This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3584775

# Rethinking Emotion Annotations in the Era of Large Language Models

Minxue Niu, Yara El-Tawil, Amrit Romana, Emily Mower Provost
University of Michigan, Ann Arbor

**Abstract**—Modern affective computing systems rely heavily on datasets with human-annotated emotion labels for both training and evaluation. However, human annotations are expensive to obtain, sensitive to study design, and difficult to quality control, because of the subjective nature of emotions. Meanwhile, Large Language Models (LLMs) have shown remarkable performance on many Natural Language Understanding tasks, emerging as a promising tool for text annotation. In this work, we analyze the complexities of emotion annotation in the context of LLMs, focusing on GPT-4 as a leading model. In our experiments, GPT-4 achieves high ratings in a human evaluation study, painting a more positive picture than previous work, in which human labels served as the only ground truth. On the other hand, we observe differences between human and GPT-4 emotion perception, underscoring the importance of human input in annotation studies. To harness GPT-4's strength while preserving human perspective, we explore two ways of integrating GPT-4 into emotion annotation pipelines, showing its potential to flag low-quality labels, reduce the workload of human annotators, and improve downstream model learning performance and efficiency. Together, our findings highlight opportunities for new emotion labeling practices and suggest the use of LLMs as a promising tool to aid human annotation.

**Index Terms**—Emotion Recognition, LLMs, Annotation, Crowdsourcing

## 1 Introduction

DEVELOPING systems that can recognize, interpret, and respond to human emotions is playing a growing role in human-centered AI systems [1], [2]. Human emotion understanding is beneficial in various fields, such as education [3], healthcare [4], and many others [5]. In recent years, we have seen significant performance advancements in emotion recognition, especially with the popularity of deep learning methods [6]. However, these models rely heavily on data with human-annotated emotion labels, which are costly in terms of time and resources and difficult to obtain due to the inherent ambiguity of emotions. Currently, there is no standard approach to annotation, as datasets often adopt different protocols at each phase, such as label selection, annotation formats, evaluation methods, etc. In the meantime, recent advances in LLMs have opened new avenues for text-based annotation. In this work, we explore emotion annotation choices within the context of LLMs, examining how LLMs perform on emotion classification tasks and how they might address existing challenges and provide new perspectives on emotion annotation processes.

Emotions are inherently ambiguous and subjective [7], [8], posing great challenges in the design of annotation studies. Low agreement is often observed among annotators [9]. Annotation outcomes are sensitive to even small changes in study design, such as the label space offered, including the size of the label space and type of emotions to be labeled (see Section 2.1), as well as how the text and labels are presented to a human annotator (see Section 2.2) [9]. These changes can all lead to different annotation outcomes [10]–[13]. This lack of consistency raises serious concerns about the reliability of emotion labels [14], [15]. Additionally, emotion perception naturally differs from person to person, influenced by individual experiences and demographic factors [16]–[19], making it difficult to identify actual errors from legitimate perceptual differences. As a result, it is also hard to apply quality control methods post hoc. Many studies have sought to improve annotation reliability by exploring factors such as label space selection [11], study design choices [12], [13], annotation interface improvement [20], and trade-offs between annotators' quality and quantity [21]. However, establishing a general pipeline for consistent and reliable emotion labeling remains an open challenge.

With the impressive advances in LLMs, there is a growing interest in using LLMs for various tasks such as generation, assessment, filtering, and annotation [22], [23]. Related work has also found that LLMs seem to possess an emerging ability to understand and interpret emotions (see Section 2.3). However, much of this research is based on individual datasets, each with its own specific label space [24]–[27], leaving questions about the generalizability of findings across different label spaces. Further, current evaluations tend to benchmark LLMs against human emotion labels [25], [28], which themselves may contain errors. In our previous work, we conducted a small-scale in-house evaluation study. We found that human evaluators often preferred GPT-4 annotations over traditional human labels, particularly on larger label spaces [29]. While these findings provide valuable insights, further verification with larger samples, more annotators, and more comprehensive analysis is needed for a deeper understanding. Lastly, beyond fully human-driven or fully automated GPT-4-based annotation, a promising and under-explored direction is to integrate GPT-4 as a supporting component within the annotation pipeline.

In this work, we focus on two Research Questions. First, we ask **how well GPT-4 performs on emotion annotation**. To address this, we conduct a human evaluation study to compare the zero-shot predictions of GPT-4 with human labels (Section 4). Interestingly, although automatic metrics indicate that GPT-4 performs no better than small supervised models trained on human labels [28], [29], evaluators consistently prefer GPT-4 labels over human labels, showing a misalignment between automatic metrics and human perspectives. A closer inspection reveals that larger label spaces enable more precise descriptions of emotions, and GPT-4 especially excels at managing a wide range of options. This study expands on our previous work [29] with a larger sample size and more evaluators within a crowdsourcing environment, providing stronger support for our findings and deeper insights into the reasons behind human preferences.

Building on this understanding of LLMs' emotion capabilities, we explore the second Research Question: **Can GPT-4 help humans annotate emotions?** We examine two strategies to incorporate GPT-4 into annotation pipelines (Section 5). While previous work has explored automatic pre-annotation as a process to narrow label choices for human annotators, these works have focused on single-label annotation and have relied on traditional text analysis tools, such as lexicons [20], [30]. In our study, we propose to leverage LLMs as a more advanced tool. We present a novel investigation into the feasibility of (1) employing GPT-4 as a pre-annotation filter to dynamically suggest appropriate labels, and (2) using GPT-4 as a post-annotation filter to flag samples with low-quality human labels. Our experiments find some clear advantages, such as enhancing model training outcomes and efficiency, reducing cognitive load on annotators, and preserving the granularity benefits of large emotion spaces. To the best of our knowledge, this is the first study to propose and evaluate the pre-filtering and post-filtering methods, showing encouraging results.

**Together, this work makes the following contributions:**

- We provide a systematic evaluation of GPT-4's zero-shot emotion annotation across datasets with varying label complexity, showing that human evaluators often prefer GPT-4's annotations over human labels.
- We offer insights into how label space size affects annotation quality, highlighting GPT-4's strength in handling fine-grained emotion categories.
- We propose and evaluate two novel strategies for integrating GPT-4 into human annotation pipelines—as a pre-annotation filter and a post-annotation quality check—demonstrating their effectiveness in enhancing annotation experience, agreement and efficiency.

Throughout our analysis, we carefully consider the complexities of emotion label spaces and the varying perspectives captured by different evaluation methods, yielding valuable insights for future emotion annotation practices. These findings advocate for thoughtful consideration of annotation design choices, highlighting the potential of LLMs as a powerful tool to leverage alongside human labelers to elevate the annotation process in emotion recognition tasks.

# 2 Related Work

## 2.1 Emotion Label Spaces

The complexity and ambiguity of emotion pose significant challenges in quantifying and labeling emotions for building emotion recognition systems. The most commonly used frameworks for describing emotions fall into two categories: categorical label spaces, where emotions are represented as one or more pre-defined categories (e.g., joy, sadness) [31], and dimensional label spaces, which conceptualizes emotions along continuous axes, such as valence (positive to negative) and activation (excited to calm) [32].

Within the emotion classification framework, selecting an appropriate set of emotion labels must be carefully considered. A common approach is to follow established theories of basic emotions. For example, Emobank [33] and DailyDialog [34] datasets adopt Ekman's theory of six basic emotions (i.e. Anger, Disgust, Fear, Happiness, Sadness, and Surprise) [31]. Other works make small modifications based on existing theories; for example, ISEAR [35] removed Surprise while adding Shame and Guilt to their label set. Another common strategy is conducting pre-annotation studies to determine the most appropriate set of emotion labels for the target data. SemEval-2018 Task 1 ran pilot annotation and included 11 emotion classes [36]. GoEmotions settled on 27 classes after an iterative refinement process [37].

## 2.2 Challenges in Obtaining Human Annotations

Obtaining high-quality, reliable human emotion annotations is a nontrivial task. It is common to see low agreement among annotators (e.g., the unanimous agreement can easily be below 10% in some datasets [9]). One reason for the low agreement lies in the inherent subjectivity of the task [16]. Research has found that demographic factors, such as gender [17], age [18] , and race [19], significantly affect how emotions are perceived. As a result, a lack of diversity among annotators may result in datasets failing to capture the full spectrum of emotional perspectives, potentially leading to biased data and models [38]. In addition, many design choices can significantly affect the annotation experience and outcomes. For example, the choice of label spaces plays an important role [11]. Larger label spaces include more diverse and nuanced options, allowing for more accurate descriptions of emotion. However, more options reduce the agreement between annotators, possibly amplifying perspective differences or causing annotation fatigue [9]. The availability of context is another key factor. Providing context during annotation generally helps reduce repetition, ease the task, and produce annotations more aligned with speakers' self-reported emotions [12], [13]. However, contextual influence can introduce inconsistencies, as variations in sample order affect annotators' judgments [10]. Finally, the effort and attention devoted to the task varies significantly by individuals. A study evaluating annotation quality across four crowdsourcing platforms revealed that roughly half of the participants failed at least one attention check, with failure rates reaching 72.9% on the least reliable platform [39]. In summary, human annotations are subjective and sensitive, and the quality is often far from perfect.

Evaluating the quality of obtained labels is also challenged by the ambiguous nature of emotion. Without ground-truth labels, agreement metrics have been used as a major quality indicator or as a criterion to remove potentially low-quality samples/annotations [40]. However, a higher level of agreement does not necessarily indicate more

TABLE 1: Summary of the datasets we use. In emotion labels, we show classes that occur in all datasets in bold and unique classes in one dataset with underline.

| | domain | #classes | multilabel | #samples (k) | emotion labels |
|---|---|---|---|---|---|
| ISEAR | Self-reports | 7 | No | 7.7 | **anger**, **disgust**, **fear**, guilt, **joy**, **sadness**, shame |
| SemEval | Tweets | 11 | Yes | 6.8/0.8/3.3 | **anger**, anticipation, **disgust**, **fear**, **joy**, love, optimism, pessimism, **sadness**, surprise, trust |
| GoEmotions | Reddits | 28 | Yes | 43.4/5.4/5.4 | admiration, amusement, **anger**, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, **disgust**, embarrassment, excitement, **fear**, gratitude, grief, **joy**, love, nervousness, optimism, pride, realization, relief, remorse, **sadness**, surprise, neutral |

meaningful labels: it can result from reduced diversity in annotations [9]. Another way to evaluate annotations is to put them in use—to train models with those labels and measure the performance on a test set [9], [21]. However, this approach relies on the assumption that "golden" labels of high quality and reliability are available in the test set — an assumption many existing datasets fail to meet. Finally, human annotations are expensive, requiring significant time and effort, often involving recruitment, training, and extensive post-analysis [7], [35]. In some cases, multiple iterations are necessary for reliable results [37]. As models grow in size, the cost of data collection increases further due to the need for more data to adequately train them [41].

### 2.3 Emotional Capability of LLMs

Previous work has found that through conversational interactions, LLMs show emerging emotional intelligence [42]: they can recognize sentiment [28], analyze the cause of emotions [43], [44], and engage in dialogues with empathy [43], [45]. The natural question that follows is whether they can be used to annotate emotions in a structured manner, adhering to predefined labels and producing consistent outputs. Existing work has examined the zero-shot emotion recognition performance of various LLMs, from smaller open-sourced models like RoBERTa [46] to larger commercial models like GPT-4 and Gemini [47], [48], generally finding reasonable performance. Instruction tuning has been shown to further improve their emotion recognition performance across a range of label spaces [49] and benefit other emotion-related tasks [50]. However, different evaluation criteria have led to different findings: many studies use human annotations as ground-truth [25], [28], [51], and find that LLMs remain inferior to human performance or fail to outperform smaller supervised models, particularly on complex tasks with numerous emotion labels. On the other hand, preliminary studies incorporating human evaluators in their assessment have shown more promising results [47]. Our own work, which conducted a small-scale human evaluation study comparing GPT-4 and human labels, also reported more positive findings on LLM performance compared to humans [29]. Further, some initial results suggest that LLMs are worse at larger label spaces than small, well-defined ones [46], [52]. Still, this effect is under-explored, and it is not clear whether this is inherent in LLMs or can be mitigated through proper prompting methods.

## 3 Datasets

We use three existing English Emotion Classification datasets. They are all commonly used datasets to evaluate emotion models, covering diverse domains, topics, and different levels of granularities of emotion classes. Table 1 shows a summary of the datasets and label spaces.

**International Survey on Emotion Antecedents and Reactions (ISEAR)** [35] was collected as part of a research project that aimed to study emotional experiences across cultures. The dataset contains more than 7000 self-reported descriptions of emotional experiences in English from participants in 27 countries, each describing emotional experiences in one of seven categories (listed in Table 1). We randomly split it into 60% train/20% dev/20% test sets.

**SemEval 2018 Task 1 (SemEval)** [36] is part of a multilingual affect analysis task released at the International Workshop on Semantic Evaluation. We take the English subset from the Emotion Classification subtask (E-c), where each tweet is annotated with zero, one or more labels from eleven emotion classes. The annotations were collected by crowdsourcing. The dataset was released with train/dev/test splits.

**GoEmotions** [37] is a large-scale multilabel emotion classification dataset consisting of over 58,000 English Reddit comments annotated for 27 emotion categories (plus a neutral category) through crowdsourcing. GoEmotions is notable for its large data size and label granularity, offering a rich resource for fine-grained emotion classification. We use its released train/dev/test splits.

## 4 GPT-4's Emotion Annotation Capability

In this section, we evaluate GPT-4's emotion annotation capabilities through a crowdsourcing-based human evaluation study, assessing its alignment with human perceptions. We provide a comprehensive analysis with a focus on the disagreements between human and GPT-4 annotations.

### 4.1 GPT-4 Prompting

To evaluate the zero-shot emotion recognition capability of GPT-4, we first query its predictions for all three datasets using the Microsoft Azure API. We use the gpt-4-1106-preview deployment, which was the latest stable version available at the time of our experiments. We employ an instruction-driven approach [53]: we prompt GPT-4 with a system prompt, which describes the task and serves as an instruction, and a user prompt, which includes solely the input text content. The instructions ask GPT-4 to identify

| Dataset | Text | GPT-4 Output |
|---|---|---|
| GoEmotions | Our father will protect us <3 | "love, optimism" |
| ISEAR | I was selected to come here (University, College) when I was least expecting it. | "joy" |
| SemEval | No man read the traffic properly!!! | "anger, pessimism" |
| GoEmotions | Nobody likes you non-human mimics and everyone knows what you are. | *rejected* |

TABLE 2: Example GPT-4 outputs for emotion annotation across different datasets. The last sample was rejected by GPT-4 due to its content policy, and thus it was excluded from our analyses.

the appropriate label(s) from the provided list and ensure its outputs follow a predefined format that can later be parsed with rule-based post-processing. The instructions are designed to mirror those given to human annotators, creating a consistent and comparable task framework. Additionally, we enhance the prompts by establishing a persona at the start, which has been found beneficial in some work [54].

We used the following system prompt for multilabel emotion classification [29] on the SemEval dataset:

> GPT-4 prompt for emotion classification
> *"You are an emotionally-intelligent and empathetic agent. You will be given a piece of text, and your task is to identify all the emotions expressed by the writer of the text. You are only allowed to make selections from the following emotions, and don't use any other words: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust. Only select those ones for which you are reasonably confident that they are expressed in the text. If no emotion is clearly expressed, reply with 'neutral'. Reply with only the list of emotions, separated by comma."*

For ISEAR and GoEmotions, we made minimal adjustments to the prompt to reflect different emotion options and task settings (i.e., whether multilabel is allowed). In rare cases where the output did not follow the specified format and could not be parsed, we retried with the same query. GPT-4 also has content policies and may refuse potentially harmful or sensitive content[1]. We excluded those samples from our analysis (ISEAR 3.7%, Semeval 4.5%, GoEmotions 2.6%). We show a few text samples and GPT-4 responses in Table 2.

## 4.2 Human Evaluation Study

Given the inherent ambiguity of emotion and the absence of absolute "truth" labels, human judgment remains essential for evaluation in this domain. We conduct a human evaluation study, engaging a separate group of humans (we refer to them as "evaluators", to differentiate from "annotators" who provided the label annotations in the datasets) to assess how accurately GPT-4 and human annotations reflect the emotions in text. All human-subject studies reported in this paper were approved by the University of Michigan Institutional Review Board (IRB), protocol HUM00250339.

### 4.2.1 Sample Selection

We selected 500 samples from the test split of each dataset for the human evaluation study, to be consistent with our evaluation-only study design. Due to the imbalanced label distributions in SemEval and GoEmotions, we applied weighted sampling with log inverse frequency as the weights to encourage a more representative inclusion of different emotions. We removed samples that were rejected by GPT-4 due to its content policy (17 in ISEAR, 12 in SemEval, 14 in GoEmotions). Since one of our main goals is to investigate the differences between their annotations, we dropped samples where the two sources gave the exact same label(s). This left 990 samples (out of 1500) for human evaluation: 124 from ISEAR, 438 from SemEval, and 438 from GoEmotions.

### 4.2.2 Crowdsourcing Experiments

We designed a human evaluation study with the goal of comparing and understanding the disagreement between human and GPT-4 annotations. We presented the evaluators with text samples alongside labels from both GPT-4 and human annotators, randomized and without revealing their source. We asked them to provide feedback on three aspects:

**Emotional Ambiguity.** We ask "Do you feel confident that you can describe the emotion expressed in the sentence(s)?" with three options "Yes", "No" and "Maybe".

**Perceived Accuracy.** We then present annotations from both sources (as Option A or Option B) and ask the evaluators to rate "How accurately do you think that the description in Option A/B reflects the text writer's emotion?" on a 7-point Likert scale (1-totally inaccurate, 7-totally accurate).

**Preference.** Finally, to make a direct comparison, we ask "If you have to choose one, which emotion description do you agree more with?"

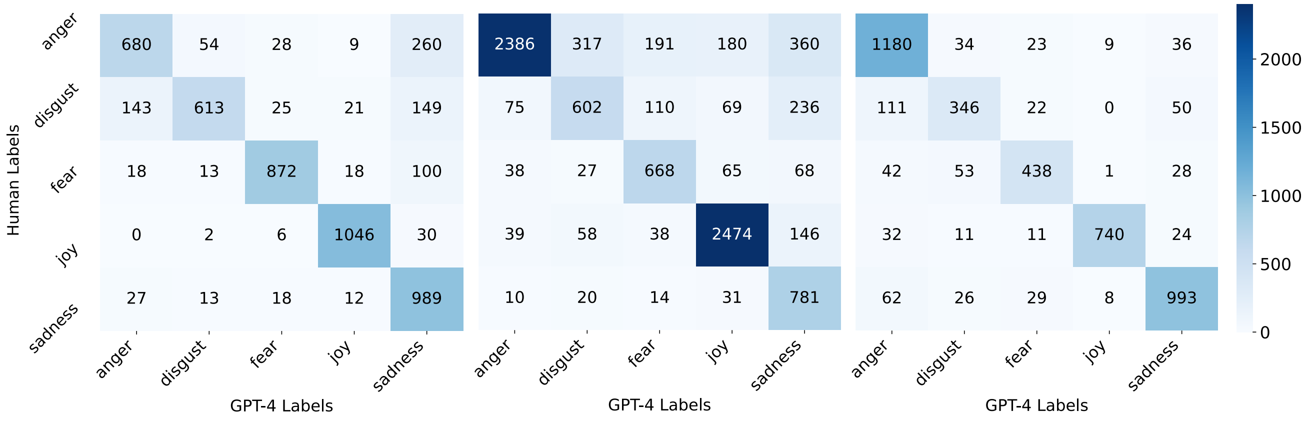
We aimed to obtain three evaluations on each sample. Each evaluator was assigned 50 samples, to keep session time manageable. We implemented the annotation interface with Potato [55], a web-based text annotation tool. We hosted the annotation webpage on an AWS server and recruited participants from Prolific. The selection criteria included: being native speakers of American English, at least 18 years old, and living in the United States. The participants were informed that the goal of this study was to understand how people interpret emotional expressions in text, and they all provided their consent to participate. We received 2948 evaluations from 59 participants (968 samples got three annotations on each, and 22 samples only got two due to connection issues). The average completion time was 20 minutes 42 seconds, resulting in an average compensation of $11.60/hour.

## 4.3 Label Distributions and Agreement Analysis

We first analyze the label distributions and disagreements between human and GPT annotations. We visualize the disagreements with confusion matrices in Figure 1. For clarity

[1] https://openai.com/policies/usage-policies/

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3584775



(a) ISEAR (b) SemEval (c) GoEmotions

Fig. 1: Disagreements between Human and GPT-4 Annotations, visualized as confusion matrices.

and to compare across datasets, we only show results with the five emotion classes that are shared in all three datasets: *anger, disgust, fear, joy* and *sadness*. For multilabel datasets, we define the confusion matrix based on the overlap and differences between human and GPT-4 annotations: if an emotion is present in both sets, we increase the count in the diagonal of the matrix for that emotion. If an emotion is present in the human labels but not in the GPT-4 annotations, and another emotion is present in the GPT-4 annotations but not in the human labels, we increase the count in the off-diagonal cell corresponding to the two emotions by one. For example, if the human labels on a sample are {*admiration, joy*} and GPT-4 set is {*joy, love, excitement*}, we record the agreement on the diagonal element of *joy-joy*, and we record confusion of *admiration-love* and *admiration-excitement*.

We see that most samples fall on the diagonal of the confusion matrices, indicating that GPT-4 annotations generally align with human annotations. Besides, as is expected, it is more common to see disagreements between similar emotion labels: confusion between a positive emotion (e.g., *joy*) and a negative one (e.g., *anger*) is less common than confusion between two negative emotions (e.g., *anger* and *disgust*). Finally, we notice that the confusion matrices are largely asymmetric. For example, in the ISEAR dataset, GPT-4 more often takes human-perceived *anger* as *sadness* (260 samples) than the reverse (27 samples). Such differences, however, do not generalize across datasets: the same *anger-sadness* confusion is shown in SemEval, but in GoEmotions the numbers are closer and the direction is reversed. These findings suggest potential perspective differences between GPT-4 and humans, specific to datasets, emotion categories, and annotation processes. This observation aligns with previous research showing a significant performance variation across emotions [25], which has been attributed to the sensitivity of LLMs to word choice and usage. We leave more in-depth explorations on this perspective difference, for example identifying the factors contributing to the directions of the difference, to future work.

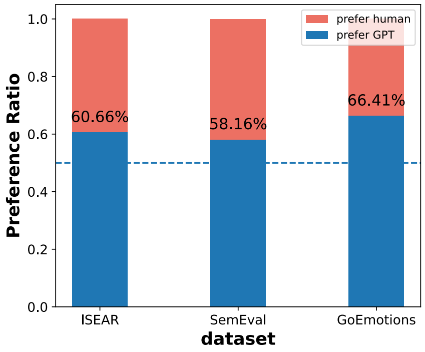


Fig. 2: Proportion of human evaluators' votes favoring human annotations versus GPT-4 annotations, across datasets.

## 4.4 Human Evaluation Results

### 4.4.1 Preference

We first look at the responses to the "preference" question. Figure 2 shows the votes for human versus GPT-4 annotations. GPT-4 annotations were significantly more preferred than human annotations (overall 62%), and this trend held across all three datasets (ISEAR 60.7%, SemEval 58.2%, GoEmotions 66.4%). We also ran per-evaluator statistics to test the between-person consistency. The vast majority (53 out of 59 evaluators, 89.8%) preferred GPT annotations on more samples, while three (5.1%) preferred human annotations more and three (5.1%) indicated equal preference. Interestingly, in our previous work, we compared GPT-4 predictions to those from smaller models finetuned on human labels and found comparable performance when human labels were used as the ground truth [29]. However, the results of this human evaluation study present an even more favorable picture for GPT-4. This conveys an important message that the common method that evaluates LLMs against human labels [25], [28] is prone to underestimate their performance and may give misleading results.

Further, comparing the datasets, we find that the preference discrepancy is larger in GoEmotions, where the label space is larger compared to the other two datasets. We hypothesize that as the label space gets larger and more

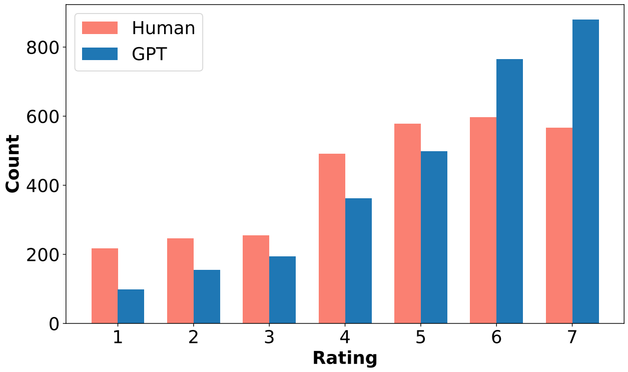


Fig. 3: Perceived accuracy ratings on a scale of 1 (totally inaccurate) to 7 (totally accurate).

complicated, humans may be more challenged and tend to make more mistakes due to the increased cognitive load [56], while GPT-4 is less affected, especially with proper prompting methods. We discuss this hypothesis in more detail in Section 4.4.3, and we run follow-up human annotation studies to further explore the complexity of the label space as a factor in human and GPT-4 performance (Section 5).

#### 4.4.2 Perceived Accuracy Ratings

We then look at the individual perceived accuracy ratings and compare those on GPT-4 versus human annotations, to gain more insights into human preference results. We compare the total number of samples that fall into each rating category, as shown in Figure 3. The results reveal a clear and consistent advantage for GPT-4 annotations: human annotators generated a greater number of labels deemed inaccurate (Rating $\leq$ 3, Human 24.3% vs. GPT-4 15.1%), suggesting a higher probability of errors. In contrast, GPT-4 demonstrates stronger performance in identifying emotions deemed totally accurate by evaluators, indicating good comprehension of the complexity of emotion labels and subtlety of emotion expressions. We further compare the accuracy ratings on each dataset in Table 3. We see that the trend also holds on each dataset, adding to the robustness of our findings. What's more, as the label space expands, both human annotators and GPT-4 are more likely to produce labels rated as fully accurate (see last row in Table 3, across all datasets). This behavior is both reasonable and desirable, as when the label space is limited, it lacks the necessary granularity to capture subtle emotional distinctions, making it impossible to provide perfectly accurate descriptions. In contrast, a larger label space increases the likelihood of encompassing the correct label(s), thus facilitating "totally accurate" outcomes. In Section 5, we will further explore the influence of label space complexities with a dataset-controlled annotation study.

#### 4.4.3 Confidence and Agreement

We assess the confidence and agreement among human evaluators to understand the perceived ambiguity of this task and perceptual differences across evaluators. When asked if they could confidently describe the emotions expressed in the text, evaluators responded "Yes" for 74% of the samples, "No" 18.2%, and "Maybe" 7.7%. Although most samples were found to convey clear emotions, evaluators disagreed a lot on their preference: among annotations marked with confidence, only 59.2% of samples with two annotations had agreement (i.e., both evaluators preferred the same label source), and 40.5% of samples with three annotations had agreement. This highlights significant variation and subjectivity in emotion perception: even when selecting between two options, agreement remains relatively low.

#### 4.4.4 GPT-4 Weakness Analysis

We also analyze the samples to understand whether and how certain text features affect GPT-4's emotion classification performance. We used Linguistic Inquiry and Word Count (LIWC) [57], a tool used to analyze text for psychological and linguistic content. It quantifies the occurrences of 73 word categories in a text, including words that convey emotional and psychological states (e.g., positive emotion, fear), as well as semantic information (e.g., adverb, conjunction) [58]. We extracted LIWC features for each sample, and we augmented the feature set with five additional semantic features commonly used for Twitter data [59], [60]: text length, word count, emoji count, hashtag count, and mention count (tagging another user with "@"). These features allow us to examine if certain types of emotional or semantic content are more likely to mislead or challenge LLMs.

We ran a Logistic Regression (LR) model (N=990) using the text features as input and the preference from the human evaluation study as the outcome variable (1 if GPT-4 labels were preferred over human labels by majority vote, 0 otherwise). We first ran independent t-tests on individual features for feature selection [61] and kept the 10 features with lowest p-values as the input to our model.

We found that higher frequencies of mentions, prepositions, future focus and interrogatives had a significant ($p < 0.05$) negative effect on GPT-4 being the preferred annotator, while the use of impersonal pronouns positively predicted the preference for GPT-4. Prior research has highlighted challenges for LLMs in understanding temporal constructs [62] and social cues beyond the text [63], which may explain some of these patterns. However, given the limited size of our data, further investigation is needed to interpret these findings meaningfully.

## 5 Feasibility of GPT-4 Aided Emotion Annotation

Our experiments in Section 4 revealed the potential of GPT-4 in emotion recognition. However, we also identified weaknesses. One notable concern is the instability and unpredictability that often accompany LLMs. They are sensitive to both training data and prompting methods, which can greatly impact their performance [64]. Therefore, using GPT-4 to perform annotations without human oversight can be risky. Furthermore, as shown in Figure 1 and as discussed in Section 4.3, there may be systematic differences between GPT-4 and human perspectives. While it is crucial to mitigate human errors, we also require that annotations accurately reflect human perspectives. Therefore, in this section, we propose and evaluate two methods for incorporating GPT-4 into emotion annotation pipelines, with the goal

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3584775

| Label | Human | GPT | ISEAR (7 classes) | | SemEval (11 classes) | | GoEmotions (27 classes) | |
|---|---|---|---|---|---|---|---|---|
| | | | Human | GPT | Human | GPT | Human | GPT |
| 1-Totally Inaccurate | 217 (7.4%) | 98 (3.3%) | 9.3% | 3.6% | 6.1% | 3.6% | 8.1% | 3.0% |
| 2 | 245 (8.3%) | 155 (5.3%) | 6.0% | 7.1% | 9.0% | 5.3% | 8.1% | 4.7% |
| 3 | 255 (8.6%) | 193 (6.5%) | 8.5% | 8.5% | 9.9% | 6.9% | 9.0% | 6.4% |
| 4 | 491 (16.7%) | 364 (12.2%) | 19.1% | 15.8% | 16.9% | 13.9% | 15.7% | 9.4% |
| 5 | 577 (19.6%) | 498 (16.6%) | 22.1% | 18.5% | 19.4% | 18.1% | 18.5% | 18.3% |
| 6 | 704 (23.9%) | 463 (15.4%) | 17.8% | 24.6% | 21.0% | 24.0% | 20.2% | 27.6% |
| 7-Totally Accurate | 566 (19.2%) | 879 (29.8%) | 16.4% | 25.4% | 18.2% | 25.9% | 21.1% | 35.1% |

TABLE 3: Percentage of rating scores Human and GPT-4 annotations receive, overall and within each dataset.

of harnessing the strength of both human and automated labelers. In our first approach, we use GPT-4 as a pre-annotation label filter to dynamically present a smaller set of classes to human annotators. In our second approach, we use GPT-4 as a sample filter to flag potentially low-quality samples. Below we describe each method and our evaluation experiments.

## 5.1 Pre-filtering, label-level

There is a trade-off between the benefit of larger label spaces and increased cognitive load (Section 4.3 and 4.4.2, also [11]). We hypothesize that we can reduce cognitive load while preserving label diversity by using GPT-4 to dynamically drop unlikely labels for each sample before presenting them to human annotators. We prompt GPT-4 in a zero-shot manner with text samples and a list of emotion options. The goal of the filter step is to include all possible classes; it is less important to avoid false positives as these can later be identified by human annotators. Therefore, instead of asking GPT-4 to make selections, we ask it to go through the emotions one by one and indicate if each is **possibly** expressed in the sample. We provide the list of emotion options along with the text samples, rather than in the system prompt. In a preliminary analysis of a small exploration set, we found that this encouraged the inclusion of more labels and significantly reduced false negatives.

> GPT-4 prompt for emotion Pre-filtering
> *"You are an emotionally intelligent and empathetic agent. You will be given a piece of text and a list of emotions. Your task is to determine which emotions are present in the text. Please go through the emotion list one by one and think about if the emotion is possibly present in the text. Please respond with each emotion plus "yes" to indicate it's possibly present, or "no" to indicate it's definitely not present. If you are not 100% sure, please select "yes". Reply with only the list of emotions words plus your response, separated by newline."*

### 5.1.1 Evaluation Setup

To evaluate the feasibility of this approach, we conducted human-annotation experiments on the same set of samples but three different label space setups:

1) **Small:** We take the 11 emotion classes from SemEval to represent a relatively small label space.
2) **Large:** We take the union of the emotion classes from SemEval and GoEmotions (30 classes in total), to represent an extensive set of emotion labels.
3) **GPT-4 Pre-filtered:** We take the large set in 2) and reduce it with GPT-4, as described above.

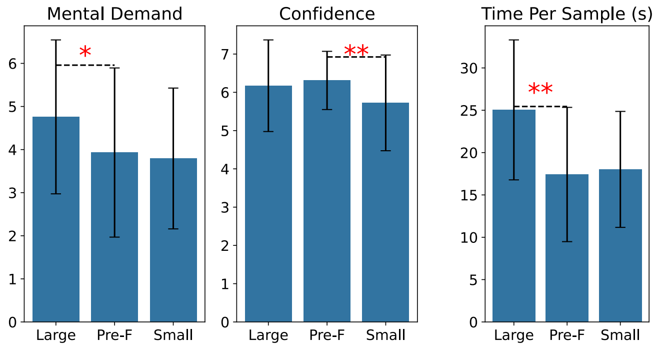
We use a between-subject design: the label sets are fixed for both the Small and Large sets, where participants may gradually gain familiarity with the labels. If we mix those setups in one annotation session, such familiarity is potentially disrupted. Therefore, we assign each participant to one of three groups, each having the same set of samples and one of the three label sets. We ask the annotators to select all applicable labels from the label list, plus an extra "None of the above / Others" option. We also include a question of whether they feel restricted by the options and would use other words to describe the emotion(s).

We used text samples from the GoEmotions dataset for the coverage of diverse emotions in its samples. We took the set of 486 GoEmotions samples we used for the evaluation study (Section 4.2.1). For samples where GPT-4 did not output any candidates (N = 4), we defaulted their labels to "neutral". We recruited 29 annotators for each group (i.e., Small, Large, and GPT-4 Pre-filtered) through crowdsourcing. We used the same crowdsourcing platform and setups as described in Section 4.2.2. Each participant annotated 50 samples, and each sample got 3 annotations.

### 5.1.2 Results

For the evaluation of the pre-filtering setup, we focus on three aspects: 1) cognitive load, indicated by subjective reports and time to completion; 2) label reliability, indicated by the agreement level among annotators; and 3) label coverage, i.e., the pre-filtered set should reasonably cover the labels human annotators selected from the Large set.

**Cognitive Load and Annotators' Experience.** We measured the cognitive load of the annotators in two aspects: perceived load [65], as a subjective measure, and time to completion [66], as an objective measure. We used the NASA Task Load Index [67] as our cognitive load scale. We excluded the question on physical demand since it was not directly relevant in our task, and we also removed the question on temporal demand, which was measured by time to completion. We asked the participants to rate their feelings on four aspects on a 7-point scale at the end of their session: Mental Demand, Confidence, Effort, and Frustration. We found that a large label space significantly increased the mental demand of the annotators compared to the small set (see Figure 4a). However, a small label space did have a drawback: annotators reported feeling more restricted by the options (11.6% of samples on the Large set, 13.9% on Pre-filter, 33.5% on Small). Consequently, significantly lower confidence was reported on the Small set compared to the other two (Small 5.72±1.25, Large 6.17±1.20, Pre-filter 6.31±0.76). No significant differences were found in Effort and Frustration. We also compared the time the annotators spent on each sample. We excluded samples that took more



(a) Subjective ratings

(b) Time per sample

Fig. 4: Comparison of the cognitive load on different label sets. The bars show the mean of the ratings, and error bars show the standard deviation. Significance tests are run between Pre-Filter and Large/Small sets. Dependent t-test is used in (a) because each annotator provided an overall rating for the whole session, and annotators were assigned different samples. Independent t-test is used in (b) because we measured the avg. time spent on each sample, and the sets share the same samples. *:p<0.1, **:p<0.05.

than 60 seconds, as they were outliers in the time distribution and likely indicated a pause in the task. Annotators spent an average of 17.41 ± 7.93 seconds on the Pre-filter set and 18.02 ± 6.85 seconds on the Small set, while much longer (25.04 ± 8.27 seconds) on the Large set (Figure 4b).

**Agreement.** We use the Jaccard Index (JI) [68] to measure the agreement between two annotators on each sample. JI is an agreement measure for multi-label classification tasks, defined by the size of the intersection of two label sets divided by their union. We calculated the average JI among pairs of annotators on each sample and the average across samples on each set. The Small set has the highest agreement of 0.29±0.34, slightly higher but not significantly different from the Pre-filter set (0.28±0.30, independent t-test p=0.36). The Large set has the lowest JI of 0.20±0.24, significantly lower than the Pre-filtered and Small sets (both p<0.05).

**Label Coverage.** Finally, we evaluate whether the filtering step retains potentially correct labels, i.e., the labels selected by humans in the Large set group. Following the approach of GoEmotions [37], we obtain aggregated labels from each set by using emotion classes that are selected by at least two annotators out of three in our new annotations. If no emotion class is selected for one sample, it is defaulted to "neutral". We first compare the chosen class labels with the Pre-filtered candidates: an average of 90.19% of the labels selected in the Large set were included in the Pre-filter set, indicating a reasonably low false-negative rate. In addition, the labels annotators chosen from the Pre-filter set have an agreement of 0.30±0.32 JI with the Large set, which is comparable to (even slightly higher than) the within-set agreement levels, and is much higher than the agreement between the Small and the Large sets (0.15±0.28, p<0.05).

Together, the results show that the Pre-filtered set can match the advantages of both the Small and the Large sets: it is less mentally demanding for annotators, takes less time to complete, and yields higher agreement among annotators.

| Model | Test Label | Human | | Filter | | Random_F | |
|---|---|---|---|---|---|---|---|
| | | F1 | UAR | F1 | UAR | F1 | UAR |
| BERT | H | **0.472** | 0.465 | 0.442 | **0.499** | 0.442 | 0.421 |
| | F | 0.578 | 0.530 | **0.620** | **0.590** | 0.535 | 0.476 |
| DBERT | H | 0.434 | 0.401 | **0.436** | **0.472** | 0.427 | 0.396 |
| | F | 0.526 | 0.462 | **0.588** | **0.551** | 0.504 | 0.441 |

TABLE 4: BERT and DistilBERT model performance, fine-tuned and tested with different label sets. Test Label: H: Human, F: Filter. For training, the Human set has 42,287 samples and the Filter set has 16,592 samples. The "Random_F" training set is a set randomly downsampled from the Human set to the size of the Filter set. Better performances are shown in bold, respectively for F1 and UAR.

## 5.2 Post-filtering, sample-level

In Section 5.1.2, we show some benefits of using GPT-4 for pre-annotation to collect new labels. In this section, we investigate a second approach: when a dataset with human-annotated labels is available, we propose to use GPT-4 as a quality checker to filter out potentially low-quality labels. Specifically, we compare the labels from human and GPT-4 annotation and drop the samples where the two sources totally disagree: i.e., they selected different labels for single-label classification datasets, or where they do not contain any overlapping labels for multi-label classification datasets. By applying this filtering step to GoEmotions, we obtained a much smaller Filtered set of 16,592 samples (out of 42,287).

### *5.2.1 Evaluation Setup*

Since the post-filtering step removes samples where GPT-4 and human annotations disagree, it is expected to remove samples with mistakes in annotations, resulting in higher-quality labels. While this generally benefits model training, this filtering step also decreases the number of samples and potentially the diversity or ambiguity in the samples. Therefore, a key question is whether this trade-off eventually enhances or hurts model training outcomes. To evaluate this, we train smaller models with either the entire labeled GoEmotions training set, or the smaller Filtered set. We measure the performance on its test set as an indicator of the usefulness of the labels. We report performance on both the whole test set and a filtered test set.

**Base model selection.** We choose two models from the BERT family for our training experiments: BERT [69] and DistilBERT [70]. BERT is one of the earliest transformer-based LLMs that gained broad attention, and it has been used as a baseline for many NLU tasks [71], including in the GoEmotions paper [37]. With 110 million parameters, BERT is significantly smaller than leading LLMs like GPT-4, making it practical for use on most modern GPUs. DistilBERT is a distilled version of the BERT model with a 40% reduction in the number of parameters while delivering comparable performance in multiple NLU tasks. We compare the models trained on the original human labeled set versus the Filtered set where samples that GPT-4 totally disagree with are dropped. We finetune the models for 30 epochs with a learning rate of 1e-5, and we select the best model measured by performance on the validation set. We report the performance on the test split of the human and Filtered set.

#### 5.2.2 Results

Results are presented in Table 4. The filtered set, despite comprising less than 40% of the samples in the full set, consistently leads to better model performance across models (both BERT and DistilBERT) and test sets (both the full and filtered test sets), with one exception of the F1 score when tested in-domain on the human-labeled set. To isolate the effect of training sample size reduction, we included a training set that was randomly sampled from the Human set and matched the size of the Filter set (16,592). As expected, this smaller set led to a performance drop, with all metrics lower compared to the Human and Filtered set. This further highlights the effectiveness of our post-filtering approach, which achieved better performance with much fewer samples. Together, these results show the potential of GPT-4 to flag possibly low-quality samples, thus improving model performance as well as training efficiency.

## 6 Discussion

Our work examines the design choices involved in emotion annotation and investigates how LLMs, specifically GPT-4, perform in this context and where they may offer new opportunities. In the first part of our study (Section 4), we evaluated GPT-4's ability to classify emotions across three datasets with varying domains and label space complexity. We found that GPT-4 predictions generally align with human-annotated labels. In addition, a human evaluation study revealed preferences for GPT-4 labels over the original human annotations, highlighting the value of human-centered evaluations and raising questions about how LLMs should be evaluated for emotion recognition tasks. We also compared GPT-4 and human labels across different label spaces. Results suggest that larger label spaces allow nuanced emotion descriptions, which are perceived as more accurate by human evaluators, while smaller spaces are less cognitively demanding and can potentially lead to fewer human mistakes.

While our results demonstrate the potential of GPT-4 as an effective annotation assistant, fully replacing human annotators with LLMs remains questionable and carries risks. Our analysis (Section 4.3) revealed systematic differences between human and GPT-4 labels, suggesting that LLMs may reflect distinct perspectives that risk narrowing the diversity of emotional interpretations. Overdependence on such models could lead to the loss of subtle, context-sensitive judgments that humans naturally bring to the task.

Based on the continued importance of human perspectives in emotion recognition, we further explored ways to integrate GPT-4 into annotation processes, focusing on the GoEmotions dataset. We found that GPT-4 can serve as a pre-annotation label filter to dynamically exclude highly unlikely labels before presenting them as options to human annotators. Our human annotation study showed that, compared to traditional methods, GPT-4 could effectively reduce more than 70% of options while preserving more than 90% of human-selected labels. This approach leverages the expressivity of larger label spaces and the reduced cognitive load and higher annotator agreement associated with smaller label spaces. What's more, on annotated datasets, GPT-4 can act as a post-annotation sample filter to flag potentially low-quality labels. Models trained on the filtered dataset, although much smaller in training data size, achieved better performance than the original full set with human annotations.

## 7 Limitations and Future Work

Emotion is a rich research topic, involving subjective, cultural, and contextual dimensions that make both recognition and evaluation inherently challenging (see Section 2.2). While our study demonstrates GPT-4's potential as an annotation assistant, we did not systematically analyze its failure cases beyond the confusion matrices in Section 4.3. Understanding when and why GPT-4 diverges from human annotations remains an important direction for future work, especially for applications in sensitive domains.

We did not empirically examine issues of bias or fairness. Prior work has shown that LLMs may reflect and amplify societal biases (e.g., [72], [73]), posing risks in sensitive domains like healthcare and mental health (e.g., [74]). These concerns highlight the need for careful oversight and further research before deploying LLM-based emotion annotation in high-stakes contexts.

We limited our discussion to classification tasks. Our previous work conducted preliminary experiments on dimensional emotion annotation using the Emobank dataset [29]. We found that GPT-4 labels had high Pearson Correlation with human labels, but the Mean Absolute Error was also high. This indicates that GPT-4 was able to compare the relative activation or positivity levels of emotional expressions but the scale or distribution of the output numbers may need further calibration, raising different challenges compared to classification tasks. Existing work has shown that language anchors help LLMs to understand dimension scales [75]. As dimensional label spaces gain more popularity [76], future research could explore ways to better leverage LLMs in dimensional emotion annotation.

Additionally, our pre- and post-filtering methods (Section 5) serve as preliminary demonstrations of the feasibility and potential of GPT-4-assisted annotation rather than as definitive solutions. Future work could incorporate more refined approaches to further improve performance. While our focus is not on comparing different perspectives or methods in human annotation processes (e.g., self-reported vs. third-person annotations, or crowdsourcing vs. in-house annotations), previous studies have examined these aspects [77]–[79]. Finally, prompting techniques are not the focus of this paper, but we acknowledge the sensitivity of the models and the importance of good prompts. We direct interested readers to related studies [64], [80]–[82].

## 8 Conclusion

In this work, we conduct a comprehensive evaluation of GPT-4's emotion classification performance and its potential to aid annotation processes. We present encouraging results along with discussions on the complexities and challenges associated with various design choices in emotion annotation studies. Our findings underscore the importance of carefully rethinking these choices with LLMs' capability in mind. We highlight the need for evaluation metrics that better align with human perspectives and the strong promise of using LLMs as tools to aid annotation efforts.

We make our prompts and code publicly available at https://github.com/chailab-umich/GPT-4-Emotion-Annotation.

## 9 Acknowledgments

This material is based in part upon work supported by the National Science Foundation (NSF IIS-RI 2230172) and National Institutes of Health (NIH R01MH130411).

## References

[1] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[2] G. Pei, H. Li, Y. Lu, Y. Wang, S. Hua, and T. Li, "Affective computing: Recent advances, challenges, and future trends," *Intelligent Computing*, vol. 3, p. 0076, 2024.

[3] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Computers & education*, vol. 142, p. 103649, 2019.

[4] Y. Liu, K. Wang, L. Wei, J. Chen, Y. Zhan, D. Tao, and Z. Chen, "Affective computing for healthcare: Recent trends, applications, challenges, and beyond," *arXiv preprint arXiv:2402.13589*, 2024.

[5] L. Tian, S. Oviatt, M. Muszynski, B. Chamberlain, J. Healey, and A. Sano, *Applied Affective Computing*. Morgan & Claypool, 2022.

[6] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[7] G. Haralabopoulos, M. Tsikandilakis, M. Torres Torres, and D. McAuley, "Objective assessment of subjective tasks in crowdsourcing applications," in *LREC 2020 Workshop on" Citizen Linguistics in Language Resource Development"*, 2020.

[8] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[9] L. Williams, M. Arribas-Ayllon, A. Artemiou, and I. Spasić, "Comparing the utility of different classification schemes for emotive language analysis," *Journal of Classification*, vol. 36, pp. 619–648, 2019.

[10] M. Jaiswal, Z. Aldeneh, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7415–7419.

[11] C. Busso, M. Bulut, S. Narayanan, J. Gratch, and S. Marsella, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction*, vol. 7, no. 17, pp. 110–127, 2013.

[12] E. Öhman, "Challenges in annotation: Annotator experiences from a crowdsourced emotion annotation task." in *DHN*, 2020, pp. 293–301.

[13] Z. Callejas and R. Lopez-Cozar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416–433, 2008.

[14] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[15] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[16] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[17] J. A. Hall, J. D. Carter, and T. G. Horgan, "Gender differences in nonverbal communication of emotion," *Gender and emotion: Social psychological perspectives*, pp. 97–117, 2000.

[18] J. W. Gurera and D. M. Isaacowitz, "Emotion regulation and emotion perception in aging: A perspective on age-related differences and similarities," *Progress in brain research*, vol. 247, pp. 329–351, 2019.

[19] A. G. Gitter, H. Black, and D. Mostofsky, "Race and sex in the perception of emotion," *Journal of Social Issues*, vol. 28, no. 4, pp. 63–78, 1972.

[20] L. Canales, W. Daelemans, E. Boldrini, and P. Martinez-Barco, "EmoLabel: Semi-automatic methodology for emotion annotation of social media text," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 579–591, Apr. 2022.

[21] A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5190–5194.

[22] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation: A survey," *arXiv preprint arXiv:2402.13446*, 2024.

[23] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 30, p. e2305016120, Jul. 2023.

[24] S. Feng, G. Sun, N. Lubis, C. Zhang, and M. Gašić, "Affect recognition in conversations using large language models," *IEEE Computational Intelligence Magazine*, Sep. 2023.

[25] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Bias in emotion recognition with ChatGPT," *arXiv [cs.RO]*, Oct. 2023.

[26] S. Latif, M. Usama, M. I. Malik, and B. W. Schuller, "Can large language models aid in annotating speech emotional data? uncovering new frontiers," *arXiv preprint arXiv:2307.06090*, 2023.

[27] Z. Zhang, L. Peng, T. Pang, J. Han, H. Zhao, and B. W. Schuller, "Refashioning emotion recognition modelling: The advent of generalised large models," *IEEE Transactions on Computational Social Systems*, 2024.

[28] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL*, 2024, pp. 3881–3906.

[29] M. Niu, M. Jaiswal, and E. Mower Provost, "From text to emotion: Unveiling the emotion annotation capabilities of llms," in *Proc. Interspeech*, 2024, pp. 2650–2654.

[30] L. Canales, University of Alicante, Alicante, Spain, W. Daelemans, E. Boldrini, and P. Martínez-Barco, "Towards the improvement of automatic emotion pre-annotation with polarity and subjective information," in *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*. Incoma Ltd. Shoumen, Bulgaria, Nov. 2017.

[31] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[32] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[33] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," *EACL 2017*, p. 578, 2017.

[34] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 986–995.

[35] H. G. Wallbott and K. R. Scherer, "How universal and specific is emotional experience? evidence from 27 countries on five continents," *Social Science Information*, vol. 25, no. 4, 1986.

[36] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 1–17.

[37] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[38] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[39] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the turk: Alternative platforms for crowdsourcing behavioral research," *Journal of experimental social psychology*, vol. 70, pp. 153–163, 2017.

[40] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2015.

[41] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[42] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, 2023.

[43] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is chatgpt equipped with emotional dialogue capabilities?" *arXiv preprint arXiv:2304.09582*, 2023.

[44] A. N. Tak and J. Gratch, "Is GPT a computational model of emotion?" in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[45] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, and M. R. Lyu, "Emotionally numb or empathetic? evaluating how llms feel using emotionbench," *arXiv preprint arXiv:2308.03656*, 2023.

[46] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[47] A. N. Tak and J. Gratch, "Gpt-4 emulates average-human emotional cognition from a third-person perspective," *arXiv preprint arXiv:2408.13718*, 2024.

[48] X. Hong, Y. Gong, V. Sethu, and T. Dang, "Aer-llm: Ambiguity-aware emotion recognition leveraging large language models," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[49] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5487–5496.

[50] Z. Liu, B. Liu, P. Thompson, K. Yang, and S. Ananiadou, "Conspemollm: conspiracy theory detection using an emotion-based large language model," in *ECAI 2024*. IOS Press, 2024, pp. 4649–4656.

[51] A. H. Nasution and A. Onan, "Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks," *IEEE Access*, 2024.

[52] B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing, "Is gpt-3 a good data annotator?" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 11 173–11 195.

[53] T. Feng and S. Narayanan, "Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 116–12 120.

[54] T. Hu and N. Collier, "Quantifying the persona effect in llm simulations," *arXiv preprint arXiv:2402.10811*, 2024.

[55] J. Pei, A. Ananthasubramaniam, X. Wang, N. Zhou, A. Dedeloudis, J. Sargent, and D. Jurgens, "Potato: The portable text annotation tool," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2022, pp. 327–337.

[56] A. Chernev, U. Böckenholt, and J. Goodman, "Choice overload: A conceptual review and meta-analysis," *Journal of Consumer Psychology*, vol. 25, no. 2, pp. 333–358, 2015.

[57] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," 2015.

[58] I. Y. Kilic and S. Pan, "Incorporating LIWC in neural networks to improve human trait and behavior analysis in low resource scenarios," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 4532–4539.

[59] R. Alharthi and A. El Saddik, "A multi-layered psychological-based reference model for citizen need assessment using ai-powered models," *SN Computer Science*, vol. 1, no. 5, p. 291, 2020.

[60] H. R. M. Bello, L. Heilmann, and E. Ronan, "Detecting fake news spreaders with behavioural, lexical and psycholinguistic features." in *CLEF (Working Notes)*, 2020.

[61] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: logistic regression," *Perspectives in clinical research*, vol. 8, no. 3, pp. 148–151, 2017.

[62] Z. Chu, J. Chen, Q. Chen, W. Yu, H. Wang, M. Liu, and B. Qin, "Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models," *arXiv preprint arXiv:2311.17667*, 2023.

[63] M. Choi, J. Pei, S. Kumar, C. Shu, and D. Jurgens, "Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 11 370–11 403.

[64] M. Loya, D. Sinha, and R. Futrell, "Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variations and hyperparameters," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 3711–3716.

[65] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[66] I. D. Wood, J. P. McCrae, V. Andryushechkin, and P. Buitelaar, "A comparison of emotion annotation approaches for text," *Information*, vol. 9, no. 5, p. 117, 2018.

[67] S. Hart, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload/Elsevier*, 1988.

[68] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[69] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*, 2019, pp. 4171–4186.

[70] V. Sanh, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[71] M. V. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[72] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[73] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," 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)*, 2019, pp. 3407–3412.

[74] T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R.-E. E. Abdulnour *et al.*, "Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study," *The Lancet Digital Health*, vol. 6, no. 1, pp. e12–e22, 2024.

[75] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[76] S. Buechel and U. Hahn, "Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation," in *ECAI 2016*. IOS Press, 2016, pp. 1114–1122.

[77] S. Biersack and V. Kempe, "Tracing vocal expression of emotion along the speech chain: Do listeners perceive what speakers feel?" in *ISCA Workshop on Plasticity in Speech Perception*, 2005.

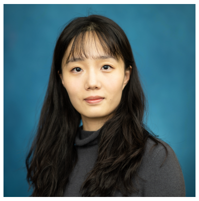
[78] C. Busso and S. S. Narayanan, "The expression and perception of emotions: comparing assessments of self versus others." in *Interspeech*, 2008, pp. 257–260.

[79] D. Y. Kim and C. Wallraven, "Label quality in affectnet: results of crowd-based re-annotation," in *Asian Conference on Pattern Recognition*. Springer, 2021, pp. 518–531.

[80] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Jan. 2023.

[81] M. Binz and E. Schulz, "Using cognitive psychology to understand gpt-3," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, no. 6, p. e2218523120, 2023.

[82] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.



**Minxue Niu** is a Ph.D. Candidate in the Department of Computer Science and Engineering at the University of Michigan. She earned a Bachelor of Engineering degree in Computer Science and Technology from Shanghai Jiao Tong University, Shanghai, China, in 2019. Her research focuses on natural language processing and speech modeling, with an emphasis on understanding human emotion and mood.

**Yara El-Tawil** is a Ph.D. Student in the Department of Computer Science and Engineering at the University of Michigan. She received a dual Bachelor of Science in Computer Science and Biopsychology, Cognition, and Neuroscience from the University of Michigan in 2020. Her research interests center around AI for healthcare, including using speech and activity data to track mood and monitor mental health symptoms, as well as patient-informed design.

**Amrit Romana** received her B.S. degree in Mathematics, and her M.S. and Ph.D. degrees in Computer Science and Engineering from the University of Michigan in 2014, 2020, and 2024, respectively. Her research interests lie at the intersection of machine learning, speech processing, and behavioral analysis, with the goal of developing more accurate and accessible speech-based technologies. She is currently working as a research scientist.

**Emily Mower Provost** (M'11, SM'17) is a Professor in Computer Science and Engineering at the University of Michigan. She received her Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2010. She is a Toyota Faculty Scholar (2020) and has been awarded a National Science Foundation CAREER Award (2017), the Oscar Stern Award for Depression Research (2015), a National Science Foundation Graduate Research Fellowship (2004-2007). She is an Associate Editor for IEEE Transactions on Affective Computing and the IEEE Open Journal of Signal Processing. She has also served as Associate Editor for Computer Speech and Language and ACM Transactions on Multimedia. She has received best paper awards or finalist nominations for Interspeech 2008, ACM Multimedia 2014, ICMI 2016, and IEEE Transactions on Affective Computing. Among other organizational duties, she has been Program Chair for ACII (2017, 2021), ICMI (2016, 2018). Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of the perception and expression of human behavior.