	4.79e-234	0.116

Table A.10: P-values of different models and settings.

Model Name 1	Model Name 2	p-value	cohen's D
Flan-t5-xxl	Llama-3.1-8B-instruct	2.33e-34	0.084
	Gemma-2-9B-it	2.32e-86	0.144
GPT-3.5-turbo	Llama-3.1-70B-instruct	6.17e-22	0.043
	Gemma-2-27B-it	1.37e-82	0.087
Gemma-2-27B-it	Gemma-2-2B-it	0.0	0.468
	Gemma-2-9B-it	2.31e-129	0.099
Gemma-2-9B-it	Gemma-2-2B-it	0.0	0.367
Llama-3.1-70B-instruct	Llama-3.1-8B-instruct	0.0	0.233
GPT-4o	GPT-3.5-turbo	2.173e-277	0.160

Table A.11: P-values of different models and settings.