Modeling Affect in Speech and Language in the Presence of Natural Inconsistency

by

Minxue Niu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2025

Doctoral Committee:

Professor Emily Mower Provost, Chair
Professor Joyce Chai
Professor David Jurgens
Professor Lu Wang

Minxue Niu
sandymn@umich.edu
ORCID iD: 0009-0007-8266-9152

© Minxue Niu 2025

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Emily Mower Provost. She has been incredibly patient and supportive. Under her mentorship, I have grown a lot both as a researcher and in many other ways. I also want to thank my committee members, David Jurgens, Joyce Chai, and Lu Wang. I am very grateful to have them on my committee and appreciate their helpful suggestions and insights on my dissertation.

I would like to thank everyone in the CHAI Lab. While I adjust to work and life in a new country, they have been my most important friends. Some of them were both mentors and friends who welcomed me when I first joined the lab: Zakaria Aldeneh, Matthew Perez, Mimansa Jaiswal, and Amrit Roamana. I am grateful for their guidance, support, and friendship. I also want to thank James Tavernor, Aneesha Sampath, and Yara El-Tawil, who were an important support network during my PhD. I always enjoyed our conversations, whether we were talking about research or small things in life.

I was also very lucky to have many fellow PhD students in the department to share experiences, frustrations, and joyful moments with. I'd like to thank Yin Lin, Dandan Shan, Shibo Chen, Junjie Xing, and my roommate Yunjie Pan for the wonderful memories we made together, traveling, sharing food, talking, and going through life in Ann Arbor. Further, I want to acknowledge many wonderful collaborators and mentors: I was lucky to have throughout my PhD journey: Ariana Orvell, Zhaobo Zheng, Liz Norred, Pichao Wang, David Fan, Mohamed El Banani. They provided thoughtful input on my research and gave helpful advice as I explored career paths.

I want to thank my boyfriend Yucheng Shi for his support and companionship over the past few years. Lastly, I want to thank my parents and my family. Their unconditional support carried me through the entire journey.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vi

LIST OF TABLES viii

ABSTRACT x

CHAPTER

1 Introduction 1

1.1 Quantifying Affect 2
1.1.1 Emotion Label Spaces 2
1.1.2 Mood Labels 4
1.2 Inconsistencies and Challenges 4
1.3 Methods and Contribution 6
1.3.1 Addressing modality inconsistency 6
1.3.2 Addressing Annotation Inconsistency 8
1.3.3 Addressing Interpersonal Inconsistency 9
1.3.4 Toward Generalizable Emotion Modeling 10
1.4 Thesis Outline 11

2 Capturing Modality Mismatch for Mood Identification 13

2.1 Introduction 13
2.2 Dataset 15
2.2.1 Emotion labels and crowdsourcing 16
2.3 Methods 16
2.3.1 Emotional MisMatch (EMM) 16
2.3.2 Mood prediction models 18
2.4 Experiments and Results 20
2.4.1 EMM feature analysis 20
2.4.2 Mood prediction 20
2.5 Conclusion 22

3 Acoustic-Semantic Factorization for Rich and Robust Speech Representation 23

3.1 Introduction 23

3.2 Datasets . . . 25
3.2.1 Upstream Finetuning . . . 25
3.2.2 Downstream Evaluation . . . 25
3.3 Methods . . . 27
3.4 Experiments and Results . . . 28
3.4.1 Experiment Setup . . . 28
3.4.2 Downstream Results . . . 30
3.4.3 Upstream Analysis . . . 31
3.5 Conclusion . . . 31
4 Impact of Design Choices on Human and LLM Annotations . . . 33
4.1 Introduction . . . 33
4.2 Datasets . . . 35
4.3 Methods . . . 36
4.3.1 GPT-4 Prompting . . . 36
4.3.2 Automatic Evaluation Metrics . . . 37
4.3.3 Supervised model: Finetuned BERT . . . 37
4.3.4 Human Evaluation . . . 38
4.4 Experiments and Results . . . 38
4.4.1 GPT-4 Zero-shot Performance . . . 38
4.4.2 Impact on Model Training . . . 41
4.5 Conclusion . . . 42
5 Enhancing Emotion Annotation with LLMs . . . 44
5.1 Introduction . . . 44
5.2 Datasets . . . 45
5.3 GPT-4's Emotion Annotation Capability . . . 45
5.3.1 GPT-4 Prompting . . . 45
5.3.2 Human Evaluation Study . . . 47
5.3.3 Label Distributions and Agreement Analysis . . . 48
5.3.4 Human Evaluation Results . . . 49
5.4 Feasibility of GPT-4 Aided Emotion Annotation . . . 53
5.4.1 Pre-filtering, label-level . . . 54
5.4.2 Post-filtering, sample-level . . . 57
5.5 Conclusion . . . 58
5.5.1 Discussion . . . 59
5.5.2 Limitation and Future Work . . . 59
6 Incorporating Annotators' Information for Robust Emotion Recognition 61
6.1 Introduction . . . 61
6.2 Datasets . . . 62
6.2.1 MSP-Podcast . . . 62
6.2.2 MSP-Podcast-Percept: Crowdsourcing an annototor-informed subset 63
6.3 Statistical Analysis . . . 66
6.3.1 Influence of Demographics . . . 66

6.3.2 Influence of Personality . . . 68
6.4 Demographic and Personality-Informed SER Adaptation . . . 70
6.4.1 Base Model . . . 70
6.4.2 Adapter Finetuning . . . 70
6.4.3 Experimental Setups . . . 71
6.4.4 SER Results . . . 73
6.5 Conclusion . . . 74
6.5.1 Discussion . . . 74
6.5.2 Limitation . . . 75
7 Distilling Generalizable Emotion Knowledge from LLMs . . . 76
7.1 Introduction . . . 76
7.2 Methods . . . 78
7.2.1 Obtaining Rich Emotion Annotations . . . 78
7.2.2 Contrastive Distillation . . . 79
7.2.3 Zero-shot Inference . . . 81
7.3 Experiments and Results . . . 81
7.3.1 Datasets . . . 81
7.3.2 Baseline Models . . . 81
7.3.3 Experimental Setup and Metrics . . . 82
7.3.4 Overall Performance . . . 83
7.3.5 Ablation Studies . . . 85
7.3.6 Emotion Space Probing . . . 86
7.4 Conclusion . . . 88
7.4.1 Discussions . . . 88
7.4.2 Limitation . . . 89
8 Conclusion . . . 91
8.1 Summary of Contributions . . . 91
8.2 Limitations and Future Prospects . . . 92
8.3 Ethical Considerations . . . 93
8.3.1 Privacy Protection and Data Use . . . 94
8.3.2 Bias and Fairness . . . 94
BIBLIOGRAPHY . . . 95

LIST OF FIGURES

FIGURE

2.1 (a) Estimated density of emotional mismatch in valence (x-axis, $v_i^{ac} - v_i^{tt}$) and activation (y-axis, $a_i^{ac} - a_i^{tt}$) of euthymic speech. (b, c) Mismatch density differences between symptomatic (b, depressed and c, manic) and euthymic speech. 19

3.1 Multitask Factorization Model Structure. 26

4.1 Disagreements between human and GPT annotations on ISEAR. 39
4.2 Human preference ratio comparing human annotations, GPT-4 classification annotations and GPT-4 generation annotations on emotion classification tasks. . . 39

5.1 Disagreements between Human and GPT-4 Annotations, visualized as confusion matrices. 47
5.2 Proportion of human evaluators' votes favoring human annotations versus GPT-4 annotations, across datasets. 49
5.3 Perceived accuracy ratings on a scale of 1 (totally inaccurate) to 7 (totally accurate). 52
5.4 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. 56

6.1 A screenshot of the crowdsourcing interface page we use to collect annotators' personality information. 64
6.2 Distribution of personality of annotators, along five dimensions (E-Extraversion, A-Agreeableness, C-Conscientiousness, N-Neuroticism, and O-Openness). The boxes show the mean (middle line), first and third quantiles (upper and lower boundaries). The whiskers extend to data points within 1.5 times the interquartile range (the range between the 25th and 75th percentiles). Points beyond indicate outliers. 65
6.3 Distribution of mean-aggregated labels of the 500 samples, from the original Podcast dataset versus our new annotations. 66
6.4 Mean valence and activation ratings across age groups. Blue shadows: 95% confidence intervals. 67
6.5 Adapter tuning model structure. 71

7.1 Our model is trained with rich emotion descriptors generated by GPT-4. During inference, this much smaller model can flexibly perform classification or regression on new label spaces. 77
7.2 Overview of our contrastive distillation model structure. 80

LIST OF TABLES

TABLE

2.1 Definition of mood classes. 15
2.2 *Emotion recognition performance where tt=textual, ac=acoustic, sep=single-task, and multi=multi-task. CCC=concordance correlation coefficient and RMSE=root mean square error. Best results are shown in bold.* 19
2.3 Mood prediction results. Lang. - language baseline, Emo. - emotion-only baseline, All - Lang. + Emo. + EMM. Best results of each task are shown in bold. Paired t-tests were performed on subject-wise performances between the language baseline and other feature sets. Significance was shown with symbols following the numbers: $^{\hat{}}$: 0.05<p<0.1, * : p<0.05. 21
2.4 List of features that are selected across all runs, and the average of their fitted coefficients in HDRS regression. 21

3.1 Summary of downstream evaluation tasks. 25
3.2 Performance on downstream tasks. mtt.- multitask, fct. - factorization. The best performance across all models are highlighted in bold. Underscore indicates the better performance with or without factorization. The "#p_train" column shows the number of trainable parameters in upstream finetuning, while "#p_inference" indicates the number of parameters for inference, to obtain the representations (excluding task-specific head). 29
3.3 Performance on upstream tasks. mtt. - multitask; fct. - factorization. '/' indicates that the model is not trained on and thus unable to perform the task. 30

4.1 Dataset details. label: C-categorical, D-dimensional. 35
4.2 GPT-4 zero-shot vs. BERT finetuned performance across four dataset. Better performances are in bold. 40
4.3 Performance (Macro-F1) of models trained and tested on different combinations of annotations. We show the best performance on each test set (per column) in bold. 42

5.1 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. 46
5.2 Percentage of rating scores Human and GPT-4 annotations receive, overall and within each dataset. 51

5.3 BERT and DistilBERT model performance, finetuned 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. Note that the performance of BERT shown in this table is slightly different from Table 4.3 due to different sampling process. 58

6.1 Linear Mixed-Effects Regression Results with Valence as the dependent variable. 68
6.2 Linear Mixed-Effects Regression Results with Activation as the dependent variable. 69
6.3 Model performance for seen and unseen annotators using different input types.We report mean(std) across runs. #params_t refers to the number of trainable parameters. Significance between the baseline model (underlined) and other models is measured by dependent t-tests across 25 runs: **: $p < 0.001$, *: $p < 0.05$. . . 72

7.1 Ten random samples selected from GoEmotions training set, comparing human-annotated emotion labels and GPT-4 generated descriptive labels. 79
7.2 Zero-shot performance across four datasets. For our model, we train the model with five random seeds and we report the mean and standard deviation (in parentheses) of the performance metrics. The best overall performance is underlined. The best zero-shot performance among comparably sized models is highlighted in bold. 84
7.3 Comparison of emotion space dimensionality. "Train" column shows the performance on GoEmotions test set with GPT-4 labels, measured by micro-F1. "G", "S", and "I" refer to performance on GoEmotions, SemEval, and ISEAR respectively, measured by macro-F1. 85
7.4 Comparison of models trained with GPT-4 generated labels versus human-labels. For models trained with human labels, we also report separate results on classes seen in the training dataset (8 out of 11 in SemEval, 5 out of 7 in ISEAR). Note that GoEmotions performance under human supervision reflects a supervised setting, while all others are zero-shot evaluations. 86
7.5 Top-4 most similar emotion terms retrieved for each target emotion using cosine similarity, using the BERT encoder, our contrastively distilled emotion model (768D), and its reduced 50D version. 87

ABSTRACT

Understanding human affect, including emotions and moods, is crucial for a wide range of applications, from improving mental health monitoring to enhancing user experiences in human-computer interaction. The complex, ambiguous, and subjective nature of human affect poses many challenges in developing affect recognition models, such as disagreements in human labels and misaligned signals across modalities (e.g., text and voice). While these inconsistencies present obstacles, they reflect fundamental aspects of human affect and can offer valuable signals for affect modeling if properly harnessed. By addressing these challenges and effectively making use of the information embedded in these inconsistencies, we can work toward building more reliable, interpretable, and human-aligned affect recognition systems.

This dissertation investigates the causes and effects of inconsistencies in affect models and explores strategies to mitigate their impact or to use them as informative signals. We focus on emotions expressed through speech and language, the most common modes of communication, particularly in interactions with intelligent systems. First, we examine **modality inconsistency**: emotions conveyed through text and vocal modalities do not always align. In the mental health domain, we demonstrate that their mismatch carries important information about people's mood and can serve as signals for mood disorder monitoring. We extend this finding and more broadly show that modeling text and acoustic modalities separately allows for disentangling their respective information, yielding richer and more robust speech representations for various speech understanding tasks. Second, we examine **annotation inconsistency** observed in human emotion labels of text data, highlighting their variability and sensitivity to annotation study designs. We compare human and Large Language Models (LLMs) generated annotations and find that LLMs achieve strong performance. Building on this insight, we propose integrating LLMs into the human annotation workflow, which improves both the annotators' experience and the quality of the labels. We then explore **interpersonal inconsistency** caused by the subjectivity of emotion perception across individuals. We show that emotion perception significantly differs across demographic and personality groups, and incorporating annotator-level information can improve personalized speech emotion recognition models. Finally, drawing on insights from annotation and interpersonal inconsistencies, we propose a contrastive distillation framework that transfers the

generalizable emotion understanding capabilities of LLMs into a lightweight text emotion recognition model. Leveraging the breadth of LLMs' pretraining, the distilled model learns an efficient, emotion-salient embedding space and can seamlessly handle unseen emotion label spaces without extra training.

Together, these findings offer a new perspective on the role of inconsistencies in affective modeling. It is important to understand their origins and implications, develop strategies to mitigate unintended ones, and leverage useful signals from them, to achieve a more comprehensive understanding and modeling of human affect.

CHAPTER 1

Introduction

Understanding human affect, including emotions, moods, and related states, is an important aspect of artificial intelligence [1]. The ability of machines to correctly interpret and respond to human affect is beneficial for a wide range of applications. In mental health, affect modeling can provide tools for detecting mood disorders or monitoring the severity of mood symptoms, offering critical support in diagnosing and managing conditions like depression and bipolar disorder [2, 3]. In human-computer interaction, recognizing emotions can make systems more responsive and empathetic, leading to better user experiences [4]. Beyond these domains, affective computing also plays a vital role in education, entertainment, automative safety, etc., where systems need to respond empathetically to human emotions [5, 6, 7, 8].

The expression and perception of human affect states occur continuously across various modalities, such as language, speech, facial expressions, and body movement [9, 10]. Among these, speech and language stand out among the most commonly used forms of communication currently in human interactions with intelligent systems [11, 12]. Language, through its semantic content (e.g., word choice, sentence meaning), and speech, through its extra acoustic characteristics (e.g., tone, pitch, prosody), provide rich cues for understanding emotional and mood states [13, 14, 15]. For these reasons, emotion recognition from speech and language data has received considerable attention from the research community [16, 17, 18]. Current emotion modeling relies heavily on datasets with human-annotated labels for both training and evaluation. Methods have evolved from manually-crafted features and traditional machine learning algorithms like Support Vector Machines (SVMs) [19, 20, 21, 22], to modern neural network-based approaches, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more recently, transformer-based models [23, 24, 25]. These methods effectively capture both acoustic and semantic features, enabling models to achieve strong performance in predicting human-annotated labels on clean datasets [26].

Despite these advancements, building reliable models that work well beyond carefully-crafted clean datasets remains an important yet challenging goal for systems designed to

be useful in real-world affect understanding. Real-world data introduce significant complications. On the expression side, affective states expressed in natural settings—whether emotions or longer-term moods—can be subtle and variable. For example, compared to samples in datasets where emotions are predetermined, acted on or elicited by prompts, real-world emotion expressions tend to be more subtle and ambiguous [27, 28]. There is also great variability in how individuals express their emotions; for instance, some are more expressive while others are more suppressed [29]; some rely more on vocal cues while others emphasize language content [30]. Individuals also differ in the variability and temporal patterns of their emotional experiences, which are closely related to mood and longer-term affective states [31]. Further, emotion expression is heavily influenced by contextual factors, such as physical or mental health conditions, the nature of the scenario, and the social dynamics with the listener [32, 33]. On the perception side, emotion perception is also highly subjective: human annotation often exhibits variability and randomness, influenced by the annotator characteristics [34, 35, 36] and the annotation protocol [37, 38, 39]. Together, the complex, dynamic, and subjective nature of human emotions introduces ambiguities and variabilities in how affect is expressed, perceived, and annotated, potentially leading to low agreements, conflicting labels, and consequentially unclear research signals and divergent conclusions. These inconsistencies can manifest across multiple stages of affect model development, making the task challenging.

This thesis focuses on the challenges posed by these inconsistencies, aiming to understand their causes, validate their impact, and effectively address them. Specifically, we investigate the sources of variability and ambiguity in affect expression, perception, and annotation, providing empirical evidence of their prevalence. We then propose methods to both mitigate unintended inconsistencies and harness meaningful variability as valuable signals for affect modeling.

1.1 Quantifying Affect

To enable machine learning algorithms to learn about emotions, it is necessary to characterize them using definite, quantitative labels. In the following subsections, we introduce existing approaches to quantifying affect and describe the label space adopted in this work.

1.1.1 Emotion Label Spaces

The most commonly used frameworks for describing emotions fall into two categories: categorical label space, where emotions are represented as one or more pre-defined categories

(e.g., joy, sadness) [40], and dimensional label space, which conceptualizes emotions along continuous axes, such as valence (positive to negative) and activation (excited to calm)[41].

Selecting an appropriate set of emotion labels is still complex and critical. A common approach is to follow established theories of basic emotions. For example, the Basic Emotion Theory proposed by Paul Ekman considers Anger, Disgust, Fear, Happiness, Sadness and Surprise as the six primary emotions that are universal across cultures [40] (which was later augmented with Contempt [42]). Other theories have proposed slightly different category lists [43], usually with fewer than 10 basic emotions. The Emobank [44], DailyDialog [45], and MSP-Podcast [46] datasets adopt Ekman's theory of basic emotions. Other works make small modifications based on existing theories; for example, ISEAR [47] 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 [48]. GoEmotions [49] settled on 27 classes after an iterative refinement process.

In addition to the lack of consensus on basic emotions, the categorical label space also faces other critiques. One notable limitation is its lack of flexibility in capturing the subtlety of emotions, such as distinguishing between emotions that are slightly present versus fully blown [50]. Furthermore, categorical labels can be ambiguous, as the interpretation of the emotion labels needs to be tied to specific language, objects, and contexts [51, 52] (e.g., "fear of physical harm" versus "fear of embarrassment").

Dimensional label spaces address these limitation by representing emotions as points in a continuous, multidimensional space [41]. While various dimensions have been proposed [53], valence (pleasant–unpleasant, or positive-negative) and activation (subdued–activated, or calm-excited) are the two mostly widely accepted dimensions, adopted by popular text and speech emotion datasets like EmoBank [44] and AffectNet [54]. Some frameworks incorporate additional dimensions like Dominance (controlled–controlling) to capture a broader range of emotional experiences [53], which are used in datasets like MSP-Podcast [46] and IEMOCAP [55]. Dimensional spaces are sometimes considered advantageous over categorical approaches for their ability to capture the continuous nature of emotional intensity and to describe emotional experiences as primitives that are not tied to specific objects [52], but they are not without limitations. Critiques include their limited descriptive power in distinguishing qualitatively different emotions [56, 57], as well as the reduced communicability of emotional states between individuals, as representing emotions as coordinates in a valence-activation space is not an intuitive or commonly used method in everyday communication [57].

Together, these two frameworks remain the most widely adopted approaches for simplify-

ing and quantifying emotions. In this thesis proposal, we use datasets with both categorical and dimensional label spaces.

1.1.2 Mood Labels

Mood is defined as "a disposition to respond emotionally in a particular way that may last for hours, days, or even weeks, perhaps at a low level and without the person knowing what prompted the state" [58]. It is generally considered to be more persistent than emotions [59]. Mood disorders are mental health conditions characterized by prolonged, pervasive disturbances in emotion regulation, such as major depressive disorder and bipolar disorder [60]. In this proposal, we focus on mood detection in the context of mood disorders, as it is one of the most important applications of mood modeling. In this context, mood labels correspond to clinical assessments of mood symptom severity, following common approach in psychiatry. Details will be included in the corresponding chapter.

It is widely acknowledged that mood is closely related to emotions: long-term mood consistently affects appraisals and shapes short-term emotions, while emotions affect the energy level monitored by mood [61]. Abnormality in emotion dynamics is a main characteristic of mood disorders. For example, individuals with depression tend to down-regulate and experience less positive emotions [62, 63]; bipolar disorder is associated with extreme fluctuations in emotional states [64]. Consequently, emotion dynamics have been used as signals for automatic detection of mood disorders [65].

1.2 Inconsistencies and Challenges

The aforementioned emotion label spaces have been effective in capturing key characteristics of emotions and providing meaningful quantification for emotion modeling. However, they do not fully account for the nuances and complexities of emotion, resulting in variability that affects modeling practices. In this thesis, we focus on three primary sources of inconsistencies:

Modality Inconsistency. Emotions are naturally expressed and perceived differently through various modalities, such as language, speech, and facial expressions. The emotional signals conveyed through each modality do not always match. A classic example is sarcasm, where vocal tone contradicts the literal meaning of words [66]. Mismatches can also occur due to conscious or unconscious regulation of emotional expression in response to social or contextual dynamics. For instance, individuals may temper their emotional expressions when delivering bad news or recalling negative memories [67, 68]. In the context of mood disorders, blunted affect—a common symptom of several mood disorders—is characterized by reduced

emotional expression in speech [69, 70]. During symptomatic episodes, patients often exhibit reduced emotional variability or intensity, despite strong self-reported emotions [70].

Such inconsistencies in emotion expressions across modalities have important implications for emotion-related speech applications. While sarcasm has garnered some attention [71], the implications of affect blunting and other nuanced emotional discrepancies for annotation and modeling remain underexplored in the speech processing community. Most current systems adopt a single label for each utterance, even though the vocal and textual components of the utterance may convey conflicting emotional information. This approach introduces ambiguity and uncertainty in both data annotation and model interpretation, as it is unclear how such joint labels should be understood. In addition, it does not explicitly account for the discrepancy between modalities, which could potentially serve an important signal for mental health.

Annotation Inconsistency. Even when the modality is fixed, many design choices in annotation experiments can significantly affect human annotators, in terms of both annotation experience and outcomes, introducing much variability and instability in this process. For example, the choice of label spaces [37, 38], the availability and order of context [72, 73, 39], as well as the language used to define emotion categories [74], can all influence annotators and lead to different annotation outcomes. Additionally, the effort and attention devoted to the task varies significantly by individuals, with fatigue being a common factor contributing to reduced performance [38]. For instance, a study evaluating annotation quality across four crowdsourcing platforms found that nearly half of the participants failed at least one attention check, with failure rates reaching as high as 72.9% on the least reliable platform [75]. Even when the same annotators were asked to re-annotate the same samples after a two-month gap, agreement rates only ranged from 66.5% to 78.5%, highlighting the instability of emotion annotations [76]. These factors contribute to label quality issues and complicate the implementation of effective quality control mechanisms. In summary, human annotations are sensitive, variable and unstable, leading to inconsistency in emotion labels.

Interpersonal Inconsistency. Another natural and significant challenge arises from different perspectives in emotion annotations. Low agreement among annotators is a pervasive issue in emotion annotations (e.g., unanimous agreement can easily be below 10% in some datasets [38]). One reason for the low agreement lies in the inherent subjectivity of the task [76]. Individuals naturally vary in their perception of emotion, both for their own and others' emotions [77]. Research has found that demographic factors, such as gender [34], age [35] , and race [36], significantly affect how emotions are perceived. Current widely-used datasets typically have 2-20 annotators per sample, which is insufficient to capture the full range of diverse human perspectives. Consequently, the randomness in annotator group com-

position often leads to low agreement within datasets and inconsistent perspectives across different datasets.

Together, these multiple sources of inconsistency remain underexplored in current emotion and mood modeling approaches. Developing a better understanding of these inconsistencies, along with effective strategies to address them, is crucial for reliable and robust affect modeling.

1.3 Methods and Contribution

This section provides an overview of our work. For each source of inconsistency, we first investigate its underlying causes and quantify its presence and impact. Based on these insights, we propose new approaches to improve the quality and reliability of emotion annotations and affect modeling.

1.3.1 Addressing modality inconsistency

We first examine the inconsistencies between acoustic and textual signals in speech and explores their implications and applications. We demonstrate how this mismatch can be quantified and can serve as a valuable signal in the mental health domain for mood tracking. Next, more broadly, we show that factorizing acoustic and textual embeddings in generalized speech representations improves performance across various downstream tasks, both within and beyond affect-related applications.

1.3.1.1 Capturing Modality Mismatch for Mood Identification

Previous studies on Speech Emotion Recognition (SER) have shown that predicting emotion using textual and acoustic features yields different performances, and combining these modalities consistently outperforms using either alone [16, 78], suggesting the presence of modality-specific information. In the mental health domain, psychiatric research has identified a lack of emotional alignment between modalities as a potential symptom of mood disorders [70]. Building on these findings, we hypothesize that the mismatch between emotions inferred from acoustic and textual modalities can serve as a valuable signal for detecting mood changes.

To explore this, we quantify the mismatch between emotion expressed through language and acoustics, which we refer to as Emotional MisMatch (EMM), as an intermediate step for mood identification. Using a longitudinal dataset collected from individuals with Bipolar Disorder (BD), we show that symptomatic mood episodes exhibit significantly higher EMM

compared to euthymic moods. We propose a fully automatic mood identification pipeline with automatic speech transcription, emotion recognition, and EMM feature extraction. We find that EMM features, although smaller in size, outperform a language-based baseline, and consistently provide improvement when combined with language and/or raw emotion features on mood classification.

Contribution: This work introduces EMM as a novel signal for mood tracking, demonstrating that episodes of symptomatic and euthymic mood exhibit different patterns of EMM with statistical significance. It also proposes a fully automated pipeline to extract and apply EMM features in mood tracking, achieving state-of-the-art performance on a BD dataset.

1.3.1.2 Acoustic-Semantic Factorization for Rich and Robust Speech Representation

Drawing from the findings that modality mismatch in speech can be a valuable signal for mood disorder tracking, we hypothesize that modeling these two modalities separately, or in other words, factorizing the two parts of information, could benefit a wide range of speech-related tasks. This project explores extending this idea to general speech representation learning.

In recent years, self-supervised pretraining has revolutionized speech representation learning by enabling models to capture general information about speech [79, 80, 81, 82]. Similar to LLMs, speech models are often pretrained using tasks like masked signal prediction or next-signal prediction, equipping them with a broad understanding of speech patterns. A common subsequent step involves finetuning these models on the Automatic Speech Recognition (ASR) task, further injecting semantic knowledge into the representations [82, 83]. Together, these pretraining and finetuning phases enable speech models to generalize across diverse downstream tasks, ranging from acoustic-oriented ones like speaker recognition to semantic-oriented tasks like ASR and keyword spotting [84]. However, empirical studies have highlighted two notable gaps. First, acoustic and semantic information are encoded together in shared representations, leading to entanglement that can compromise model robustness. For instance, acoustic variations, such as those found in emotional speech, can negatively impact ASR performance [85]. A second challenge involves layer selection. Like LLMs, speech models consist of deep neural networks, with their typical depths ranging from 6 to 20 layers. Studies have shown that acoustic and semantic information is often encoded in different layers, requiring a task-specific selection process to identify the most relevant layer for optimal performance [86]. This additional step can be computationally expensive and adds complexity to model deployment, especially when tasks require varying degrees of acoustic and semantic information.

To address these issues, we propose a two-branch multitask finetuning strategy that integrates ASR and transcript-aligned audio reconstruction, designed to preserve and disentangle semantic and acoustic information in the final layer of a pretrained model. Experiments with the pretrained Wav2Vec 2.0 [82] model demonstrate that this factorized approach surpasses ASR-only finetuning across multiple downstream tasks, with notable improvements in emotion recognition and ASR robustness on emotional speech.

Contribution: This work introduces a factorized multitask finetuning strategy for learning robust, general-purpose acoustic and semantic speech representations. Experiments demonstrate improved performance across multiple downstream speech tasks, showcasing its ability to preserve information from both modalities and improve robustness in challenging conditions.

1.3.2 Addressing Annotation Inconsistency

This section examines the disagreements and instability that arise from various design choices in annotation processes and explores new LLM-supported approaches to obtain higher quality annotations.

1.3.2.1 Impact of Design Choices on Human and LLM Annotations

Training emotion recognition models has relied heavily on human annotated data, which is sensitive to various design choices involved in annotation. In the meantime, LLMs have shown remarkable performance on many Natural Language Understanding tasks, emerging as a promising tool for text annotation. To understand their feasibility and limitations, it is critical to understand the quality of LLM annotations compared to human annotations and the sources of disagreements, especially under the influence of different design choices. Previous studies have examined the impact of individual annotation design factors (e.g., label space [37, 38], annotation platforms [75]) on human annotations, or some have evaluated the performance of LLM against human labels as ground-truth [87, 88]. However, few studies have considered both aspects together, leading to conclusions that may lack generalizability and fail to capture the interplay between annotation design choices and annotation quality.

To bridge this gap, we conduct a comprehensive evaluation study comparing a leading LLM, GPT-4, with supervised models and humans under different label space setups across in three aspects: agreement with human annotations, alignment with human perception, and impact on model training. Our human evaluation experiment reveals a consistent preference for GPT-4 annotations over humans across multiple datasets and evaluators, especially on a dataset with large, complex label spaces. Furthermore, models trained on GPT-4 anno-

tations demonstrate superior performance in predicting human-adjudicated emotion labels, underscoring the potential of LLMs in emotion recognition applications.

Contribution: This work provides a comprehensive evaluation of GPT-4 versus human text emotion annotations across different label spaces, both dimensional and categorical, and with both objective metrics and human evaluation. Our findings highlight the promise of LLMs in emotion annotation tasks and underscore the need for refined evaluation methodologies.

1.3.2.2 Enhancing Emotion Annotation with LLMs

We have demonstrated that GPT-4 achieves high ratings in human evaluation studies, indicating its potential for emotion annotation. However, it remains unclear what specific strengths GPT-4 offers and how it can best support annotation workflows. To this end, we first extend the experiments in the previous study (Chapter 1.3.2.1) with a larger sample size and more detailed human evaluation feedback. Our results reaffirm GPT-4's strong performance, as indicated by human ratings. At the same time, we identify notable differences between human and GPT-4 emotion perceptions, highlighting the ongoing 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.

Contribution: This work expands on the work in 1.3.2.1 with a larger sample size, more annotators, and more comprehensive analysis, strengthening the findings on LLMs' emotion recognition capability. Further, it proposes and evaluates two methods to integrate GPT-4 in human emotion annotation pipelines, demonstrating their potential to improve annotation quality and efficiency.

1.3.3 Addressing Interpersonal Inconsistency

This section explores whether individual differences among annotators affect their perception of emotion in speech, and whether modeling these differences can lead to more robust and personalized SER systems.

1.3.3.1 Modeling Subjectivity in Speech Emotion Recognition

Low agreement among annotators is a common challenge in speech emotion annotation. While some of this variability can be attributed to unintended human errors, a significant portion stems from the inherent subjectivity of emotion perception [89, 90]. Existing approaches typically aggregate annotations (e.g., averaging ratings across annotators) to approximate a "general population" perspective. However, with limited annotations and a lack of diversity in many datasets, these aggregated labels often reflect biases toward certain populations or groups [91]. Moreover, variations in annotator composition across datasets can lead to label inconsistencies, potentially confusing models, and hindering training if the source of the inconsistency is not accounted for.

To address these challenges, this work investigates the interpersonal inconsistency present in speech emotion annotations. We curated a new dataset with approximately 10,000 samples by crowdsourcing additional annotations for a small set of speech data. We ran statistical tests and models to examine group-wise differences and trends. Through statistical analyses, we find that emotion annotations vary significantly across groups defined by demographic and personality traits. Building on these findings, we propose an adapter-tuning approach that incorporates annotator-level information into SER models. Results show that annotator demographics can improve personalized predictions for annotators seen during training, but generalization to unseen annotators remains limited and is prone to overfitting. These findings highlight both the promise and the challenges of annotator-aware modeling.

Contribution: This work offers a comprehensive investigation into how annotator characteristics shape SER annotation and modeling. We show that annotations vary systematically with demographics and personality, and that modeling this subjectivity through annotator-aware architectures improves personalized emotion label predictions.

1.3.4 Toward Generalizable Emotion Modeling

This section builds on insights from earlier sections on the tradeoffs between label spaces and interpersonal subjectivity in emotion perception. We explore new approaches to develop flexible and broadly applicable emotion recognition systems that can better adapt to the diversity and nuances of human affect.

1.3.4.1 Distilling Generalizable Emotion Knowledge from LLMs

Prior sections have shown that fine-grained emotion label spaces enable a more accurate description of emotion states, but they adds cognitive burden to human annotators. Subjectivity in emotion perception can also lead to variability in human labels. While Large

Language Models (LLMs) demonstrate strong zero-shot emotion understanding across diverse taxonomies, their scale and output format pose practical limitations for integration into real-world systems.

Building on these insights, this project explores how the rich affective knowledge embedded in LLMs' language spaces can be transferred into a smaller, dedicated model without the use of human emotion annotations. We use GPT-4 to generate descriptive emotion annotations, offering rich supervision beyond fixed label sets. By aligning text samples with emotion descriptors in a shared embedding space with contrastive training, our method enables zero-shot prediction on different emotion classes, granularity, and label schema. Our approach is effective across multiple ER datasets and label spaces, outperforming strong baselines of similar size and approaching the zero-shot performance of GPT-4 while being over 10,000 times smaller. Additionally, the learned representation space captures emotional similarity and nuance more effectively than general language embeddings.

Contribution: This work presents a framework for distilling generalizable emotion knowledge from LLMs into a lightweight model. Evaluation across multiple datasets with both categorical and dimensional emotion labels shows that the model enables seamless zero-shot inference on a variety of unseen label spaces. The learned representations also capture emotional nuance and relational structure, offering a compact and flexible alternative to LLMs for emotion recognition.

1.4 Thesis Outline

This rest of this dissertation is organized as follows. Chapters 2 and Chapter 3 examine modality inconsistencies. Chapter 2 presents our work on leveraging emotional mismatch between acoustic and textual modalities for mood identification, while Chapter 3 explores modality factorization in general speech representation learning. Chapter 4 and Chapter 5 address annotation inconsistencies. Chapter 4 focuses on evaluating text emotion annotation performance by humans and GPT-4, with an emphasis on understanding the impact of annotation design choices. Chapter 5 describes two methods for integrating LLMs into annotation pipelines and evaluates their effectiveness. Chapter 6 focuses on interpersonal inconsistencies in speech emotion annotations. It investigates how annotator demographics and personality information affect their annotation, and if annotator information can be incorporated for more personalized and generalizable SER models. Chapter 7 presents a contrastive distillation framework that transfers rich, generalizable emotion knowledge from LLMs into a lightweight model, enabling zero-shot emotion recognition across diverse and unseen label spaces. Chapter 8 concludes this dissertation, where we summarize our contributions and

discuss the limitations and future work. Discussions around ethical considerations are also presented in 8.

CHAPTER 2

Capturing Modality Mismatch for Mood Identification

2.1 Introduction

Affective signals expressed through speech involve the modalities of vocal tone and linguistic content, and they do not always convey the same emotional message. These modality inconsistencies, while frequently treated as noise, may reflect meaningful variations in how emotions are expressed and perceived in natural settings. In this work, we explore modality inconsistency in speech, with a focus on its applicability to mood disorder monitoring.

Providing adequate care for individuals living with mental health conditions represents one of the major challenges in society. Mood disorders are often chronic and recurring, and require ongoing monitoring with consistent access to specialty mental health care providers. There is a substantial gap in the availability of such providers and the demand for service [92]. This motivates the need to develop tools to monitor mood and prioritize individuals for care.

Researchers have proposed automatic mood tracking techniques using various types of data. Speech and language data have been widely used, the benefit being that speech can be recorded in natural environments and has shown promise in automated mood identification [93]. Commonly used linguistic feature sets include frequency counts on categories of words that encode linguistic styles and semantic content [94] (e.g., Linguistic Inquiry and Word Count (LIWC) [95]). Commonly used acoustic feature sets include aspects that capture the frequency content or prosodic patterns of an individual's vocalizations [96]. Another line of research considers the dynamics of emotions as a signal for mood. Survey methodologies (e.g., Ecological Momentary Assessment [97]) have been adopted, where participants complete surveys in which they report their feelings and behaviour in their daily lives. Research has demonstrated the close connection between emotion variations and mood change, supporting the hypothesis that mood identification models benefit from the explicit modeling of emotion characteristics [63]. However, the use of these tools requires that individuals take

actions outside of their day-to-day life, which is difficult to sustain in the long term. There have also been works where emotion recognition models were applied to obtain emotion ratings on speech or text data for mood prediction, eliminating the need for surveys [65, 98]. Those works mostly rely on model training processes to learn the relationship between emotion and mood variations, seldom incorporating domain knowledge obtained from clinical research.

In this work, we aim to bring together the benefits of emotion recognition models and clinical insights. We propose a set of emotion-centered features inspired by the phenomena of Affect Blunting (AB) and Affect Exaggeration (AE). AB refers to when individuals show reduced emotional expressivity despite strong self-reported emotion and is a symptom associated with schizophrenia and depression [70]. Subjects suffering from AB can fail to show normal vocal emphasis patterns [99], where the text content may be reflective of their emotions but the acoustic content is not. AE, the symptom where individuals show exaggerated affect, is a typical symptom in mania [100]. Building on clinical insights, we hypothesize that the inconsistency between the emotions conveyed through text and voice modalities, which we referred to as Emotional MisMatch (EMM), carries important signal for mood and can be used as signals for mood identification models.

We study EMM in a longitudinal dataset collected from individuals with Bipolar Disorder (BP) [101]. We use existing acoustic emotion labels and crowdsource text emotion annotations on a consented and transcribed subset of the dataset [65]. Speech segments from depressed or manic states display statistically significantly more EMM, compared to those from euthymic states. Further, we find that emotions measured from acoustics are more negative and less energetic than emotions measured from text in depressed speech, supporting clinical findings of affect blunting. We then explore how EMM can be incorporated into mood recognition systems. We define a set of EMM features over emotion scores, which aim to capture the dynamic interactions between acoustic and text emotions. We propose a fully automatic pipeline composed of speech recognition to generate transcripts and separate acoustic and text emotion recognition models to automate the extraction of EMM features. We find that with a simple linear model, EMM features outperform emotion-only features and a state-of-the-art language-feature baseline in multiple mood prediction tasks. Further, we find that they contribute to better performance when combined with baseline features. Our results show EMM contains strong and consistent signals of mood change. These informative and interpretable features have great potential to be used in mood tracking applications.

2.2 Dataset

Mood	Scale Ratings
Euthymia	HDRS<=7 and YMRS <=9
Sub-clinical Depression	YMRS<=9 and 8<=HDRS<=16
Clinical Depression	HDRS>=17
Sub-clinical Mania	HDRS<=7 and 10<=YMRS<=20
Clinical Mania	YMRS>=21

Table 2.1: Definition of mood classes.

The Predicting Individual Outcomes for Rapid Intervention (PRIORI) Dataset is a longitudinal collection from individuals with BP [101]. The dataset consists of smartphone conversations made or received by study participants over the course of six to twelve months. Calls were recorded using a secure app installed on participants' smartphones, capturing the participant's side of the conversation. Transcripts of the recordings were obtained with Microsoft Azure speech-to-text transcription service.

There are two types of calls in the dataset: 1) weekly assessment calls, in which clinicians conduct mood assessment interviews with the participant and 2) personal calls, which include all other calls. Prior work has shown that mood signals are more clear in the assessment calls than personal calls [65], and thus in this paper we focus on assessment calls.

Each assessment call contains two structured interviews using the Hamilton Depression Rating Scale (HDRS) [102] and the Young Mania Rating Scale (YMRS) [103]. We define five mood classes based on the HDRS and YMRS scores following previous work [104, 105] (Table 2.1). We use the terms "depressed" or "manic" mood to mean the combination of both the sub-clinical and clinical classes. We apply the same exclusion criterion as in previous work [106], which yielded a final dataset with 1022 calls from 47 subjects, with an average of 82.4 segments (continuous speech separated by silence) per call. The average length of a segment is 6.25 seconds, with a 4.24 seconds std and minimum length of 3 seconds, as in [65]. The distribution of mood labels is 559 euthymia, 389 sub-clinical depression, 129 clinical depression, 57 sub-clinical mania and 22 clinical mania. We measure performance on HDRS score regression, depression severity classification (Euthymic vs. Sub-clinical depression vs. Clinical depression) and 3-way mood classification (Euthymic vs. depression vs. mania). We don't run regression on YMRS scores due to insufficient number of manic samples.

2.2.1 Emotion labels and crowdsourcing

The PRIORI-Emotion Dataset is a subset of PRIORI that contains data from 19 of the subjects who consented to have researchers listening to their phone call recordings [65]. We build emotion recognition models with this subset.

Acoustic Emotion Labels. Previous work has obtained acoustic emotion annotations assessed on a 9-point Likert scale on two dimensions [52]: valence (1 very negative, 9 very positive) and activation (1 very subdued, 9 very energized). The annotators were trained to rate solely based on the acoustic aspects of the recordings, ignoring their lexical content.

Text Emotion Labels. We obtained text emotion annotations with transcripts using the Amazon Mechanical Turk (AMT) crowdsourcing platform. All experiments were approved by University of Michigan Institutional Review Board. We asked annotators to read segment transcripts and select the valence and activation labels that best fit. We again used 9-point Likert scales to evaluate the valence and activation, as in the acoustic annotations. We required AMT workers to be in the US and to have a more than 98% history approval rate. Before the annotation started, workers read a short instruction note explaining valence and activation, and took a brief qualification test to ensure they understood the concepts. Each sample was assigned to three different workers.

We identified unreliable AMT workers by calculating the percentage of "outliers" in their annotation [107]. For each segment, we compared the annotation from the target worker ann_{worker} with the other two annotations ann_1 and ann_2. If $\min(|ann_{worker} - ann_1|, |ann_{worker} - ann_2|) > \max(2, |ann_1 - ann_2|)$, we consider ann_{worker} an outlier. This resulted in the removal of 190 annotations from 17 workers (out of 173) who had a >15% outlier rate, one standard deviation higher than the mean outlier rate. Our annotated subset has 8,033 segments with an average of 2.97 annotations per segment. We use the averaged score across annotators as the final emotion label.

2.3 Methods

2.3.1 Emotional MisMatch (EMM)

2.3.1.1 Quantifying EMM from emotion scores

We define segment-level EMM as the difference between acoustic and text emotion scores. We define a set of statistical descriptors on all segments from a call as the call-level EMM features.

Formally, we consider n segments in a call, $i \in \{0, n-1\}$, each associated with four

emotion scores: activation and valence labels on both modalities. We use a and v for activation and valence, and ac and tt for the acoustic and text modalities, respectively. Then, a_i^{ac} represents the acoustic activation rating on the ith segment, and v_i^{tt}, the textual valence rating on the ith segment. We use m to denote the mismatch between acoustic and textual emotion, i.e., $m_i = (v_i^{ac} - v_i^{tt}, a_i^{ac} - a_i^{tt})$. We then define the EMM feature set with the following components:

Mismatch statistics (25-dim). We calculate 5 statistics (min, median, max, mean, and standard deviation) over the valence mismatch $v^{ac} - v^{tt}$, its absolute value $|v^{ac} - v^{tt}|$, the activation mismatch $a^{ac} - a^{tt}$, its absolute value $|a^{ac} - a^{tt}|$, and the mismatch distance $\|m\|$.

Covariance (4-dim). To measure the extent to which text and acoustic emotion vary together, we calculate covariance on the emotion ratings (i.e., lists of the emotion scores over time) for each call. Specifically, we calculate the covariance between v^{tt} and v^{ac}, a^{tt} and a^{ac}, v^{tt} and a^{ac}, and a^{tt} and v^{ac}.

Interaction with text emotion (20-dim). We hypothesize that the same mismatch values can indicate very different symptoms when the emotion content differs. For example, a person sounding neutral when recalling an exciting experience could convey different information than when sounding sad talking about emotionally neutral content, although they could have the same negativity $v^{ac} - v^{tt}$. To capture this difference, we add 5 statistics on the interaction terms between the emotion and EMM features: $v^{tt} \times (v^{ac} - v^{tt})$, $a^{tt} \times (v^{ac} - v^{tt})$, $v^{tt} \times (a^{ac} - a^{tt})$, and $a^{tt} \times (a^{ac} - a^{tt})$).

2.3.1.2 Automatic emotion recognition

In this section, we describe the emotion recognition methods we use to automatically predict the text and acoustic emotion labels from raw speech.

Text Emotion Recognition (TER). Recently, transformer-based methods have achieved state-of-the-art results. We use Bidirectional Encoder Representations from Transformers (BERT) [108] as our text feature extractor. BERT is a large pretrained language model with outstanding performance in various natural language understanding tasks. We use the pretrained BERT model "bert-base-uncased" from Huggingface[1] as the embedding layer, on which we add two regression layers: one for activation and another for valence.

Acoustic Emotion Recognition (AER). Large transformer-based neural networks (such as Wav2vec 2.0 [109]) have also shown competitive AER performance. However, these learned representations also embed language information [110], which is at odds with our

[1] https://huggingface.co/bert-base-uncased

desire to separately measure text and acoustic emotional content. In this work we use a Convolutional Neural Network (CNN) with Mel Filterbank (MFB) features for AER. For the CNN, we adopt the same architecture that was found to be effective on these data in previous work [65]: We train a CNN that takes 40-dimensional MFB features as input. The CNN itself has with two convolutional layers, each with a kernel size of 4, and output channels of 120 and 360, respectively. The convolutional layers are followed by max pooling over time, a fully connected layer with an output size of 240, and the same two output heads, as the TER model. Each layer, other than the output layer, is followed by a ReLU activation function.

For both TER and AER, we experiment with two model architectures: the base model, which predicts activation and valence independently, and a multi-task model which jointly predicts valence and activation, with the objective function being the average error of valence and activation predictions.

Emotion Recognition Performance. We randomly split our data into nine subject-independent folds. Eight of the folds contain data from two subjects, and the last fold contains three. For each experiment, we train on seven folds, validate on one fold and test on one fold, to ensure that all data from test subjects have not been seen by the model. We measure the average performance over five random seeds and over all folds.

We train both models with an AdamW optimizer with average Root Mean Square Error (RMSE) loss on activation and valence. For TER, we train 10 epochs with a hyperparameter search across learning rates of {5e-5, 1e-5, 5e-6}, with a final selection of 5e-6. For AER, we train 15 epochs with a learning rate of 1e-4 as in [65]. We use Concordance Correlation Coefficient (CCC) as model selection criterion on the validation set.

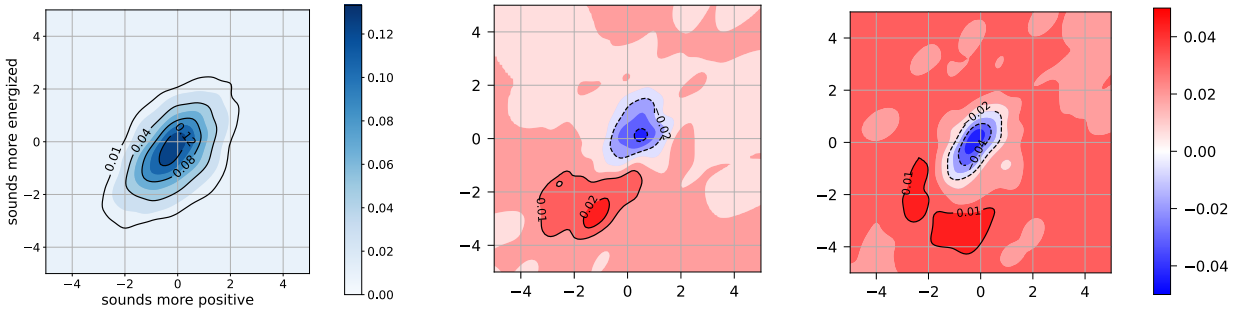
Consistent with previous findings [111], the text modality is better at capturing valence, while acoustics perform better for activation (Table 2.2). We find that multi-task learning generally improves performance and reduces standard deviation, especially for the "weaker" modality. We apply the trained multi-task models to predict the valence/activation scores on segments. We use the average over the five random seeds as predicted emotion scores and use these to extract the EMM features (see Section 2.3.1.2).

2.3.2 Mood prediction models

We evaluate the potential of mismatch features for mood recognition with comparison to two baseline feature sets: 1) language and 2) emotion-only. The **language baseline** follows prior work which achieves state-of-the-art performance on mood severity prediction (HDRS) on the PRIORI dataset [106]. They proposed a set of language features to capture linguistic

Model	Activation		Valence	
	CCC	RMSE	CCC	RMSE
tt-sep	0.301 (.01)	0.242 (.00)	0.620 (.01)	**0.203 (.00)**
tt-multi	**0.328 (.00)**	**0.234 (.00)**	**0.621 (.00)**	**0.203 (.00)**
ac-sep	**0.579 (.01)**	**0.243 (.00)**	0.316 (.01)	0.248 (.00)
ac-multi	0.556 (.01)	**0.243 (.00)**	**0.397 (.02)**	**0.238 (.00)**

Table 2.2: *Emotion recognition performance where tt=textual, ac=acoustic, sep=single-task, and multi=multi-task. CCC=concordance correlation coefficient and RMSE=root mean square error. Best results are shown in bold.*



(a) Euthymic (b) Depressed - euthymic (c) Manic - euthymic

Figure 2.1: (a) Estimated density of emotional mismatch in valence (x-axis, $v_i^{ac} - v_i^{tt}$) and activation (y-axis, $a_i^{ac} - a_i^{tt}$) of euthymic speech. (b, c) Mismatch density differences between symptomatic (b, depressed and c, manic) and euthymic speech.

style, semantic content (Term Frequency Inverse Document Frequency (TF-IDF)), speech intelligibility, along with speaker timing information. We extract these features using their released code[2], which yields 111-dim language features plus tens of thousands of TF-IDF features.

The **emotion-only baseline** first extracts segment-level valence and activation for both acoustics and text and then calculates call-level statistics using the same five statistics described in the mismatch section (20-dim). We augment this with in-call emotion variance (4-dim) and the within-modality covariance (2-dim). All features are normalized by the mean and standard deviation on all euthymic segments in the training set. For both baselines, we perform feature selection with f-regression based correlation ranking. We decide the number of features to keep through cross-validation on the training set, as in [106]. We focus on three tasks: 1) linear regression to predict the HDRS scores, 2) multi-class logistic regression

[2]https://github.com/kmatton/Feature-Extraction/tree/master/text_features

to predict depression severity (euthymia vs. sub-clinical-depression vs. depression), and 3) multi-class logistic regression for 3-way mood prediction (depression vs. euthymia vs. mania, see Table 2.1). We report the mean and standard deviation of performance across subjects using leave-one-subject-out cross-validation.

2.4 Experiments and Results

2.4.1 EMM feature analysis

We first assess whether EMM features differ between mood classes. We run independent t-test between the mismatch distances of euthymic and symptomatic speech segments from the annotated set as a preliminary test of the utility of EMM for mood prediction. We find that speech from individuals in depressed or manic moods show significantly more mismatch than those from euthymic moods: $\mu_{euthymia}(1.04) < \mu_{depress}(1.24)$ ($p < 0.001$), $\mu_{euthymia} < \mu_{mania}(1.35)$ ($p < 0.001$). We then apply kernel density estimation on the 2-d mismatch space (m_i) for each mood state. We visualize the differences between the estimated density of the symptomatic and euthymic states in Figure 2.1. We find that in euthymic speech, the distribution is well centered around the origin, which represents similarity between acoustic and text emotion (Figure 2.1a). Alternatively, depressed speech tends to sound less energetic and more negative than the conveyed content (see the red section in the lower-left area of Figure 2.1b), while manic speech shows more mismatch across the space (note the dispersed red, Figure 2.1c). Those mismatch distributions are consistent with the clinical findings on AB and AE.

2.4.2 Mood prediction

We measure the performance of the regression task with Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE), and we measure classification performance with Unweighted Average Recall (UAR). As shown in Table 2.3, EMM features, when used alone, consistently outperform the baselines in the HDRS regression task: EMM features achieve 0.47 PCC on HDRS, improving over language (0.38), and also over emotion (0.41) baselines. In classification tasks, EMM features bring more improvement when used together with language and/or emotion features.

We conduct an investigation of the feature selection process with the HDRS regression model. We find that eight features are consistently selected across all cross-validation runs and these features are listed, with their coefficients, in Table 2.4. Of these eight features, four are mismatch features: the standard deviation of valence mismatch and three covariance

	Lang.	Emo.	EMM	Emo+EMM	All
HDRS Regression					
PCC	.38 (.31)	.41 (.22)	.47 (.25)ˆ	**.50 (.22)***	0.48 (.27)*
RMSE	5.9 (2.2)	5.6 (2.4)	5.5 (2.2)	5.5 (2.2)	**5.4 (1.9)***
Depression Severity Classification (3-class)					
UAR	.46 (.14)	.45 (.12)	.47 (.16)	.48 (.12)	**.54 (.16)***
Mood Classification (3-class)					
UAR	.43 (.05)	.42 (.07)	.42 (.07)	**.45 (.08)**	.43 (.08)

Table 2.3: Mood prediction results. Lang. - language baseline, Emo. - emotion-only baseline, All - Lang. + Emo. + EMM. Best results of each task are shown in bold. Paired t-tests were performed on subject-wise performances between the language baseline and other feature sets. Significance was shown with symbols following the numbers: ˆ : $0.05<p<0.1$, * : $p<0.05$.

Feature type	Feature name	Coefficient
EMM	std($\lvert v^{ac} - v^{tt} \rvert$)	0.025
	cov(v^{tt}, v^{ac})	-0.064
	cov(v^{tt}, a^{ac})	-1.389
	cov(v^{ac}, v^{ac})	-1.134
Emotion	min(v^{tt})	-0.934
Language	(LIWC) positive emotion	-0.415
	(TFIDF) "yes"	1.200
	(TFIDF) "yeah yes"	0.341

Table 2.4: List of features that are selected across all runs, and the average of their fitted coefficients in HDRS regression.

values. Other features include the count of positive emotion words from LIWC, minimum text valence, and two affirmation words/bigrams from TF-IDF: "yes" and "yeah yes". Consistent with clinical findings on AB, EMM covariance features have negative coefficients. Lower covariance indicates higher mismatch, which is related to higher HDRS scores. Lower text valence score and lower LIWC positive emotion word counts are also indicators of higher depression severity. The last two TF-IDF features are associated with answers in structured clinical interviews. A positive coefficient for these features is capturing the nature of a clinical interaction where individuals answer "yes" to questions on symptoms they may be experiencing. As a result, we suspect that these features may not generalize well to natural speech [112, 106].

2.5 Conclusion

Inspired by research on Affect Blunting and Affect Exaggeration, we propose a set of EMM features quantifying the mismatch between text and acoustic emotion expressions. We show that EMM feature extraction can be fully automated on raw speech recordings using ASR and uni-modal emotion recognition models. Symptomatic speech shows significantly more mismatch than euthymic speech. Our EMM features outperform existing language-only and emotion-only mood baselines, and they can also be used together with existing features to further improve performance. Our results support findings from previous work that mood recognition models benefit from learning emotion variations as an intermediate step and can benefit from multimodal input by observing the mismatch and interactions between them. Looking forward, we will investigate how these features perform in other tasks that relate to mental health. We further plan to explore how models can be guided to attend to meaningful mismatch patterns between modalities, to promote the learning of EMM patterns.

CHAPTER 3

Acoustic-Semantic Factorization for Rich and Robust Speech Representation

3.1 Introduction

Advancements in Large Speech Models through self-supervised pretraining have significantly improved the efficacy and generalizability of learned speech representations [113]. These pretrained speech models can produce general speech representations, which can be applied to a wide array of speech tasks, often approaching or even outperforming small models trained on individual tasks [84]. Recently, representations from prominent models such as Wav2Vec 2.0 [82] and HuBERT [83] have frequently been used as baselines when exploring new speech applications. However, Large Speech Models vary in their performance across tasks, and creating robust, re-usable representations that can generalize across many tasks remains an ongoing challenge [114, 84].

Various self-supervised tasks have been used for learning speech representations, including reconstructing the original audio signal (VQ-VAE [79]), contrastive next-token prediction (CPC [80], Wav2Vec [81]), masked token prediction (Wav2Vec 2.0 [82], HuBERT [83], WavLM [115]), among others. Though these pretraining tasks rely solely on the raw audio signal to construct the prediction target, the contextualized nature of these tasks enables models to capture higher-level semantic information from low-level acoustic patterns. This enables downstream adaptation for semantic tasks like Automatic Speech Recognition (ASR) [82]. To more directly inject semantic information into the model, subsequent finetuning on ASR is often deployed as a second phase of the pretraining (we refer to this phase as "ASR finetuning"), significantly improving the performance on semantic tasks [82, 83].

Speech understanding is inherently complex, as different tasks require different amounts of lexical and para-linguistic information. For instance, speaker recognition depends heavily on local acoustic features, enabling models to identify voices from brief audio snippets, while tasks like ASR necessitate higher-level abstractions focused on the semantic content.

Motivated by our findings from Chapter 2 that acoustic and semantic information encoded in speech can provide different and complimentary information in the mental health domain, we hypothesize that factorizing their representation can broadly benefit speech representation learning, leading to more informative embeddings.

Previous work investigating the localization of acoustic and semantic information within the layers of speech models has shown that the early layers in pretrained models typically capture more general acoustic characteristics, while semantic information emerges in the middle and later layers [86, 116]. The information encoded in the last layer is dominated by the last seen task. For instance, ASR finetuning encourages a better preservation of semantic information in the Wav2Vec 2.0 model, but it causes a significant loss of acoustic information in the final layer: finetuning Wav2Vec 2.0 with 100 hours and 960 hours of ASR data reduces acoustic information preservation by about 50% and 75%, respectively [116]. As a result, downstream applications often have to handle the extra complexity of model and layer selection for optimal performance [117], where the optimal layer is specific to individual model, task, and dataset [86]. It remains unclear whether this acoustic-semantic trade-off is inherent to the tasks themselves or can be mitigated through novel training techniques.

Moreover, semantic and acoustic information is entangled in the representations at each layer, which reduces model robustness when acoustic variations or text domain shift occur. For example, emotional speech often contains acoustic variations (e.g., change of pitch or tone) that can cause extra challenges for ASR systems [85]. Factorization techniques have been helpful in many speech tasks, where learning benefits from disentangling one or more information components. For instance, i-vectors and x-vectors are well-known methods for factorizing speaker representations [118, 119]. More applications include target speaker extraction [120], speech recognition [121], voice conversion [122], and even learning version-invariant music features [123], beyond speech. However, factorization has not been sufficiently explored in learning general speech representations.

To address the challenge of preserving both acoustic and semantic information in speech representations, we propose a factorized (two-branch) multitask finetuning framework. This framework is built on top of the Wav2Vec 2.0 model but can be easily adapted for other speech representation models. Our approach simultaneously performs ASR and audio reconstruction, utilizing a novel design that factorizes semantic and acoustic information into two separate representations in the last layer of the speech model. Specifically, the semantic branch follows the traditional ASR finetuning approach, while the acoustic branch is tailored to capture the rich, non-semantic features of speech by reconstructing the audio signal based on the output of the semantic head, which we call "transcript-aligned audio reconstruction". Leveraging this two-branch architecture, we train each branch to specialize in

different aspects of the speech signal, encouraging the separation and preservation of both types of information. Experiments demonstrate that our method benefits both semantic and acoustic tasks, surpassing both the base and the ASR-finetuned model on three out of four downstream tasks. This indicates a good preservation of both semantic and acoustic information as well as enhanced robustness.

Together, our work highlights two key contributions: 1) we show the feasibility of preserving both semantic and acoustic information in the same layer of a speech model through multi-task training, and 2) we show the promise of semantic-acoustic factorization for more robust general speech representation learning, benefiting various downstream tasks.

3.2 Datasets

3.2.1 Upstream Finetuning

Automatic Speech Recognition (ASR) The Librispeech dataset consists of read English speech from audiobooks, along with rich labels including transcripts and speaker information. The full version contains 960 hours of speech (and has been used in Wav2Vec 2.0 pretraining). We use its 100 hour train-clean split for finetuning our upstream model.

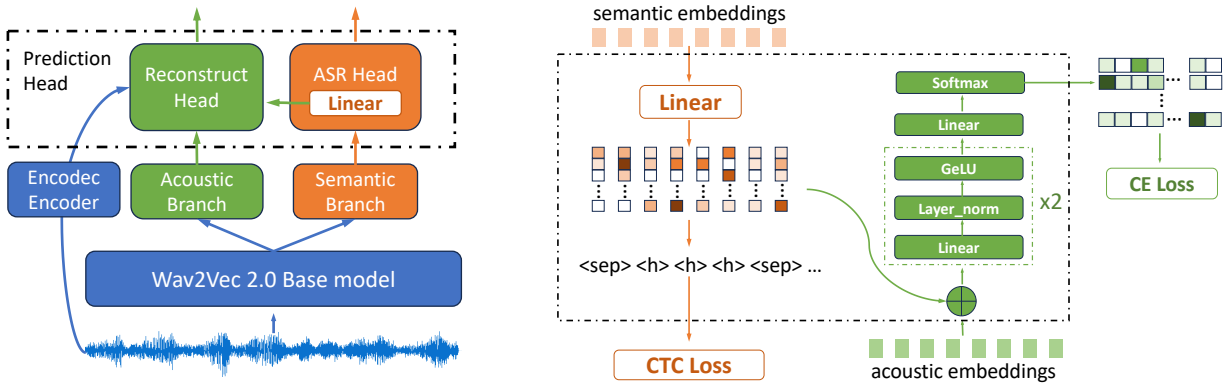
3.2.2 Downstream Evaluation

task	dataset	metric	#train/#test	lr	#steps
SID	Voxceleb1	ACC	149k/5k	0.01	50k
SER	IEMOCAP	CCC	6k/2k	0.01	5k
E-ASR	IEMOCAP	WER	/10k	/	/
KS	Speech Commands	ACC	51k/3k	0.001	50k

Table 3.1: Summary of downstream evaluation tasks.

We evaluate our learned semantic and acoustic representations on several downstream tasks. Details of the downstream datasets and finetuning setups are summarized in Table 3.1.

Speaker Identification (SID) SID aims to classify a given audio into one of many previously-seen speakers, and we use this task to evaluate the preservation of acoustic information in our representations. We use the VoxCeleb1 dataset [124], which contains over 100,000 utterances from 1,251 celebrities extracted from Youtube Videos. We follow the



(a) Full model structure and modules.

(b) Details of the multitask factorization head.

Figure 3.1: Multitask Factorization Model Structure.

train-validation-test split provided in its official release. The performance is measured by Accuracy (ACC).

Speech Emotion Recognition (SER) Emotional Activation is a measure of emotion, ranging from calm to excited. It is highly linked to acoustic signals in speech [125]. Therefore, we use the Activation regression task to measure the preservation of acoustic information. We use the IEMOCAP dataset [55], which contains approximately 12 hours of speech data and emotion activation labels annotated on a 5-point scale. We use Concordance Correlation Coefficient (CCC) as the evaluation metric, which assesses the agreement between predicted and true levels. The dataset contains five recording sessions, and we follow previous work [126] to run cross validation and report CCC mean and standard deviation across sessions.

ASR for Emotion Speech (E-ASR) To evaluate the robustness of semantic representations in acoustically challenging situations, we compare the ASR performance on emotional speech using speech transcripts from IEMOCAP, using Word Error Rate (WER) as the metric. Note that the models are trained to perform ASR, so no extra downstream training is involved.

Keyword Spotting (KS) KS is a semantic-dominant speech task that involves detecting specific keywords or phrases within an audio. We use the Speech Commands dataset [127] v0.01, which has speech utterances containing 31 words such as “bird”, “house”. ACC is used as the evaluation metric.

3.3 Methods

We propose a two-branch multitask framework to learn factorized acoustic and semantic representations of speech, illustrated in Figure 3.1a.

Semantic Branch and ASR Supervision The orange blocks show our semantic branch, which is also the baseline ASR-only finetuning approach as used in the Wav2Vec 2.0 model [82]. The output frame-level representations from the base model are fed into a linear layer with layer normalization to obtain the semantic representations. Then, another linear classification head is applied as a simple decoder to predict the logits. The ASR head is trained with a Connectionist Temporal Classification (CTC) Loss [128]. Given a batch of target sequences y and predicted sequences of logits x,

$$\mathcal{L}_{ASR} = \frac{1}{N}\sum_{i=1}^{N}\left(-\log\sum_{s\in\mathcal{S}(y_i)} p(s \mid x_i)\right) \tag{3.1}$$

$\mathcal{S}(y_i)$ is the set of all possible alignments of y_i with x_i. Note that pretrained Language Model decoders have been applied and can achieve better results on ASR [82]. However, since our goal here is to preserve more information in the semantic embedding rather than the decoder, we use a simple linear layer as the decoder.

Acoustic Branch and Reconstruction Supervision Similarly, we obtain the acoustic representations with a linear layer from the base model output. To supervise the learning of acoustic information, individual tasks may pay more attention to certain aspects like voice characteristics (Speaker Identification) or prosody (Emotion Recognition), but non-semantic audio components themselves are multi-faceted and difficult to define. In order to give a comprehensive representation of the acoustic information, we design the task of transcript-aligned audio reconstruction. As shown in Figure 3.1a, the acoustic embeddings (output of the acoustic branch) are concatenated with the output character logits from the ASR head, and are together used to reconstruct the audio signal through a reconstruction decoder. The idea is that the acoustic representations need to embed all information about the original speech other than the semantic information, which can be learned from the character logits. For the acoustic decoder, we use a stack of two linear layers with normalization and GeLU activation [129].

Reconstruction Target Another design choice of the reconstruction head is what to use as the reconstruct target. One straightforward approach is to reconstruct the original waveform or spectrogram [130]. However, doing this usually involves a much more complicated decoder to handle the sampling rate mismatch between the features and original signals, which adds to the model complexity and the difficulty of learning. Another natu-

ral approach is to continue the masked prediction pretraining task to recover local acoustic features, but that has the drawback that it is model-dependent: for example, Wav2Vec 2.0 is pretrained to reconstruct low-level local acoustic features learned by its CNN feature extractor [82], while HuBERT [83] uses more intermediate codebooks learned by two-phase clustering. If we simply keep the pretrain task as another head, the learned representations may not be consistent in terms of abstraction levels across models. Therefore, we chose Encodec [131], an off-the-shelf model-agnostic audio signal representation as our reconstruction target. Encodec is a neural audio codec model that compresses audio into sequences of discrete class tokens. We use its 6kbps model, which learns to quantize and represent each frame with tokens in eight codebooks, each of size 1024. Since Encodec provides trained encoder and decoder and has good reconstruction performance, being able to predict the Encodec tokens indicates sufficient information for reconstructing the audio.

Therefore, the reconstruction loss on each frame of sample i at frame t is a Cross Entropy Loss between the actual and predicted class, summed across C codebooks.

$$l_{i,t} = -\sum_{c=1}^{C}\sum_{k=1}^{K} y_{i,t,c,k} \log p(x_{i,t,c,k}) \tag{3.2}$$

The reconstruction loss weighs each frame equally and is averaged across sequence lengths T and batch size N:

$$\mathcal{L}_{Rec} = \frac{1}{N \cdot T}\sum_{i=1}^{N}\sum_{t=1}^{T} l_{i,t} \tag{3.3}$$

Together, we train the model with a weighted sum of the ASR loss and Reconstruction loss:

$$\mathcal{L} = \mathcal{L}_{ASR} + \lambda \mathcal{L}_{Rec} \tag{3.4}$$

3.4 Experiments and Results

3.4.1 Experiment Setup

Training. We use the pretrained base Wav2Vec 2.0 checkpoint provided in the transformers library[1]. The acoustic and semantic branch each has hidden size of 768. We finetune the base model with our multitask factorization head with $\lambda = 1$ and an AdamW optimizer

[1]https://huggingface.co/facebook/wav2vec2-base

	#p_train	#p_inference	E-ASR (WER↓)	SER (CCC↑)	SID (ACC↑)	KS (ACC↑)
w2v2-base	/	94.4M	/	0.462±0.010	0.366	0.805
w2v2-100h	90.2M	94.4M	43.29%	0.406±0.009	0.134	**0.874**
ours, mtt.+fct.	114.0M	95.6M	**40.90%**	**0.500±0.011**	**0.387**	0.869
ours, mtt.-only	113.4M	95.0M	41.63%	0.485±0.006	0.335	0.852

Table 3.2: Performance on downstream tasks. mtt.- multitask, fct. - factorization. The best performance across all models are highlighted in bold. Underscore indicates the better performance with or without factorization. The "#p_train" column shows the number of trainable parameters in upstream finetuning, while "#p_inference" indicates the number of parameters for inference, to obtain the representations (excluding task-specific head).

	ASR (WER↓)		Reconstruction (ACC↑)	
	test-clean	test-other	test-clean	test-other
w2v2-100h	6.10%	13.3%	/	/
ours, mtt. + fct.	5.65%	13.90%	0.349	0.286
ours, asr-only	5.81%	14.33%	/	/
ours, rec-only	/	/	0.372	0.319
ours, mtt. only	5.68%	13.84%	0.348	0.285

Table 3.3: Performance on upstream tasks. mtt. - multitask; fct. - factorization. '/' indicates that the model is not trained on and thus unable to perform the task.

(learning rate = 1e-4) for 10,000 steps with a batch size of 128. We select the model with the lowest WER on the Librispeech dev-clean set.

Downstream evaluation. For the E-ASR task, we directly run model inference with the upstream model and decode the output logits from the semantic branch. We use the output from the acoustic branch for the SID and SER, and the output from the semantic branch for the semantic-focused task KS. For all utterance-level tasks (SID, SER, KS), we use a simple pooling layer followed by a linear classification head as the downstream mode, following previous approach [84]. The upstream model is frozen during evaluation. All representations we evaluate have the same dimension of 768. We tune the number of training steps and learning rate for each task to ensure convergence, and the optimal setups we report are detailed in Table 3.1.

3.4.2 Downstream Results

We compare our model to the pretrain-only Wav2Vec2 model (w2v2-base) and the same model finetuned on 100 hour of speech (w2v2-100h) on downstream tasks (E-ASR, SER, SID and KS) in Table 3.2. Note that the base model represents a stronger baseline for acoustic tasks while the 100h model is stronger at semantics [116]. As an ablation study, we also include a multitask-only model, where only one branch is trained with both heads.

First, the performance of baseline models (w2v2-base, w2v2-100h) verifies our assumption that **ASR-only finetuning does cause a loss of acoustic information**. Performance on acoustic tasks (SER and SID) both see a significant drop after finetuning with 100 hour of ASR data. On the other hand, the semantic task KS is improved by the finetuning, showing a trade-off between acoustic and semantic information.

Then, our model outperforms both baseline models on both acoustic tasks (SER ours 0.500 vs. base 0.462, SID ours 0.387 vs. base 0.366). On the downstream semantic task KS,

our model approaches the better performance (ours 0.869, base 0.805, 100h 0.874). Those results indicate that **our model can successfully preserve both acoustic and semantic information** in the last layer of a speech model. What's more, **the factorization design provides an extra bump in the models performance**: it performs better across all four downstream tasks compared to multitask-only finetuning. Notably, on the challenging task of E-ASR, where semantic extraction can be biased by acoustic variations, using factorized representation further reduces WER from 41.63% to 40.90%, indicating enhanced robustness. Further, we note that our approach only adds a small computation overhead (1.3% more parameters) for downstream inference, compared to the baseline models.

3.4.3 Upstream Analysis

Although our goal is to learn a general representation that benefits various downstream tasks, we also analyze the upstream tasks' performance to better understand the model's behavior.

As shown in Table 3.3, **our model can achieve reasonable reconstruction performance without hurting ASR performance**, compared to the model solely finetuned on ASR. On the clean speech test set (the same domain as training data), our model gets 5.65% WER– lower than the semantic-only baseline (5.81%) and the w2v2-100h model (6.10%). It has a slightly worse but comparable performance on the noisier test-other set, despite not being trained with data augmentation (ours 13.90%, wav2vec2 13.30%), outperforming the ASR-only model (14.33%). Comparing the multitask model with models trained solely with ASR or Reconstruction, we find that the two tasks have a small conflict with each other to achieve the best performance, but can both achieve reasonable performance through multitask training. Additionally, factorization doesn't make a significant difference on the upstream performance, compared to the multitask-only model. However, as we show in Section 3.4.2, this extra factorization constraint brings robustness to downstream tasks.

3.5 Conclusion

In this work, we study the challenge of learning factorized acoustic and semantic information in speech representations. We proposed a two-branch multitask finetuning framework that integrates ASR and transcript-aligned audio reconstruction to factorize and preserve these different types of information in the final layer of a pretrained model. Our results demonstrate superior performance and enhanced robustness across multiple downstream tasks. Furthermore, our analysis shows the feasibility of co-training ASR and reconstruction. We

believe our experiments offer valuable insights for building more informative and robust speech models. While our work focuses on factorization at the final layer, earlier separation of the branches may offer further benefits. Future research will explore factorization across different layers.

CHAPTER 4

Impact of Design Choices on Human and LLM Annotations

4.1 Introduction

The training of emotion models has relied heavily on datasets with human annotations. However, obtaining emotion annotations is challenging at all stages. In addition, many design choices can significantly affect the annotation experience and outcomes. For example, the choice of label spaces plays an important role [37]. 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 [38]. 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 [72, 73]. However, contextual influence can introduce inconsistencies, as variations in sample order affect annotators' judgments [39]. 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 [75]. 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 [132]. However, a higher level of agreement does not necessarily indicate more meaningful labels [38]: it can result from reduced diversity in annotations. 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 [133, 38]. 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 [47, 134]. In some cases, multiple iterations are necessary for reliable results [49]. As models grow in size, the cost of data collection increases further due to the need for more data to adequately train them [135].

Recently, the progress in LLMs brings new alternatives. With remarkable proficiency in language modeling across a wide range of scenarios, LLMs show emerging common sense reasoning capability [136]: they can answer a wide range of natural language reasoning questions through zero- or few- shot prompting, matching or even outperforming supervised models [137, 138, 139]. What's more, LLMs display an understanding of human emotion and can respond differently to emotional content [140, 141]. This has inspired research into leveraging LLMs as emotion models to conduct or aid emotion annotation. Yet, they are generally found to be inferior to humans: a few works that developed benchmarks for assessing emotional intelligence consistently indicate a notable gap in complex emotion reasoning between state-of-the-art LLMs and human performance [142, 141, 143]. There have been a few works that evaluate GPT's zero- or few-shot capability of emotion recognition from text or speech input [88, 144, 145, 146]. However, the diversity of emotion label spaces are rarely discussed. Besides, existing works adopt evaluation criteria that rely on automatic metrics against human annotations as the ground truth.

In this work, we comprehensively assess GPT-4's potential to perform emotion annotations in a zero-shot manner. We first measure its emotion recognition performance and find that it performs comparably to established supervised models as baselines, using human annotations as the ground truth. We then reflect on the differences between GPT-perception and human-perception and evaluate how those differences are perceived by a separate set of human evaluators (we refer to them as "evaluators", to differentiate from "annotators" who provided the label annotations in the datasets). Surprisingly, we find that human evaluators consistently prefer the GPT-4 annotations over human annotations. These findings raise important open questions about the suitability of conventional "ground truth" concepts and evaluation practices, especially as models begin to approach human-level performance. Further, we analyze how label formats (categorical vs. dimensional) affect GPT-4's performance, and we explore the feasibility of applying GPT-4 as a quality checker for existing annotations. We demonstrate that GPT-4 can identify potentially low-quality annotations and help with curating a cleaner and more efficient training set.

In summary, our research reveals the great potential of utilizing LLMs for emotion annotation tasks, offers new insights into their capabilities across label formats, and highlights

Dataset	Domain	Label (dim)	Multilabel	Individual label available
ISEAR	self reports	C (7)	No	No
SemEval	tweets	C (11)	Yes	No
GoEmotions	reddits	C (28)	Yes	Yes
Emobank	multi-genre	D (3)	N/A	Yes

Table 4.1: Dataset details. label: C-categorical, D-dimensional.

the challenges involved in their evaluation. We also release the GPT-4 annotations on the entire GoEmotions dataset, along with our code and prompts[1].

4.2 Datasets

We use four publicly available emotion recognition datasets for our analysis, encompassing a variety of label representations and diverse text domains (Table 4.1). Considering the substantial volume of these datasets, we first select a subset of 500 samples from each for GPT-4 annotation and subsequent analysis.

International Survey on Emotion Antecedents and Reactions (ISEAR) [47] is an outcome of a psychological study aiming to understand the antecedents and reactions to seven basic emotions (joy, fear, anger, sadness, disgust, shame, guilt). It consists of 7.6k samples from firsthand emotional reports in text form. We randomly select 500 samples for our analysis.

SemEval 2017 Task 4 (SemEval) [147] is part of the International Workshop on Semantic Evaluation. It consists of Twitter text samples, each annotated with one or more of 11 emotion classes. Since this dataset is very unbalanced, we conduct weighted sampling to select the 500 samples by applying log inverse frequency weighting to the labels, in order to include more emotion labels in our analysis. If a sample carries multiple emotions, the weighting is determined by the rarest label.

GoEmotions [49] is a comprehensive dataset with 58k samples derived from Reddit comments, designed for fine-grained emotion detection. It is characterized by its extensive range of 27 distinct emotion categories, including admiration, remorse, gratitude, etc. Each sample can be assigned one or more emotion labels, as well as an extra "neutral" option. We also apply log inverse frequency weighting in our selection of 500 samples, to address the label imbalance.

Emobank [148] consists of 10k English sentences balancing multiple genres (newspapers,

[1] https://github.com/chailab-umich/GPT-4-Emotion-Annotation

blogs, etc.). The samples are annotated with dimensional emotion labels in the Valence-Arousal-Dominance (VAD) space on a 5-point scale. We focus on the valence score in this study, as it is most commonly included in related literature [149, 146]. Notably, EmoBank distinguishes between the emotional perceptions of writers and readers [150]; we use the reader's annotations, to be consistent with the perspective of GPT-4. We weight the samples by their log deviation from neutral score, to encourage the inclusion of stronger emotional content. I.e., $w_i = log|V_i - 3|$.

4.3 Methods

4.3.1 GPT-4 Prompting

Prompting methods. It is widely acknowledged that LLMs are sensitive to the format and word choices in the prompts [151], making prompts the key factor in the successful application of LLMs. There are two common ways of prompting [152]: cloze prompts, which involve a fill-in-the-blank approach (e.g.,"I feel [X]. I finally got that promotion!"), and prefix prompts, where the model extends a given prompt (e.g., "'I finally got the promotion!' What is the speaker's emotion?"). Given GPT-4's pretraining on generation tasks, our study employs prefix prompts. There have been a lot of work exploring different techniques of prompting that could bring a significant improvement in the models' responses [152, 153, 154]. The efficiency of different prompting techniques is not the focus of this paper. We follow the common effective practices without dedicated prompt engineering.

For each of the three emotion classification datasets, we collect two sets of GPT-4 annotations. In the first set of annotations, we ask GPT-4 to conduct classification annotations by making selections from a pre-determined set of emotion classes. We follow an instruction-based prompting method, which is consistent with the tasks given to human annotators. We try to give GPT-4 similar instructions as those given to humans, based on the descriptions in the GoEmotions paper [49]. Additionally, we set up a persona in the beginning, which has been found to be effective in our preliminary experiments. As an example, the prompt we use for multi-label classification datasets (GoEmotions and SemEval) is shown below.

"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: [Options]. Only select those ones for which you are reasonably confident that they are expressed in the text. If no emotion is clearly expressed, select 'neutral'. Reply with only the list of emotions, separated by comma."

We make minimal modifications as needed for other tasks/datasets and all prompts we use across tasks/datasets can be found in our released code.

We then ask GPT-4 to freely generate descriptors of the expressed emotion, without a given range of options. We compare the generated descriptors with the classification results to understand how the granularity of emotion labels affect GPT-4's performance (Section 4.4.1). For Emobank, we use a similar prompt with the expected response being a integer number from 1 to 5, indicating the perceived valence of the expressed emotion. Using these prompts, we obtain GPT-4 emotion annotations on the 2000 samples selected from four datasets. We additionally obtain classification annotations on all of the GoEmotions dataset using GPT-4 for our model training analysis (Section 4.4.2).

4.3.2 Automatic Evaluation Metrics

Following common approaches in previous work, we evaluate GPT-4's performance on two aspects: 1) agreement with human annotations [149, 155], and 2) potential to improve model performance when GPT-4's annotations are used as training data [145, 146] to train smaller models, in this case implemented by fine-tuning BERT. For classification, we use Unweighted Average Recall (UAR) and Macro-averaged F-1 (Macro-F1) scores as the metrics. UAR measures a model's ability to correctly identify instances of each class with equal importance, while Macro-F1 assesses the balance between precision and recall for all classes. For regression, we use Pearson Correlation Coefficient (PCC) to measure the strength and direction of the linear correlation, and Mean Absolute Error (MAE) to reflect the average error magnitude.

4.3.3 Supervised model: Finetuned BERT

We finetune BERT [108] models on the full training set of each dataset to serve as a supervised baseline. BERT is one of the most commonly used models for text classification tasks [156] and has been used as a benchmark model for the GoEmotions dataset [49]. Besides, its smaller size means it can be run on a single GPU, so we also use it as our base model when comparing the training efficiency of different annotation sources.

We use the same finetuning settings across all experiments: we use the "bert-base-uncased" model implemented in the transformers library, which has 110M parameters. We add a linear layer on top of the base model, and finetune the whole model on a training set with an AdamW optimizer and learning rate = 1e-5. We optimize a Binary Cross Entropy loss for multi-label classification tasks, a Cross Entropy loss for the single-label classification task, and a Mean Squared Error loss for the regression task. We train the model for 10

epochs, and use the model with best performance on a validation set for testing. For the regression task, we find the model not yet converged after 10 epochs, so we train the model for 30 epochs.

4.3.4 Human Evaluation

Human annotations often contain inaccuracies [157], thus metrics based solely on human annotations can be biased. Therefore, we conduct a human evaluation study on samples where GPT-4 and the human evaluators do not agree, aiming to incorporate human perspectives into our evaluation.

We recruited four students from the University of Michigan as our evaluators, aged between 19 to 28 and including two females. They were presented with annotations from two sources (i.e., human vs. GPT-4 classification or GPT-4 classification vs. generation) without identification, and were asked to choose the one which they thought "better and more accurately describes the emotion expressed in the text". Each sample was evaluated by two evaluators, who were given an option to indicate uncertainty on each sample. For the classification tasks (ISEAR, SemEval and GoEmotions), we first found all samples that were annotated with disjoint sets of labels by the two sources. Note that we did not adjudicate annotations if they contained overlapping emotion labels as the differences can be subtle (e.g., "anger" vs. "anger, annoyance"). The annotations were randomized and mixed from three different datasets to reduce the likelihood that evaluators could recognizing patterns associated with a specific source. For the regression task (Emobank), it is harder for evaluators to decide whether a given number is a more or less accurate valence rating for a given sample, especially when the ratings are close. Thus, we adopted a relative evaluation schema [158]. We found pairs in disagreement where one annotation source assigns sample A a significantly ($>$ 1 standard deviation) higher rating than sample B, while the other indicates reversed significance. We asked evaluators to indicate which of the two samples in each pair should have the higher valence. The order of the samples was randomized.

4.4 Experiments and Results

4.4.1 GPT-4 Zero-shot Performance

We first compare GPT-4 and human classification annotations, with a focus on their disagreements. We visualize the disagreements on the ISEAR dataset as a confusion matrix in Figure 4.1 (we select ISEAR because it has the fewest number of classes and is the clearest to show). GPT-4 aligns well with human annotations on most samples, as indicated by the

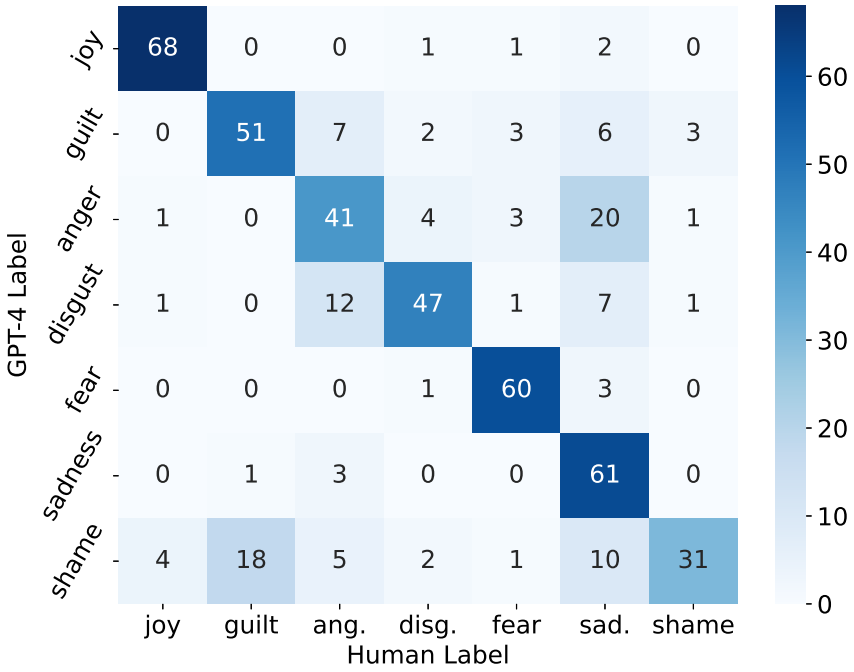
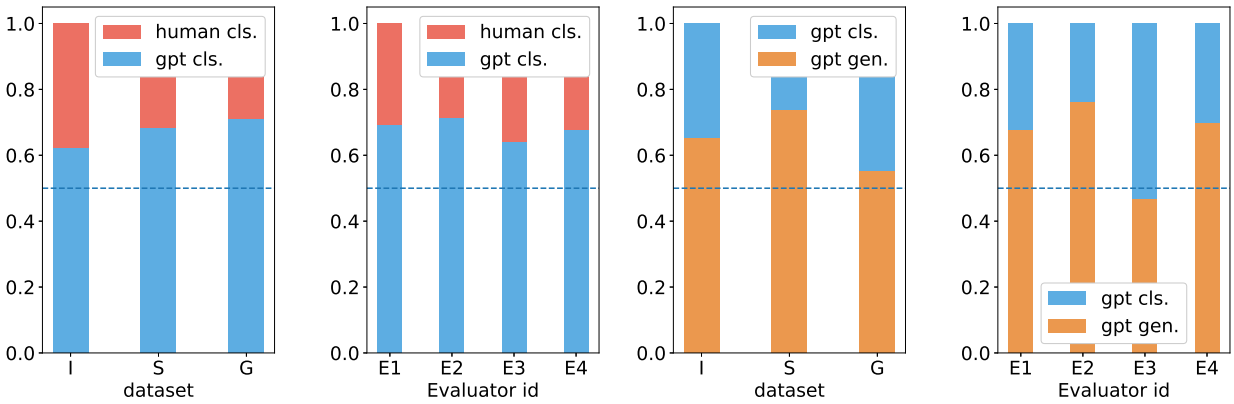


Figure 4.1: Disagreements between human and GPT annotations on ISEAR.



(a) Human vs. GPT-4 classification.

(b) GPT-4 classification vs. generation.

Figure 4.2: Human preference ratio comparing human annotations, GPT-4 classification annotations and GPT-4 generation annotations on emotion classification tasks.

	Macro-F1 ↑		UAR ↑	
	GPT-4	*BERT*	*GPT-4*	*BERT*
ISEAR	**0.739**	0.726	**0.747**	0.727
SemEval	0.511	**0.548**	0.476	**0.495**
GoEmotions	0.375	**0.521**	**0.485**	0.469
	PCC ↑		**MAE ↓**	
	GPT-4	*BERT*	*GPT-4*	*BERT*
Emobank	**0.764**	0.321	0.645	**0.442**

Table 4.2: GPT-4 zero-shot vs. BERT finetuned performance across four dataset. Better performances are in bold.

numbers on the diagonal. It's worth noting that confusions mostly happen among similar emotions, and the confusion between a positive emotion and a negative one is rare. Further, it shows that the confusion between classes is not symmetric, indicating some systematic differences between human and GPT-4 annotations. For example, GPT-4 tends to perceive more shame than guilt (18), but seldom marks human-perceived shame as guilt (3).

We then quantitatively evaluate the zero-shot efficacy of GPT-4, and compare its performance to a BERT model fine-tuned to predict the human evaluations. Our findings are in line with prior work in this space [155] that the two approaches perform comparably, and GPT-4 performs slightly better than BERT on the easier 7-class classification dataset ISEAR, but was more challenged on the multi-label classification datasets SemEval and GoEmotions (Table 4.2).

However, the subsequent human evaluation reveals a different trend and suggests that the automatic metrics may have underestimated GPT-4 performances. As shown in Figure 4.2 (a) with the colored bars representing the ratio of human preference obtained on each annotation source (Human vs. GPT-4 classification), human evaluators prefer labels from GPT-4 on more samples than those from human annotators, consistently across datasets: ISEAR 62.3%, SemEval 68.2%, GoEmotions 71.1%. This trend holds for each individual annotator, ranging from 64.1% to 71.2%.

Further, Figure 4.2 (b) shows that GPT-4 generated emotion descriptions are preferred to GPT-4 classification annotations by human evaluators, indicating that without the predefined classes as a restriction, GPT-4 generates emotion descriptions that were more often preferred by human evaluators. This trend is more significant when the label set is small, like ISEAR (7-classes 65.4%) and SemEval (11 classes, 73.8%), compared to GoEmotions (28 classes, 55.2%). This comparison indicates that it's beneficial to have a larger label space,

which is more likely to encompass the precise emotion labels needed for accurate annotation. The results in Figure 4.2 (a) highlight the proficiency of GPT-4 in navigating a wide range of labels, further demonstrating its utility in complex emotion recognition tasks.

On the valence regression task, GPT-4 significantly outperforms fine-tuned BERT when measured by PCC, but it has a larger MAE (Table 4.2). The large MAE can be explained by the highly centralized distribution of human annotations (standard deviation for human evaluations was 0.54, vs. 1.16 for GPT-4) and the fact that GPT-4 predicts integer-valued numbers while the human evaluations are continuous (e.g., averages of multiple evaluators). However, the large PCC value (0.764) indicates that GPT-4 can identify relative emotional valence. Human evaluation also finds an overall 56% agreement with GPT-4 rather than the original human annotations.

4.4.2 Impact on Model Training

We then investigate whether the labels resulting from GPT-4 classification annotations can be used to train emotion recognition models. We focus on the GoEmotions dataset for this study, using its original train/val/test split.

We compare the performance of a BERT model when it is fine-tuned on the whole training set (N_{train} = 42,278) with human annotations to one trained using the GPT-4 annotations. Additionally, we downsample the training data, retaining only data where the original human evaluations agree with the GPT-4 annotations. We refer to this set as the filtered human set (Human-F, N_{train} = 19,130). Note that this set potentially contains easier samples, compared to both the original human evaluations and GPT-4 annotation labels, because ambiguous samples are more likely to receive different annotations from human and GPT-4, and would thus be filtered out. We test the model on the original human evaluation data (N_{test} = 5,283), the GPT-4 annotations (N_{test} = 5,283), and the Human-F test set (N_{test} = 2,409). We add an extra test set that we refer to as the "adjudicated" test set (N_{test} = 405), which is a subset of the 500 samples used in the human preference evaluation experiment. The set is first populated with samples that have overlapping labels from the original human evaluations and GPT-4 (N = 217), and either the human or the GPT-4 label is selected by random. The remaining samples exhibit disagreement between the two sources. We select the subset of samples where humans exhibited a clear preference for either the original human evaluation or GPT-4 label[2] as the final label (N = 188). The performance on the adjudicated test set is our main metric, because the annotations have been adjudicated and are considered to be more reliable than the raw human or GPT-4 annotations.

[2]Samples where the human evaluators did not agree on the preferred annotation were not included in this sample, 19% of the samples.

Train \ Test	Human	GPT-4	Human-F	Adjudicated
Human (42k)	**0.486**	0.304	0.568	0.392
GPT-4 (42k)	0.343	**0.517**	0.533	**0.524**
Human-F (19k)	0.478	0.367	**0.617**	0.430

Table 4.3: Performance (Macro-F1) of models trained and tested on different combinations of annotations. We show the best performance on each test set (per column) in bold.

In Table 4.3, the models perform most accurately when trained and tested on the same type of annotation. When models are trained on human annotations and tested with GPT-4 annotations (and vice versa) there are notable performance decreases. This indicates that the models learn a systematic difference between human and GPT-4 annotations, which echos our findings in Section 4.4.1. On the adjudicated test set, we find that the model trained on GPT-4 annotations outperformed the model trained on human annotations by a large margin (0.524 vs. 0.392, respectively), again pointing to the systematic differences between the two annotation sources. We find that models trained on the filtered subset of the original human evaluations estimate the labels of the adjudicated data more accurately than models trained on the full set of the original human evaluations (0.430 vs. 0.392, respectively). This result is notable given that the Human-F set is only 45% of the size of the original Human set ($N = 19,130$ vs. $N = 42,278$, respectively).

4.5 Conclusion

In this work, we evaluate GPT-4's emotion recognition capability and find that its zero-shot performance is comparable to supervised models. Our human evaluation study reveals that GPT-4 annotations are preferred to human annotations by our human evaluators, and GPT-4 is good at handling a wide range of options in emotion classification tasks. We also show that models trained on GPT-4 annotations are subsequently better at predicting the labels amongst the adjudicated subset of data. These results highlight the potential of LLMs to be applied in emotion recognition applications.

Several factors may contribute to the observed preference for GPT-4 annotations. First, humans make mistakes, and the increased cognitive load on more complex label spaces could have increased the vulnerability [159]. Additionally, given the inherent subjectivity and ambiguity of emotion annotations [76], different preferences could indicate variations in annotation perspectives or reflect a lack of diversity in the annotation process. Further

exploration is needed to identify the underlying reason. Our findings emphasize the need to reconsider conventional notions of "ground truth" and explore novel evaluation metrics as LLMs approach and surpass human-level performance.

CHAPTER 5

Enhancing Emotion Annotation with LLMs

5.1 Introduction

In the previous chapter, 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 [160]. While these findings provide valuable insights, further verification with larger samples, more annotators, and more comprehensive analysis is needed for a deeper understanding. Additionally, 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 continue our exploration on **how well GPT-4 performs on emotion recognition**. Following our experiments in Chapter 4, we extend our human evaluation of GPT-4 emotion labels to a crowdsourcing setup, to engage more annotator, more diverse perspectives, and larger sample size. Interestingly, although automatic metrics indicate that GPT-4 performs no better than small supervised models trained on human labels, 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. With detailed human feedback, this study provides 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 5.4). 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 [161, 162]. 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.

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.

5.2 Datasets

We use three existing English Emotion Classification datasets. They are a subset of datasets we used in Chapter 4, with a focus on categorical emotion recognition: ISEAR, SemEval and GoEmotinos. A detailed description of those datasets can be found in Section 4.2. For easier reference and more details, we show a summary of the datasets with detailed label information in Table 5.1 .

5.3 GPT-4's Emotion Annotation Capability

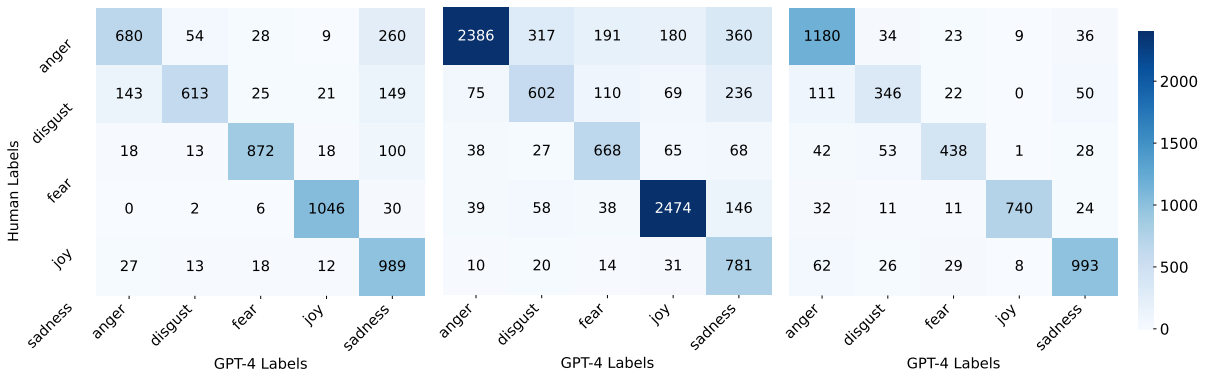
In this section, we evaluate GPT-4's emotion annotation capabilities through a crowdsourcing-based human evaluation study, assessing its alignment with human perceptions. We provide a comprehensive analysis with a focus on the disagreements between human and GPT-4 annotations.

5.3.1 GPT-4 Prompting

We follow the same prompting methods as in Chapter 4. Please refer to Section 4.3.1 for details and examples.

Table 5.1: 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



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

Figure 5.1: Disagreements between Human and GPT-4 Annotations, visualized as confusion matrices.

5.3.2 Human Evaluation Study

Folling the setup in Chapter 4, we conduct a human evaluation study, engaging a separate group of humans to assess how accurately GPT-4 and human annotations reflect the emotions in text. Our sampling method and human evaluation process have some differences from the previous work in Chapter 4, and we detail them below.

5.3.2.1 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.

5.3.2.2 Crowdsourcing Experiments

We design a human evaluation study with the goal of comparing and understanding the disagreement between human and GPT-4 annotations. We present the evaluators with text samples alongside labels from both GPT-4 and human annotators, randomized and without revealing their source. We ask 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. To minimize potential confounds, participants were restricted to a single entry (across all studies in this paper). We implemented the annotation interface with Potato [163], a web-based text annotation tool. We hosted the annotation webpage on an AWS server and recruited participants from Prolific. The participants were selected to be native speakers of American English, at least 18 years old and live in the United States. They 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.

5.3.3 Label Distributions and Agreement Analysis

We first analyze the label distributions and disagreements between human and GPT annotations. We visualized the disagreements with confusion matrices in Figure 5.1. For clarity 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 would have been

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 [88], 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, for future work.

5.3.4 Human Evaluation Results

5.3.4.1 Preference

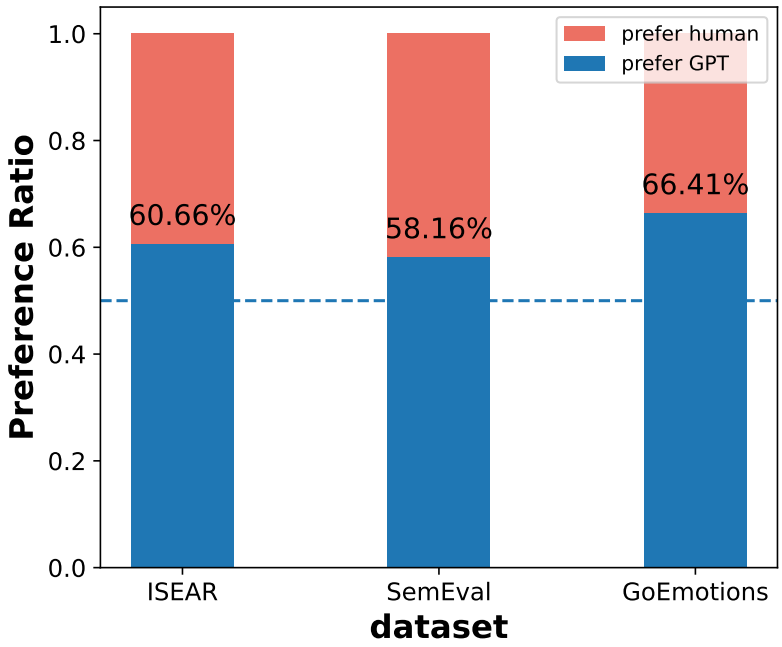


Figure 5.2: Proportion of human evaluators' votes favoring human annotations versus GPT-4 annotations, across datasets.

We first look at the responses to the "preference" question. Figure 5.2 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%) indicate 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 [160]. However, the results of this human evaluation study present a more favorable picture for GPT-4. This conveys an important message that the common method that evaluates LLMs against human labels [87, 88] 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. We hypothesize that as the label space gets larger and more complicated, humans may be more challenged and tend to make more mistakes due to the increased cognitive load [164], while GPT-4 is less affected, especially with proper prompting methods. We discuss this hypothesis in more detail in Section 5.3.4.3, 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).

5.3.4.2 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 5.3. The results reveal a clear and consistent advantage for GPT-4 annotations: human annotators generated a greater number of labels deemed inaccurate (Rating ≤ 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 5.2. 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 5.2, 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

Table 5.2: Percentage of rating scores Human and GPT-4 annotations receive, overall and within each dataset.

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%

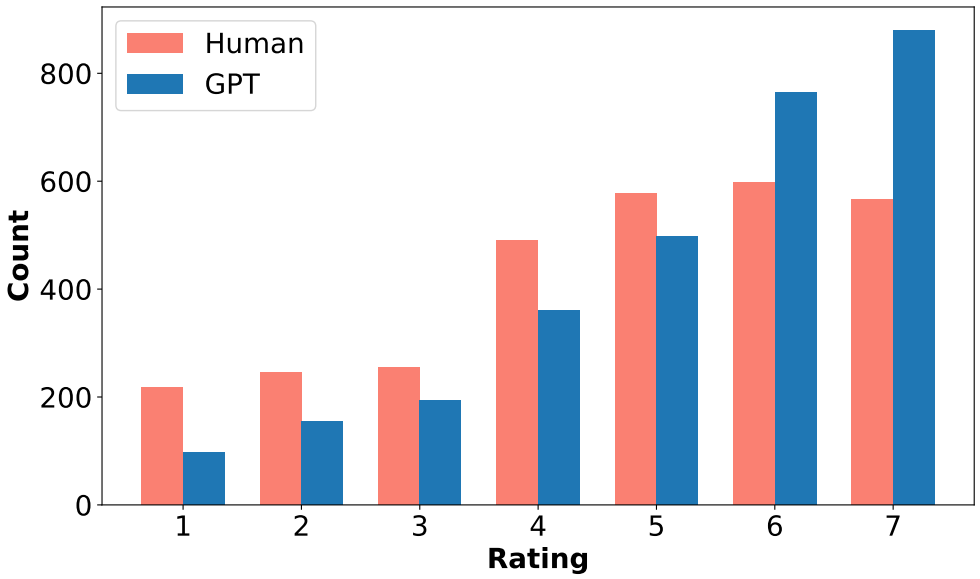


Figure 5.3: Perceived accuracy ratings on a scale of 1 (totally inaccurate) to 7 (totally accurate).

correct label(s), thus facilitating "totally accurate" outcomes. In Section 5.4, we will further explore the influence of label space complexity with a dataset-controlled annotation study.

5.3.4.3 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 in emotion perception: even when selecting between two options, agreement remains relatively low.

5.3.4.4 GPT-4 Weakness Analysis

We also analyze the samples to understand whether and how certain text features may affect GPT-4's emotion classification performance. We extracted the Linguistic Inquiry and Word Count (LIWC) [165] frequencies from each sample. LIWC is 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) [166]. We augmented LIWC with five additional semantic features commonly used for Twitter data [167, 168]: 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 [169] 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 [170] and social cues beyond the text [171], which may explain some of these patterns. However, given the limited size of our data, further investigation is needed to interpret these findings meaningfully.

5.4 Feasibility of GPT-4 Aided Emotion Annotation

Our experiments in Section 5.3 reveal the potential of GPT-4 in emotion recognition. However, we also identified its weaknesses. One notable concern is the instability and unpredictability that often accompany LLMs. They are known to be sensitive to both training data and prompting methods, which can greatly impact their performance [151]. Therefore, using GPT-4 to perform annotations without human oversight can be risky. Furthermore, as shown in Figure 5.1 and Section 5.3.3, there may be systematic differences between GPT-4 and human perspectives. While it is crucial to avoid human errors, we also want annotations to 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 of harnessing the strength of both sources. We focus on the GoEmotions dataset for the coverage of diverse emotions in its samples.

We consider two ways to use GPT-4 in emotion labeling pipelines: using GPT-4 as a pre-annotation label filter to dynamically present a smaller set of classes to human annotators, and, on existing datasets, using GPT-4 as a sample filter to flag potentially low-quality samples. Below we describe each method and our evaluation experiments.

5.4.1 Pre-filtering, label-level

There is a trade-off between the benefit of larger label spaces and increased cognitive load (Section 5.3.3 and 5.3.4.2, also [37]). 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 again prompt GPT-4 in a zero-shot manner with text samples and a list of emotion options. Note that since humans will make selections from GPT-4 filtered labels, the goal of the filter step is to include all possible classes; it is less important to avoid false positives as they can later be identified by human annotators. Therefore, instead of asking it 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 instead of in the general instruction. In a preliminary analysis of a small exploration set, we found that those changes in our prompting methods 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."

5.4.1.1 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.

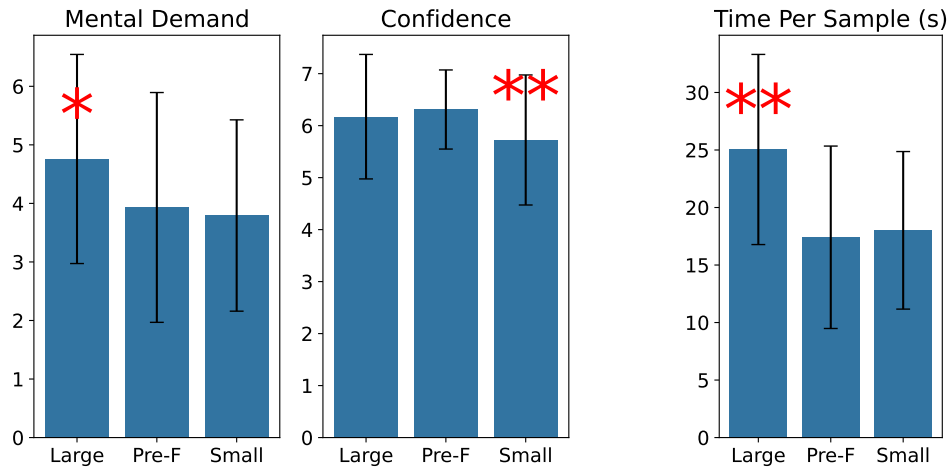
We use a between-subject design: the label sets are fixed for both the Small and Large sets, where participants may gradually gain familiarity with the labels. If we mix those setups in one annotation session, such familiarity cannot be reflected. Therefore, we assign each participant to one of three 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 took the set of 486 GoEmotions samples we used for the evaluation study (Section 5.3.2.1). We further removed samples that GPT-4 didn't output any candidates (N = 4) and defaulted the 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 5.3.2.2. Each participant annotated 50 samples, and each sample got 3 annotations.

5.4.1.2 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. Following common approaches, we measured the cognitive load of the annotators in two aspects: perceived load [172], as a subjective measure, and time to completion [159], as an objective measure. We used the NASA Task Load Index [173] as our cognitive load scale. We removed two questions that were not directly relevant in our task (physical demand and temporal demand), and 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 5.4a). However, a small label space did have a drawback: annotators reported feeling restricted by the options in 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 (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 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 seconds on the Pre-filter set and 18.02 seconds on the Small set, while much longer (25.04 seconds) on the Large set (Figure 5.4b).



(a) Subjective ratings

(b) Time per sample

Figure 5.4: 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. *:$p<0.1$, **:$p<0.05$.

Agreement. We use the Jaccard Index (JI) [174] 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 [49], we obtain aggregated labels from each set by using emotion classes that are selected by at least two annotators out of three. If no emotion class is selected for one sample, it is defaulted to "neutral". We first compared the chosen class labels with the Pre-filtered candidates, and found 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 maintains a low cognitive load and shorter time to completion (which often indicates lower cost) while allowing more descriptive accuracy and confidence of the annotators.

5.4.2 Post-filtering, sample-level

In Section 5.4.1.2, 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).

5.4.2.1 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 whole 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 [108] and DistilBERT [175]. 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 [176], including in the GoEmotions paper [49]. 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

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 5.3: BERT and DistilBERT model performance, finetuned 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. Note that the performance of BERT shown in this table is slightly different from Table 4.3 due to different sampling process.

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.

5.4.2.2 Results

Results are presented in Table 5.3. The filtered set, despite comprising less than 40% of the samples in the full set, consistently led to better model performance across models (both BERT and DistilBERT) and 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 full Human 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.

5.5 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.

5.5.1 Discussion

Our work examines many design choices involved in emotion annotation and investigates how LLMs, specifically GPT-4, perform in this context and where they may offer new possibilities. In the first part of our study (Section 5.3), 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 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.

In the second part, 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.

5.5.2 Limitation and Future Work

Through our studies, we encountered several challenging factors that can influence the collection and evaluation of emotion labels. First, our evaluation study (Section 5.3) reveals that using human labels as ground truth is not always reliable. However, we do not have a good alternative evaluation approach that can account for the inherent subjectivity in emotion. This limitation influenced our quantitative model evaluation in Section 5.4.2, where we

relied on a human-annotated set created through multi-annotator discussions. Although we believe that this annotation method greatly reduces the possibility of mistakes, we also note the ambiguity of many samples, and thus reasonable alternative interpretations are possible. Developing new evaluation metrics that address these complexities would be a valuable, though challenging, direction for future research.

We also limited our discussion to classification tasks. Our previous work conducted preliminary experiments that included dimensional labels with the Emobank dataset [160] and found that the scale of GPT-4 output did not quite align with human labels, raising more questions for evaluation. LLMs are generally good at language-based interactions, and language anchors are helpful for them to understand dimension scales [177]. As dimensional label spaces gain more popularity [178], future research could explore ways to better leverage LLMs in dimensional emotion annotation.

Additionally, our pre- and post-filtering methods (Section 5.4) 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 [179, 180, 181]. 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 [151, 152, 182, 154], and we encourage new explorations with our prompts and code publicly available [1].

[1] https://github.com/chailab-umich/GPT-4-Emotion-Annotation

CHAPTER 6

Incorporating Annotators' Information for Robust Emotion Recognition

6.1 Introduction

In this chapter, we explore another source of inconsistency in human labels: the inherent subjectivity of emotion perception. Emotions are not universally expressed or interpreted, which leads to variations in annotations and challenging the notion of a single, objective ground truth [183]. Previous research has identified various factors that contribute to such variation, including characteristics of both the speaker and the perceiver, as well as contextual influences. Speaker characteristics such as gender [184], age [185, 35], culture [186], as well as perceivers' demographics like gender [187, 188], age [189, 190], personality traits [191] and affective states [192], can all influence how certain expressions are interpreted. These sources of variations pose great challenges in training SER models. Models can be confused by inconsistencies in the labels [193] or biased towards a dominant group in the data collection process [194, 195].

Several studies have recognized the inherent uncertainty and subjectivity in emotion annotations and have proposed modeling approaches to address these challenges. Common approaches include representing labels as soft distributions, capturing the proportion of annotators choosing each emotion rather than relying on one-hot encodings [196, 197, 198, 199], or adding annotation uncertainty or agreement prediction as additional prediction targets [200, 201]. Some studies have attempted to personalize SER by tailoring models to individual annotators. For example, separate models or modules have been trained for each annotator [202, 203], or annotator indicators have been included as inputs to a shared model [202, 204]. Additionally, speaker embeddings have been introduced to make annotator-informed predictions [205]. Although these methods acknowledge the existence of subjectivity and capture uncertainty and variation in human perception, they often require access to historical annotations from each annotator, which are not always available. Moreover, they

do not seek to quantify how annotator characteristics influence the resulting annotations. In other words, they implicitly treat the label distribution as a property of the input sample, not accounting for the characteristics of the annotators who generated those labels. As a result, any biases in the annotator pool on each sample remain unmodeled.

To the best of our knowledge, few studies have explored SER with a focus on annotator characteristics. Chien and Lee [194] proposed training separate emotion spaces for male and female perspectives. However, this approach does not directly scale to more attributes or interactions among attibutes, as the number of required modules would grow rapidly. Beyond SER, Ding et al. [195] demonstrated that annotators' gender and age can lead to shifts in sentiment label distributions across modalities, ultimately impacting model performance. Similarly, Yao et al. [206] found that incorporating personality features improved the performance of personalized abusive language detection. Despite these insights, the integration of annotator characteristics (beyond gender) into SER models remains underexplored.

In this work, we study subjectivity in emotion perception. We examine how demographics and personality influence annotators' interpretation of emotion expressions in speech, and how such insights can be used to enhance model performance. Most public datasets do not have detailed annotator information, so we augment a subset of the widely used MSP-Podcast dataset ([46], details in Section 6.2.1) with 20 annotations per sample, along with annotator demographics (age, gender, race, education) and personality traits (measured by the Big Five dimensions [207]). Our statistical analyses show that these annotator-specific factors have significant impact on their labels (Section 6.3). Furthermore, we propose a simple adapter-tuning method to finetune a trained SER model with a small annotation-aware set, as a baseline for incorporating subjectivity into SER (Section 6.4). Our results show that even with a small number of examples, conditioning on annotator characteristics can significantly improve performance for seen annotators, highlighting the potential of personalization in SER. However, this improvement does not extend to unseen annotators. We discuss potential reasons, highlighting the difficulty and risk of generalizing subjectivity modeling beyond the training set.

6.2 Datasets

6.2.1 MSP-Podcast

We use the MSP-Podcast dataset release 1.11 (May 31, 2023) [46], a large, publicly available SER corpus, for our experiments. It features spontaneous, naturalistic speech segments from podcast recordings across diverse speakers and contexts. Each segment is annotated

with emotional attributes, including both categorical labels (e.g., Happiness, Anger) and dimensional labels (Valence, Arousal and Dominance). In this work, we focus on the most commonly used dimensional labels of valenve (1-Very negative to 7-Very positive) and arousal (1-Very calm to 7-Very active). The dataset provides mean-aggregated labels, i.e., ratings averaged across annotators. For consistency, we adopt the data splits provided in this dataset. This includes 84,030 segments from 1,409 speakers for training, 19,815 segments from 454 speakers for validation, and 30,647 segments from 237 speakers in the most commonly used Test Set 1.

6.2.2 MSP-Podcast-Percept: Crowdsourcing an annototor-informed subset

MSP-Podcast does not contain detailed annotator information. We introduce the Personality and Demographic Emotion Perception subset (MSP-Podcast-Percept), a small, balanced subset of MSP-Podcast augmented with emotion annotations as well as demographic and personality information of the annotators, to enable the exploration of how individual differences shape emotion perception.

Sampling. We applied a sampling process to select a small yet representative subset of 500 samples for annotation. First, to ensure a balanced distribution of emotion labels across annotators, we divided the sampling goal into 10 buckets, each containing 50 samples. Each bucket was balanced in the valence-activation (VA) space by dividing the two-dimensional VA space into a 3×3 grid of squares. We sampled five samples from each square and selected an additional five samples randomly, without repetition across buckets. We sampled 60%, 20%, and 20% (6, 2, 2 buckets, respectively) from the train, validation, and test splits.

Emotion Annotations. We collected dimensional emotion annotation, similar to the approach in MSP-Podcast [46]. At the beginning of the session, we presented an instruction page explaining the task and the concept of valence and activation (arousal), using the self-assessment manikins for visual illustration of the scales [53]. In the main annotation task, we presented the annotators one audio segment at a time, with the option to replay as many times as they needed. We asked them to rate valence and activation, respectively, on the same 7-point scales as in MSP-Podcast (Section 6.2.1).

Demographics. We collected demographic information at the beginning of the survey. Following prior work that identified demographic factors affecting annotations in related tasks [195, 208], we included questions about gender, age group, race, and level of education. To be consistent with common practices, we adopted the demographic categories used in

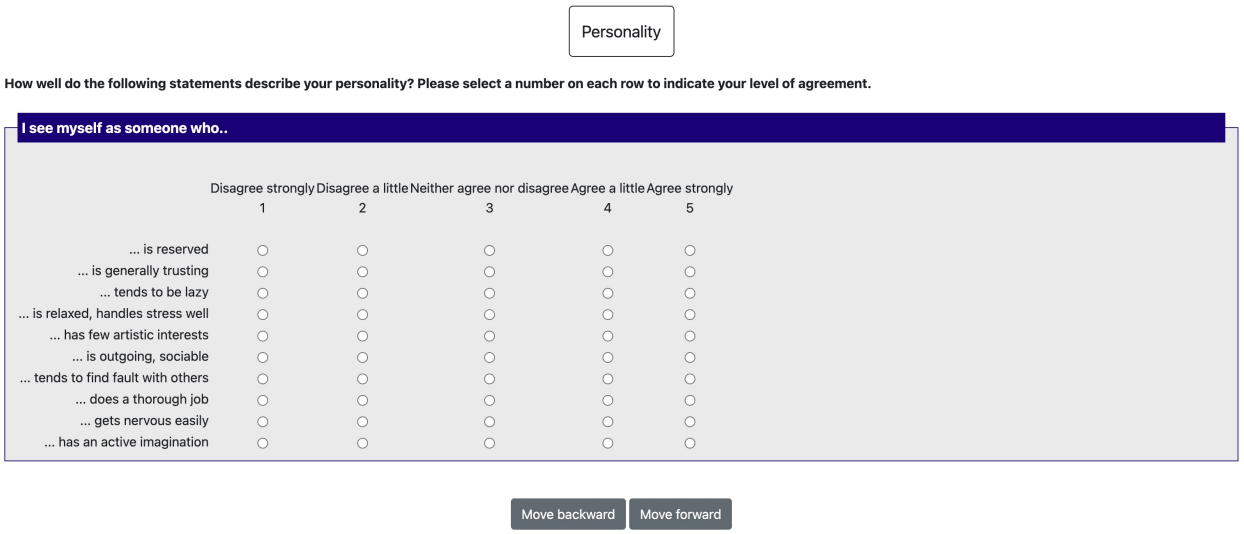


Figure 6.1: A screenshot of the crowdsourcing interface page we use to collect annotators' personality information.

Prolific's screener questions[1].

Personality. We followed the most widely adopted Big Five framework to measure personality, which quantifies personality along five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness [207]. To minimize the burden on participants, we used a short version of the Big Five Inventory, BFI-10, which is significantly shorter than the full 44-item BFI-44, but has shown strong correlations with the full inventory and good reliability [209]. Participants rated 10 statements describing their personality on a 5-point scale ranging from 1 (Disagree strongly) to 5 (Agree strongly). These ratings were then converted into scores for each of the five dimensions. For example, Extraversion was measured using two statements: "(1) I see myself as someone who is reserved" and "(6) I see myself as someone who is outgoing, sociable," with the Extraversion score computed as the rating for (6) plus the reverse of the rating for (1) (i.e., 5 - (1)). We presented the personality scale at the end of each session. We present the scale we use within the corwdsourcing interface in Appendix A.

Survey Flow and Responses. Each annotation session consisted of five sequential parts: Consent, Demographics, Instructions, Annotation, and Personality. We recruited 200 annotators and assigned 50 samples to each of them, with the goal of collecting 20 annotations per sample across a total of 500 samples. We used the quota sampling feature of Prolific to ensure gender balance, obtaining 50% of responses from female participants.

We encountered minor data loss: two annotators and their 100 annotations were excluded

[1] https://researcher-help.prolific.com/en/article/412c0a

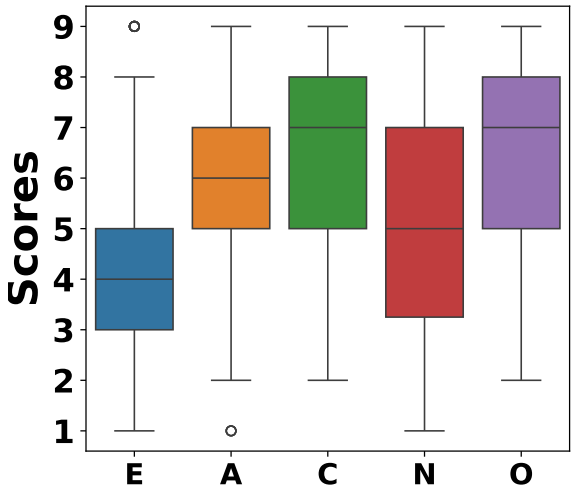


Figure 6.2: Distribution of personality of annotators, along five dimensions (E-Extraversion, A-Agreeableness, C-Conscientiousness, N-Neuroticism, and O-Openness). The boxes show the mean (middle line), first and third quantiles (upper and lower boundaries). The whiskers extend to data points within 1.5 times the interquartile range (the range between the 25th and 75th percentiles). Points beyond indicate outliers.

because of missing demographic information, and five annotators did not complete the full 50 samples, resulting in an additional loss of 25 annotations. The final MSP-Podcast-Percept set consists of 9,875 valid annotations from 198 annotators. Each sample got at least 18 annotations. The median time to complete the task was 16 minutes and 13 seconds, with a median payment rate of $11.1 per hour.

Annotator Distributions. We achieved a relatively balanced gender distribution among annotators, with 50.5% identifying as male, 46.5% as female, and 3% as non-binary. The majority of annotators identified as White (62.7%), followed by 16.7% Black or African American, 9.1% Asian, 2.0% American Indian or Alaska Native, and 9.6% identifying as multi-racial. The age distribution was more evenly spread, with 20.7% aged 18-24 years, 35.4% aged 25-34, 21.7% aged 35-44, 19.2% aged 45-64, and 3.0% aged 65 and above. Regarding education, 38.9% of annotators held an undergraduate degree, 22.2% a high school diploma, 18.2% a graduate degree, 15.2% a technical or community college qualification, and 5.6% identified as having other educational backgrounds. We also had a good coverage of personality along all dimensions, as shown in Figure 6.2.

Label Distributions and Agreement Check. We compare the distributions of the original and new labels within the MSP-Podcast-Percept subset. We show a scatter plot of the original and new annotations in the valence–activation space Figure 6.3. Our new annotations generally show good coverage of the emotion space. We also validate the newly

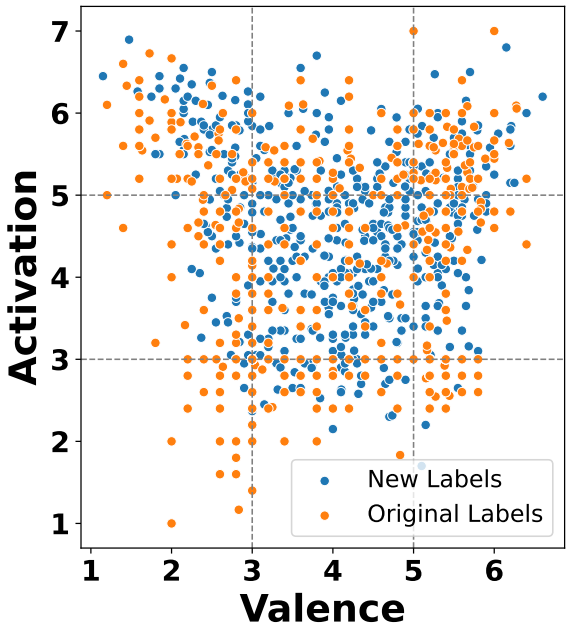


Figure 6.3: Distribution of mean-aggregated labels of the 500 samples, from the original Podcast dataset versus our new annotations.

collected annotations by calculating Krippendorff's Alpha using the ordinal distance metric [210]. We calculate Alpha values of 0.401 for activation and 0.481 for valence, comparable to those in the MSP-Podcast corpus (0.426 for activation and 0.459 and for valence) [46]. Alpha values in emotion annotations often interpret 0.4 to 0.6 as fair to moderate agreement [211]. Given that the annotations were collected approximately five years apart with different crowdsourcing platforms, different groups of participants, and non-identical instructions, we believe the observed agreement level falls within a reasonable range.

6.3 Statistical Analysis

In this section, we present statistical analyses to measure the relationship between annotator's demographics and personality on annotation outcomes.

6.3.1 Influence of Demographics

We first examined whether annotations differed significantly across demographic groups. To ensure meaningful comparisons, we focused on groups with at least 20 annotators in our dataset.

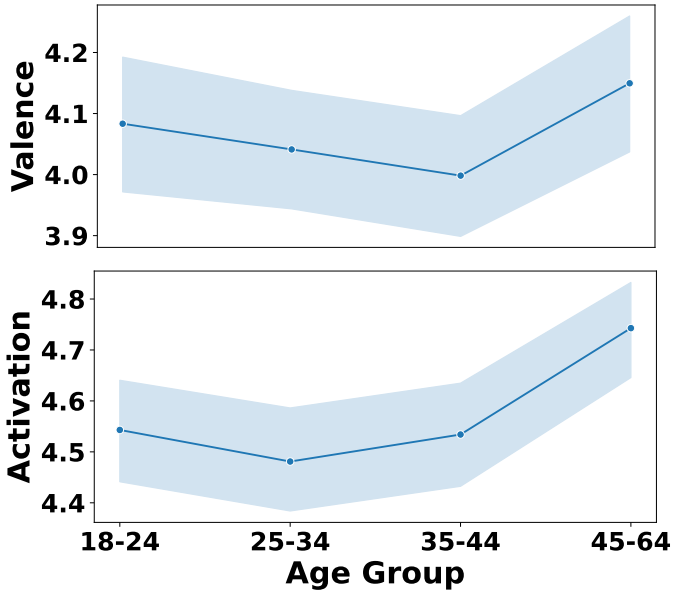


Figure 6.4: Mean valence and activation ratings across age groups. Blue shadows: 95% confidence intervals.

For gender, we focused on the female and male groups, as there were only three annotators identified as non-binary in the responses. We calculated the average valence and activation ratings for each sample, separately for female and male annotators. To assess the significance of the differences, we conducted paired t-tests on these sample-level averages between the two groups. Overall, female annotators tended to perceive slightly lower valence (F 4.03 ± 1.14, M 4.10 ± 1.07, $p < 0.005$). No significant difference was observed in activation (F 4.57 ± 1.12, M 4.59 ± 0.99, $p = 0.562$).

For education, we followed previous work [208] to combine smaller categories into four groups: College degree, Graduate degree, High School Degree, and others. We compared the mean valence and activation ratings among the groups. First, one-way ANOVA [212] analysis indicated significant differences in the average ratings among the four groups, for both valence and activation (both $p < 0.05$). A closer pairwise comparison with the dependent t-test showed that annotators with a Graduate degree perceived significantly higher valence and higher activation than those in any other education group ($p < 0.05$ for all comparisons), while differences among the remaining groups were not statistically significant.

For race, we excluded the American Indian or Alaska Native group from our analysis due to the small sample size ($n = 4$). ANOVA revealed significant differences among racial groups ($p < 0.001$ for both valence and activation). Given the small sample sizes of non-white groups in our data, we do not report results from pairwise t-tests between racial groups to avoid unreliable or misleading conclusions. Future work can explore this with larger, more

Variable	Coef.	Std.E.	$P > \|z\|$
Intercept	4.093	0.058	0.000
C(gender), base Male			
Female	-0.049	0.025	0.055
C(age), base 18-24 years			
25 to 34 years	-0.109	0.034	0.001
35 to 44 years	-0.189	0.040	0.000
45 to 64 years	-0.045	0.041	0.270
C(education), base College			
Graduate degree	0.235	0.034	0.000
High School degree	0.006	0.031	0.855
Others	-0.003	0.059	0.963
C(race), base White			
non-White	0.123	0.026	0.000
Personality Traits			
Extraversion	0.026	0.014	0.056
Agreeableness	0.030	0.013	0.021
Conscientiousness	0.089	0.013	0.000
Neuroticism	0.004	0.015	0.771
Openness	-0.017	0.013	0.165

Table 6.1: Linear Mixed-Effects Regression Results with Valence as the dependent variable.

representative samples.

For age, we excluded > 65 years old group ($n = 6$). ANOVA showed a more significant age-wise different for activation ($p < 0.001$) than valence ($p = 0.033$). As shown in Figure 6.4, valence ratings declined from the 18–24 group to the 35–44 group, followed by a significant increase in the 45–64 group in the 45–64 group ($p < 0.001$). This trend aligns with previous findings suggesting that older adults tend to report more positive feelings [35, 195]. For activation, ratings were significantly higher in the 45–64 group compared to all others ($p < 0.001$).

6.3.2 Influence of Personality

We ran Linear Mixed-Effects Regression (LMER) to investigate the influence of personality traits on annotation outcomes. Annotators' five personality traits (z-normalized across all annotators) were included as fixed effects, with their valence and activation ratings being the dependent variables. We included sample_id as a random effect to account for sample-wise variability. To control for potential confounding factors, we also included all demographic features as covariates in the regression. This ensures that any observed associations between personality traits and annotation outcomes are not driven by correlated demographic

Variable	Coef.	Std.E.	$P > \|z\|$
Intercept	4.619	0.059	0.000
C(gender), base Male			
Female	0.008	0.028	0.781
C(age), base 18–24 years			
25 to 34 years	-0.112	0.038	0.004
35 to 44 years	-0.154	0.044	0.001
45 to 64 years	0.113	0.045	0.013
C(education), base College			
Graduate degree	-0.052	0.038	0.170
High School degree	-0.218	0.035	0.000
Others	-0.173	0.066	0.008
C(race), base White			
non-White	0.124	0.029	0.000
Personality Traits			
Extraversion	-0.042	0.015	0.006
Agreeableness	0.023	0.015	0.123
Conscientiousness	-0.000	0.015	0.996
Neuroticism	0.032	0.017	0.058
Openness	-0.008	0.014	0.577

Table 6.2: Linear Mixed-Effects Regression Results with Activation as the dependent variable.

variables. For race, due to the limited samples in each non-White groups, we adopt a coarse-grained White/Non-white group setup, following previous work [195].

The results are summarized in Table 6.1 and Table 6.2. We found that annotators with higher Agreeableness and Conscientiousness tended to give higher valence ratings ($b_A = 0.030$, $p < 0.05$; $b_C = 0.089$, $p < 0.001$). Similarly, lower Extraversion is associated with higher activation ratings ($b_A = 0.027$, $p < 0.05$; $b_N = 0.033$, $p < 0.05$). However, the effect sizes of the personality features are quite small, leaving questions about whether they can provide sufficient signal to meaningfully influence model performance when used in model training. Results from our LMER models on demographic features are largely in line with the patterns observed in our t-test analyses (Section 6.3.1). For example, the LMER results also show that male annotators tend to give higher valence ratings than female annotators. This consistency adds additional support to our findings.

6.4 Demographic and Personality-Informed SER Adaptation

Building on the insights from Section 6.3, we hypothesize that incorporating annotator information into SER models can improve their ability to predict individual emotion labels. To test this, we implement a lightweight annotator-aware adapter on top of a pretrained SER model, conditioning the predictions on annotator characteristics. We evaluate its effectiveness on both seen and unseen annotators.

6.4.1 Base Model

We follow common approaches for training SER models [26]. We use a pretrained Wav2Vec 2.0 model [109] as the encoder and train a multitask regression head, which takes the mean-pooled representation from the encoder and jointly predicts valence and activation. The regression head consists of a simple dropout layer followed by a linear layer, trained using Concordance Correlation Coefficient (CCC) loss [26, 213]. We finetune the model with the mean-aggregated labels on the full MSP-Podcast training set and select the best checkpoint based on performance on the development set. This model serves as the base model for all our finetuning experiments.

We initialize our model from the Hugging Face wav2vec2-base-960h checkpoint. It has been finetuned on 960 hours of ASR data, which enables it to better capture linguistic content in speech. After finetuning, our base model achieves CCC scores of 0.657 for valence and 0.597 for activation on the test set, comparable to the performance reported in previous work [26].

6.4.2 Adapter Finetuning

We introduce a lightweight adapter module to embed and incorporate annotator information into a pretrained SER model. As illustrated in Figure 6.5, annotator information is first encoded with linear layers for continuous personality traits or embedding layers for categorical demographic features. The resulting embeddings are concatenated and projected into a single annotator feature vector, which is then concatenated with the audio representation from the frozen Wav2Vec2 encoder. This combined vector is fed into the regression head. The adapter is randomly initialized, while all other weights are initialized from the base model. During training, only the adapter and regression head are updated.

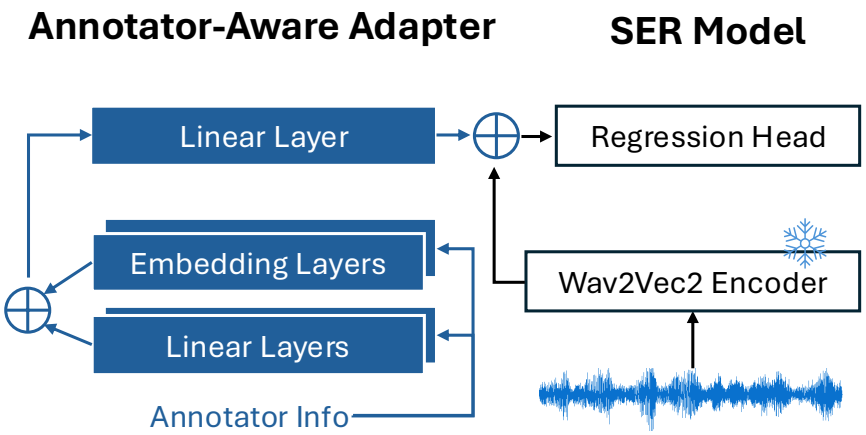


Figure 6.5: Adapter tuning model structure.

6.4.3 Experimental Setups

We evaluate our approach in two scenarios: (1) on seen annotators, to assess personalization capability, and (2) on unseen annotators, to measure generalization capability.

Seen Annotators. In our crowdsourcing experiments, we ensured that each annotator annotated 30 samples from the training split of MSP-Podcast dataset, 10 from the development set and 10 from the test set. The three splits share the same set of annotators but different audio samples, enabling evaluation of models' personalization capability.

Unseen Annotators. To evaluate whether annotator information can help the model generalize to unseen annotators, we assign each annotator to one of five folds and run cross-validation experiments.

Metrics. We use Concordance Correlation Coefficient (CCC) as the evaluation metrics. We report the CCC mean and standard deviation (std) on valence and activation respectively, as well as their average. We run experiments with five random seeds for the cross-validation experiments, with a total of 25 runs (5 folds). We run experiments over 25 random seeds for seen annotators.

Baselines and Model Selection. We experiment with demographic and personality information as additional input, as well as annotator id as a comparison. We also include a baseline model where the regression head is further finetuned on the emotion ratings of each individual, without using any annotator information. All models converged during training, and we report results on test sets using the best-performing model on the validation sets.

Method	**CCC_avg**	**CCC_V**	**CCC_A**	**#params_t (k)**
	Seen Annotators (25 random seeds)			
base model: mean-finetuned	0.546 (0.005)**	0.495 (0.008)**	0.596 (0.006)*	90172.8
baseline: adapter-tuned, no extra info	0.558 (0.001)	0.489 (0.002)	0.628 (0.002)	1.5
adapter_annotator_id	0.567 (0.003)**	0.496 (0.003)**	0.638 (0.003)**	14.2
adapter_demographic	**0.575 (0.002)****	**0.504 (0.002)****	**0.645 (0.002)****	12.4
adapter_personality	0.558 (0.001)	0.491 (0.002)	0.626 (0.002)	4.5
	Unseen Annotators (5 random seeds × 5-fold cross validation)			
base model: mean-finetuned	0.595 (0.032)**	0.625 (0.036)*	0.565 (0.033)**	90172.8
baseline: adapter-tuned, no extra info	0.614 (0.030)	0.629 (0.034)	0.600 (0.030)	1.5
adapter_demographic	**0.616 (0.027)**	**0.630 (0.032)**	**0.603 (0.024)**	12.4
adapter_personality	0.614 (0.030)	0.628 (0.035)	0.600 (0.031)	4.5

Table 6.3: Model performance for seen and unseen annotators using different input types.We report mean(std) across runs. #params_t refers to the number of trainable parameters. Significance between the baseline model (underlined) and other models is measured by dependent t-tests across 25 runs: **: $p < 0.001$, *: $p < 0.05$.

6.4.4 SER Results

Results are presented in Table 6.3. We first observe that the baseline model, trained on individual annotations without any annotator information, outperforms the base model trained on mean-aggregated labels. This performance gain holds despite the model receiving noisier supervision (i.e., the same audio input paired with varying labels). We tuned the base model's learning rate and trained it for additional epochs to ensure this was not due to differences in training effort. These results suggest that learning from individual-level variation, even if noisy, can be more informative than relying solely on mean-aggregated labels.

Next, we compare models that incorporate annotator information via adapters to the baseline. We find that adding demographic information yields statistically significant improvement in the seen-speaker setting (paired t-test, $p < 0.001$). Incorporating personality traits does not lead to noticeable gains over the baseline—consistent with our statistical analyses, which showed relatively weak effects of personality on emotion ratings. Interestingly, in the seen-speaker setting, models informed by demographic attributes significantly outperform those using annotator ID ($p < 0.001$ for both valence and Activcation). This suggests that the model benefits from shared structure across annotators with similar characteristics, rather than learning isolated representations for each individual.

However, the improvements from incorporating annotator information do not generalize to unseen annotators. When applying the same model architecture used in the seen-speaker setting, we observed substantial overfitting: models achieved higher training CCC but degraded test performance. To counter this, we introduced strong dropout (p = 0.5) on the annotator representation layer, which mitigated overfitting performed similarly to the baseline model (shown in Table 6.3, $p > 0.1$). These findings suggest that while annotator-informed modeling can support personalization, it does not yet yield robust generalization. Despite using lightweight adapters with relatively few trainable parameters, the model appears to memorize individual annotators' demographic information and associated labels rather than learning transferable patterns. This can lead to unintended consequences, for example, the model may over-rely on annotator-specific signals, limiting generalization and potentially entangling predictions with demographic characteristics in undesirable ways.

We examine the degree to which annotators are identifiable based on their demographic information to better understand the model's tendency to overfit in the generalization setting. Demographic attributes can possibly serve as near-unique identifiers given the large number of possible combinations and the limited number of annotators in our dataset. Among 200 annotators, we found 94 distinct demographic combinations, with 50 annotators having a combination that is unique within the dataset. This suggests that demographic embeddings may act as proxies for annotator IDs, enabling effective personalization while also encoding

some similarities. However, this property poses a challenge for generalization: unseen annotators have no meaningful representation in this embedding space. As a result, the learned demographic representations may be too specific to support generalizable emotion prediction for new annotators.

We discuss other possible reasons for this limited generalization in Section 6.5, but we emphasize that generalization can pose a significant risk of overfitting. We present these results to strongly caution against these risks.

6.5 Conclusion

This work explores modeling annotator subjectivity in speech emotion recognition. Using a dataset of nearly 10k emotion annotations with associated demographic and personality information, we found, through statistical analyses, that emotion perception is significantly influenced by annotator characteristics. Our modeling experiments demonstrate that incorporating demographic features via lightweight adapters improves performance in predicting individual-level labels for seen annotators. However, this approach does not improve generalization to unseen annotators and poses risks of overfitting and overly relying on annotator information. Overall, these findings highlight both the promise and the challenge of annotator-informed SER modeling.

6.5.1 Discussion

This work contributes to the gap between the growing interest in studying the subjectivity of emotion and the lack of exploration on the perception side. Statistical testing showed significant differences among the emotion labels assigned by people with different demographic background, laying a foundation and showing the promise of annotator-aware SER. For SER models, we found that integrating a lightweight annotator-aware adapter leads to better performance on personalized label prediction, even outperforming models trained with annotator IDs.

However, our model failed to learn generalizable patterns to predict unseen annotators given their demographics. Several factors may contribute to this limitation. First, while we focused on demographic and personality features, these likely capture only a small portion of the variability in emotion perception. As a result, the data are highly variable even with similar annotators, adding to the difficulty of model learning. A large and diverse dataset may better capture the full spectrum of subjective perspectives and help mitigate this variability.

Overall, the significance observed in statistical analyses and the gains in personalization demonstrate the promise of modeling subjectivity in SER. More broadly, our findings align with and contribute to a growing body of work on the influence of annotator characteristics in dataset construction and model behavior [208, 195]. Future work may investigate whether larger, more diverse datasets and alternative model architectures can support better generalization—while being mindful of the risk of stereotyping or learning spurious associations.

6.5.2 Limitation

Following our discussion, we note again that our dataset consists of a relative small group of annotators (n=200) and number of unique samples (n=500). As such, our goal is to highlight the importance and potential benefits of annotator-aware SER, rather than to draw strong conclusions about specific demographic groups. Additionally, both our statistical analyses and modeling approaches focused on the effects of individual annotator factors in isolation, without examining potential interactions between them. We also expect that perception-level factors may interact with characteristics of the speaker—for example, people may respond differently to speakers who share their gender or racial background [214]. These limitations point to promising directions for future work, including larger-scale studies and more nuanced models that jointly consider multiple sources of variation.

CHAPTER 7

Distilling Generalizable Emotion Knowledge from LLMs

7.1 Introduction

Previous chapters have shown that human emotion annotations are expensive to obtain and often variable, subjective, and sensitive to factors such as annotator characteristics and annotation design. Another important drawback is that current ER models are mostly trained on a fixed label set (e.g., 11-category classification) and struggle to generalize beyond them. As Emotion Recognition (ER) systems become integral to many applications [215, 6, 216], there is a growing need for models that can flexibly adapt to application-specific requirements. For instance, customer service systems often focus on identifying coarse-grained emotions such as positive, neutral, or negative emotions to understand customer experiences [217, 218], while mental health monitoring can benefit from more fine-grained emotional signals to capture subtle psychological states [215, 219]. Consequently, ER datasets have been developed to meet different needs. Common design choices include **label schema** (e.g., whether to use categorical or dimensional labels), **label granularity** (whether to adopt an extensive set of fine-grained emotions or a smaller set of distinct categories), and **label selection** (which specific emotions or dimensions to include in the label set) [49, 220, 47, 148]. While models can be trained to support multiple emotion label spaces in multitask setups [221], they are not designed to handle novel labels at inference. This limits their application across different downstream scenarios, resulting in a need for dedicated data collection and retraining for each new application.

On the other hand, LLMs achieve strong performance on various emotion spaces in a zero-shot manner without any extra training, especially on large, fine-grained ones [222, 160]. Moreover, previous work has found that LLMs can be prompted to generative descriptive emotion labels, which provides the flexibility beyond traditional fixed emotion categories and supports more accurate and nuanced emotion recognition [160, 223, 224, 225].

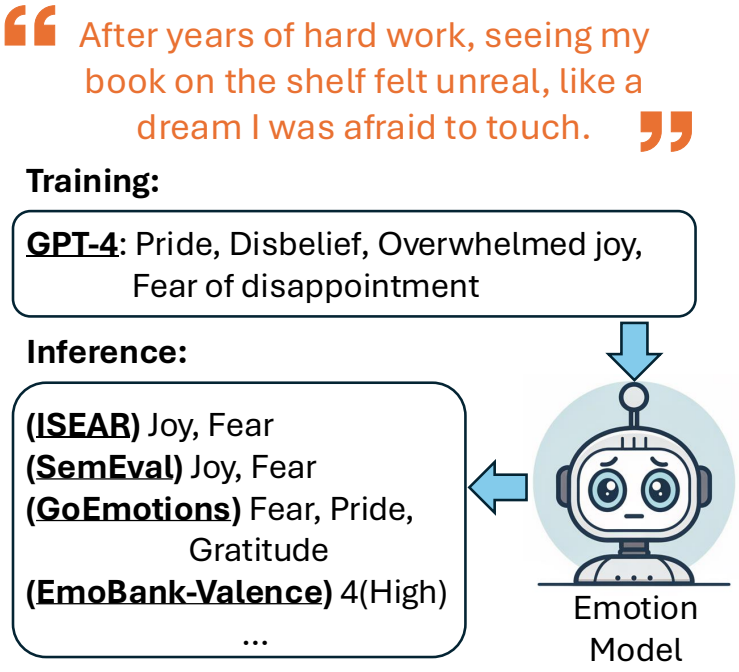


Figure 7.1: Our model is trained with rich emotion descriptors generated by GPT-4. During inference, this much smaller model can flexibly perform classification or regression on new label spaces.

Despite their desirable generalizability, directly applying LLMs to real work applications is limited by multiple factors. First, their scale is often prohibitive: leading LLMs can contain hundreds of billions of parameters, making it infeasible to deploy them on personal or edge devices [226]. Yet, many good or even state-of-the-art ER models rely on much smaller architectures [227, 228], suggesting that extensive computational overhead may not always be necessary for many applications. Additionally, the general-purpose nature of LLMs enables them to retain rich linguistic content, which can raise privacy concerns in sensitive settings [229]. Finally, LLMs produce free-form text outputs rather than structured emotion predictions. As a result, existing methods rely on prompt engineering to elicit responses in a specific format, requiring extra processing to extract usable emotion representations [230].

In this work, we propose a training approach that distills emotion knowledge from GPT-4's general language understanding capabilities, enabling generalizable zero-shot emotion recognition while preserving a compact model size. The overall structure of our training framework is illustrated in Figure 7.1. Instead of relying on fixed categorical labels, we prompt GPT-4 to generate nuanced, free-form emotion descriptors that offer rich supervision beyond predefined label sets. Our framework is compatible with any LLM capable of text-based interaction. We train our model using a contrastive learning framework inspired by

CLIP [231], to align text samples and emotion terms in a shared latent emotion embedding space. Our model has good zero-shot performance across multiple datasets with diverse label spaces, outperforming strong baseline models of comparable size. It also approaches GPT-4 zero-shot performance on multi-label emotion classification while being much smaller. We perform a nearest-neighbor analysis to show that our approach results in meaningful clustering of semantically related emotions, enhancing its interpretability and reliability.

In summary, we make the following contributions: 1) We introduce a method for generating rich, descriptive emotion annotations using GPT-4 without additional human annotation. 2) We propose a contrastive framework that extends CLIP to multi-label scenarios, enabling zero-shot ER across variable label spaces. 3) We evaluate our model against multiple supervised and zero-shot baselines, showing strong performance with a compact model size. We release the annotations for the GoEmotions dataset and our code to facilitate further research.

7.2 Methods

7.2.1 Obtaining Rich Emotion Annotations

We adopt a zero-shot prompting strategy following previous work [160] to encourage GPT-4 to generate multiple descriptive terms (words or phrases) that capture the nuanced emotion states conveyed in the text. We use the following system prompt to obtain annotations for the GoEmotions dataset (see Section 7.3.1): *You are an emotionally-intelligent and empathetic agent. You will be given a piece of text, and your task is to identify the emotions expressed by the writer of the text. Reply with only the emotion descriptors (words or phrases), separated by commas. If no emotion is clearly expressed, reply with "neutral".*

We use the "gpt-4-1106-preview" version of GPT-4 deployment through the Azure API. Across 43.4k samples in the GoEmotions training split, GPT-4 generated $2,173$ unique emotion terms, in contrast to 27 labels in human annotations. Each sample has 1.91 (standard deviation 0.65) terms on average and 13.0 (standard deviation 5.9) characters per term. We show the human classification and GPT-4 generated labels on 10 randomly selected samples in Table 7.1. Upon manual inspection, we notice that most terms in human labels also occur in GPT-4 generations, while GPT-4 tends to produce more complex or less common terms, such as "reminiscence" and "foreboding". It also uses extent descriptors like "faint optimism" and "a little bummed", or specifies emotional nuances or causes, e.g., "fear of abandonment", "fear of embarrassment", and "fear of failure". Finally, the format of the generated terms is not normalized; e.g., it includes a mixture of nouns ("fear") and adjectives

Text	Human Label	GPT-4 Label
At least they can make some good pizza	Neutral	Contentment, Satisfaction
Embrace the feels my friend. Glad you found happiness even if it is fleeting.	Joy	Supportive, Glad, Accepting
Congrats! Vegan baking is still daunting to me but I will conquer it one day!	Gratitude	Encouragement, Determination
I AM CALLING THE POLICE	Neutral	Urgency, Fear, Anger
Whew lad they're bleeding employees, reminds me of [NAME] before they bit the dust	Neutral	Concern, Reminiscence, Foreboding
Refrigerators? That's cool	Neutral	Amusement, Pun-Intended
I'm not sure what you mean by 18th century family here but otherwise thanks.	Confusion, Gratitude	Neutral
Thanks for the link. She actually likes some of the stuff!	Gratitude	Appreciation, Happiness
It sounds like you're the one who is afraid of the internet. Relax, bud. You're on r/cringe	Fear	Condescension, Annoyance
Forget new buildings, I work at the Chase Plaza building and am a little bummed that Seagram Tower made it but this didn't.	Disappointment, Neutral	Bemused, A Little Bummed

Table 7.1: Ten random samples selected from GoEmotions training set, comparing human-annotated emotion labels and GPT-4 generated descriptive labels.

("fearful") for the same emotion. We intentionally retain this variability, hypothesizing that exposure to diverse expressions may improve the generalizability of the model.

7.2.2 Contrastive Distillation

Our model structure is illustrated in Figure 7.2. During training, a batch of text samples is encoded into hidden representations using a text encoder. We collect all the emotion labels associated with the texts in the batch and embed them through a separate label encoder. We adopt pretrained BERT models for both the text and label encoders, keeping the label encoder frozen during training. BERT is chosen due to its strong text representation capabilities among models of similar sizes (around 100 million parameters). Also, BERT-based models are frequently used in zero-shot ER baselines (Olah et al. [232], Del Arco et al. [233], Bareiß et al. [234]), making them a strong and representative base model for our experiments.

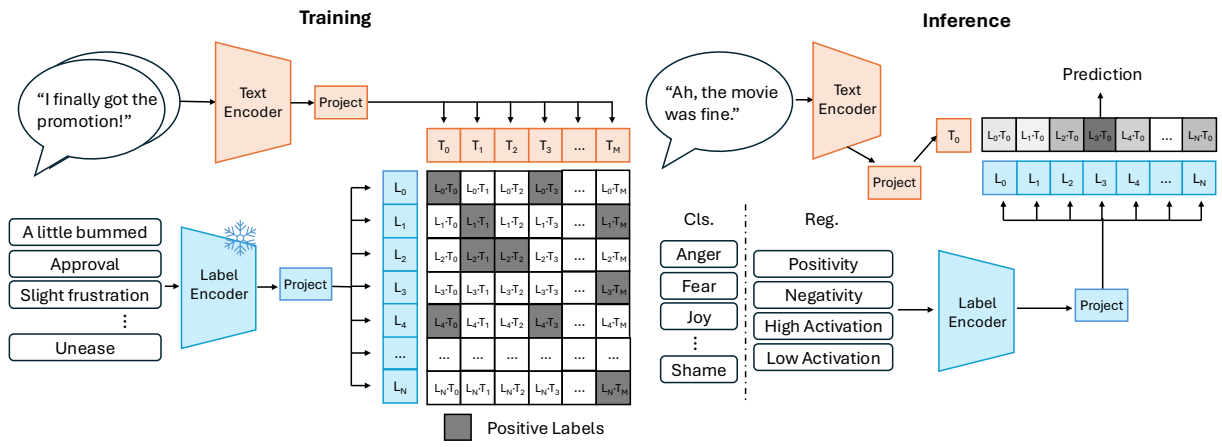


Figure 7.2: Overview of our contrastive distillation model structure.

Then, two separate projection modules transform the text and label embeddings into the same space, which we consider as our emotion representation space. On the text side, we adopt a Q-Former [235] to capture salient emotional information from the contextualized word representations. On the label side, since labels are relatively short, we apply a simple linear projection to their [CLS] token embedding [236]. We denote $\mathbf{T} \in \mathbb{R}^{B \times d}$ as the text embeddings after normalization, and $\mathbf{L} \in \mathbb{R}^{N \times d}$ as the label embeddings. B is batch size, N is the number of labels in that batch (varies across batches). The dimension of the learned emotion space d can be tuned as a hyperparameter. We obtain an alignment matrix $\mathbf{M} = \mathbf{T}\mathbf{L}^\top / \tau$. Since $\mathbf{T}$ and $\mathbf{L}$ are both normalized, the dot product returns the cosine similarity for each text-label pair. τ is a temperature scalar. We set $d = 768$ for fair comparison with baseline models (Section 7.3.2), and we set $\tau = 0.1$.

We modify the loss function in CLIP to enable multi-label training. We first apply a sigmoid function to the alignment matrix to obtain the predicted probabilities. Then, using the ground-truth binary label alignment matrix $\mathbf{Y} \in \{0, 1\}^{B \times N}$, we aim to maximize the log probability for matched (positive) text-label pairs, while minimizing the log probability for unmatched (negative) pairs. The resulting contrastive sigmoid loss is defined as:

$$\mathcal{L} = -\frac{1}{\sum \mathbf{Y}} \sum_{i=1}^{B} \sum_{j=1}^{N} \mathbf{Y}_{ij} \log(\sigma(\mathbf{M}_{ij})) - \frac{1}{\sum (1 - \mathbf{Y})} \sum_{i=1}^{B} \sum_{j=1}^{N} (1 - \mathbf{Y}_{ij}) \log(1 - \sigma(\mathbf{M}_{ij}))$$

where $\sigma(\cdot)$ denotes the sigmoid function. The alignment matrices are generally very sparse, so we average the log probability for positive and negative pairs separately before adding them together to balance the precision and recall.

7.2.3 Zero-shot Inference

We consider three common formulations in text-based ER: single-label classification, multi-label classification, and regression. Our trained model can flexibly adapt to different emotion label spaces at inference time by changing the set of candidate labels, as shown in Figure 7.2 (Inference).

As discussed in Section 3.2, the model is trained to predict the GPT-4 generated emotion labels. However, during inference, these labels are replaced with the fixed set of human labels that are specific to each test dataset. For single-label classification, we output the label from the fixed set that has the highest probability. For multi-label classification, we apply label-specific thresholds, calibrated on a small validation set (see Section 7.3.3) and output all labels that are above this threshold. For regression tasks we convert the original valence and activation labels to four discrete categories: "Positivity", "Negativity", "High Activation", and "Low Activation". The predicted valence score is then computed as the difference $M[\text{text}, \text{Positivity}] - M[\text{text}, \text{Negativity}]$, and the predicted activation score is computed as $M[\text{text}, \text{High Activation}] - M[\text{text}, \text{Low Activation}]$.

With contrastive training, our model learns a compact and reusable emotion space from GPT-4's descriptive supervision, and it can generalize seamlessly to various label spaces.

7.3 Experiments and Results

7.3.1 Datasets

We use the same text-based ER datasets detailed in Section 4.2 for training and evaluation, covering diverse emotion label spaces. The training set is constructed using text samples from the GoEmotions dataset paired with rich descriptive annotations generated by GPT-4 (see Section 7.2.1). GoEmotions is selected for training due to its diverse emotional content and its fine-grained human annotation schema, which allows a controlled comparison between human and GPT-4 supervision. We evaluate the trained model's zero-shot performance on three additional datasets, each with distinct label spaces.

7.3.2 Baseline Models

Upper-bound models. We contextualize our model's performance with two upper-bound baselines. First, we evaluate **GPT-4 zero-shot (GPT-ZS)** performance by prompting GPT-4 to perform zero-shot ER. The prompts are adapted from prior work [160], formatted similarly to our generation prompts (Section 7.2.1) but modified to elicit dataset-specific

outputs. Given that GPT-4 is over 10k times larger than BERT[1], and our model is distilled from its outputs, we consider GPT-ZS as a performance upper bound. Second, we include a **finetuned BERT model (BERT-FT)**, which shares the same backbone as our model but is trained with direct supervision. We train a separate model for each evaluation dataset and select the best checkpoint based on validation performance (macro-F1 for classification; PCC for regression).

Baseline Models Several studies have explored zero-shot emotion recognition in unseen label spaces. Earlier approaches have used soft labels on a set of seen emotions as features to predict unseen ones, effectively learning correlations between emotion categories [237, 238]. These approaches are typically applied to coarse-grained emotion classes and often benefit from few-shot training. Similarity-based methods are also popular. For example, Olah et al. [232] proposed to obtain semantic embeddings of the emotion label and text input and use cosine similarity to measure their alignment. Stanley et al. [224] applies a similar approach to speech ER. Another line of work is entailment-based ER, which frames the task as predicting sentence entailment using crafted input like "this text expresses [emotion]" [239, 233, 234]. In general, zero-shot ER remains highly challenging: many studies report a gap of 0.3-0.5 in macro-F1 between finetuned and zero-shot setups [232, 233, 240].

We compare our model against three baseline models. The first is a BERT model finetuned on GoEmotions and directly applied to other datasets without further adaptation (**BERT-ZS**). For classification tasks (SemEval and ISEAR), we reuse the prediction heads for overlapping emotion labels and randomly initialize new heads for unseen labels. For regression tasks (EmoBank), the prediction heads are entirely randomly initialized. This serves as a conservative and relatively weak baseline. We also include two strong zero-shot baselines with comparable model sizes from prior work. The first is a **similarity-based** method [232], where cosine similarity is computed between the embedding of the input text and each emotion label augmented with its definition. The second is an **entailment-based** method [234], where a BERT model is finetuned on multiple NLU tasks and predicts the plausibility of the crafted input "[text] This text expresses [emotion]." We use BERT as the base model across the baselines and our methods.

7.3.3 Experimental Setup and Metrics

Training. To ensure that no validation or test samples from GoEmotions are seen during the training of our model, we reserve 20% of GoEmotions training set for model selection.

[1]Estimate based on public parameter counts: GPT-4 > 100B vs. BERT-base around 110M.

Additionally, to assess model fit on the training task itself, we report micro-averaged F1 scores on the GoEmotions test set. (Macro-averaging is not feasible due to the extensive number of training labels.)

Evaluation. For classification datasets, we evaluate performance using the macro-averaged F1-score to reflect overall classification effectiveness across classes. For regression tasks, we report Pearson correlation coefficient (PCC) to measure the linear relationships between predicted and ground-truth scores, and Spearman's rank correlation coefficient (ρ) to assess monotonic relationships. For EmoBank, we report regression performance separately for valence and activation dimensions.

It is important to note that comparisons on the GoEmotions test set are not fully controlled, as models differ in the additional information available during training: our model is exposed to the text samples in the GoEmotions training set, while the entailment-based baseline is finetuned on a broad range of natural language understanding (NLU) tasks, including some emotion-related tasks. The similarity-based baseline relies on emotion label definitions from WordNet. We still report their performance on GoEmotions as zero-shot since none of the models have seen the GoEmotions human annotations or test set text samples, but we note that the results should be interpreted with care.

For multi-label classification tasks, we perform an additional threshold calibration step on a validation set for our model and the baseline methods. Specifically, we determine the optimal threshold for each emotion class using the validation split of each dataset by searching between 0 and 1 (in increments of 0.05), selecting the value that yields the best performance on the validation set. While this setup is not strictly zero-shot, it is consistent with prior zero-shot SER work [232]. We believe this step is necessary because of the nature of multi-label classification: in this setup, the goal is not to output the most likely emotion, but instead to identify all emotions that are present. In this case, a threshold is needed to make this judgment. We further discuss this limitation in the Limitation Section. The regression and single-label classification tasks are evaluated in a strictly zero-shot manner.

For all baseline models, we report results based on our own replication to ensure consistency across datasets and evaluation setups. Our replicated results are comparable to those reported in the original papers. Additional training and hyperparameter details can be found in our released code.

7.3.4 Overall Performance

We compare the zero-shot performance of our model and all baseline methods in Table 7.2.

First, we find that dataset-specific fine-tuning provides significant benefits: BERT-FT

Model	GoEmotions macro-F1	SemEval macro-F1	ISEAR macro-F1	EmoBank-V Rho (ρ)	EmoBank-V PCC	EmoBank-A Rho (ρ)	EmoBank-A PCC
BERT-FT	<u>0.477</u>	<u>0.563</u>	0.677	<u>0.764</u>	<u>0.790</u>	<u>0.470</u>	<u>0.556</u>
GPT-4-ZS	0.319	0.486	<u>0.728</u>	0.678	0.718	0.369	0.324
BERT-ZS	/	0.205	0.336	-0.033	-0.004	-0.011	-0.022
Similarity	0.137	0.391	**0.504**	0.428	0.401	0.110	0.119
Entailment	0.073	0.430	0.222	0.534	0.485	0.102	0.105
Ours	**0.299 (0.005)**	**0.480 (0.003)**	0.479 (0.015)	**0.613 (0.008)**	**0.593 (0.003)**	**0.155 (0.028)**	**0.202 (0.040)**

Table 7.2: Zero-shot performance across four datasets. For our model, we train the model with five random seeds and we report the mean and standard deviation (in parentheses) of the performance metrics. The best overall performance is underlined. The best zero-shot performance among comparably sized models is highlighted in bold.

achieves the best performance on most tasks. Although GPT-4 is substantially larger, it only outperforms BERT-FT on the ISEAR dataset. This is possibly due to the specificity of emotion datasets: label sets, annotation instructions, and cultural or contextual assumptions can vary widely, making it difficult even for powerful general-purpose models to fit specific emotion distributions without adaptation.

Our model shows competitive performance on multi-label classification tasks. It outperforms both strong zero-shot baselines and approaches GPT-4 performance (e.g., 0.486 for GPT-4 vs. 0.466 for ours on SemEval). Although our model has been exposed to GoEmotions training texts (but not the specific label space, see Section 7.3.3) and may be more familiar with the text domain, the substantial performance margin over baselines suggests a genuine ability to generalize to new label spaces. This is further validated on SemEval, where all models are fully zero-shot, and ours achieves 0.466 F1, outperforming 0.391 for the similarity-based method and 0.430 for the entailment-based method. We observe that the performance of entailment-based method varies substantially across datasets (e.g., strong on SemEval but nearly-random on GoEmotions), likely due to their sensitivity to domain shifts and prompt formulations [239].

Our model performs slightly worse than the similarity-based baseline on the single-label ISEAR dataset (macro-F1: Ours 0.481, Similarity 0.504, Entailment 0.222). We suspect this is due to our model's multi-label training setup, which encourages capturing multiple plausible emotions rather than selecting the most dominant one. For instance, it often confuses guilt and shame, which do naturally co-occur. Further analysis and targeted experiments may help clarify this behavior or improve the model for single-label scenarios.

Finally, although not explicitly trained for regression tasks, our model and both base-

Dim	Train	GPT-4 distillation		
		G	S	I
50	0.428	0.282	0.461	0.451
100	0.443	0.289	0.461	0.433
200	0.447	0.294	0.466	0.472
768	**0.454**	**0.307**	**0.466**	**0.481**

Table 7.3: Comparison of emotion space dimensionality. "Train" column shows the performance on GoEmotions test set with GPT-4 labels, measured by micro-F1. "G", "S", and "I" refer to performance on GoEmotions, SemEval, and ISEAR respectively, measured by macro-F1.

line models achieve surprisingly strong results on valence regression. Our model obtains a Spearman correlation (ρ) of 0.563 and a PCC of 0.550, outperforming 0.534 ρ and 0.485 PCC for the entailment-based model, and 0.428 ρ and 0.401 PCC for the similarity-based model. Performance on activation prediction is notably lower across all models. Activation is generally more difficult to infer from text alone [148], and activation-related terms such as "high activation" or "low activation" are less commonly included in both human-annotated labels and GPT-4-generated descriptions. Overall, our approach shows encouraging results for zero-shot regression tasks, but further research is needed to close the performance gap between zero-shot and supervised models.

7.3.5 Ablation Studies

Since the standard deviation of performance metrics was found to be small during training (see Table 7.2), we use a fixed random seed (42) for all ablation studies to reduce computational cost.

Dimension Size. We investigate the impact of the emotion space dimension d. In the Table 7.2, we use $d = 768$ to maintain consistency with the baseline models for fair comparison. However, smaller d is desirable for downstream applications due to lower computational costs and smaller model sizes.

To explore this trade-off, we conducted additional experiments with $d \in \{200, 100, 50\}$ across the classification datasets. As shown in Table 7.3, performance generally drops slightly as the dimension decreases, but the decline remains small even under aggressive reductions (e.g., $d = 50$). This suggests that our model can maintain strong performance even with a compact emotion space, making it suitable for resource-efficient applications.

GPT-4 Labels. We next probe the effect of using GPT-4 generated labels for supervision. We compare our model with a variant trained on the same samples but supervised

	Train	G	**S**	**I**
Ours (GPT-4 labels)	0.307	0.466	0.481	0.563
Human labels				
overall	0.587	0.475	0.414	0.410
seen classes	/	0.475	0.509	0.488
unseen classes	/	/	0.161	0.215

Table 7.4: Comparison of models trained with GPT-4 generated labels versus human-labels. For models trained with human labels, we also report separate results on classes seen in the training dataset (8 out of 11 in SemEval, 5 out of 7 in ISEAR). Note that GoEmotions performance under human supervision reflects a supervised setting, while all others are zero-shot evaluations.

with human annotations, rather than GPT-4 generated annotations. As Table 7.4 shows, models trained on human labels perform better on the GoEmotions dataset itself—as expected, due to direct supervision—but exhibit worse zero-shot generalization to SemEval and ISEAR. The model trained on human labels performs well on seen classes, even outperforming our GPT-4-distilled model on SemEval, but its performance drops sharply on unseen classes, with macro-F1 scores of only 0.161 for SemEval and 0.215 for ISEAR. Yet, both models share the same contrastive architecture, making it theoretically possible for the human-supervised model to generalize to unseen labels because of BERT's existing semantic embedding space. However, the richness of the supervision makes a substantial difference. Notably, GoEmotions already provides one of the most extensive categorical label sets among text-based ER datasets. These results suggest that it remains difficult to learn a generalizable emotion space from a fixed and limited set of labels, underscoring the advantage of distilling from rich, descriptive annotations.

7.3.6 Emotion Space Probing

Finally, to interpret the learned emotion space, we examine its nearest-neighbor structure. We use all 27 classes from the GoEmotions dataset as target emotions and GPT-4-generated emotion description terms on its test split as the candidate pool ($N = 684$). We only use the test set to ensure that the retrieved structure reflects generalization rather than overfitting to the training data.

We compare our model against a BERT encoder baseline. For BERT, we extract the [CLS] token embeddings of all emotion labels/terms. For our model, we encode these terms using our trained Label Encoder and Projector (as shown in Figure 7.2. Since the Label

Target	BERT (768D)	Ours (768D)	Ours (50D)
Admiration	Sympathy, Gratitude, Satisfaction, Reluctant Admiration	Reverence, Amazement, Awe, Attraction	Accomplishment, Reverence, Amazement, Pride
Gratitude	Satisfaction, Admiration, Affection, Jealousy	Appreciation, Grateful, Thankful, Politeness	Appreciation, Grateful, Thankful, Recognition
Approval	Disapproval, Recommendation, Fairness, Acceptance	Positive Surprise, Positive, Validation, Recommendation	Positive Interest, Positive Surprise, Positive, Favorability
Annoyance	Distraction, Reluctance, Frustrated, Avoidance	Irritation, Annoyed, Irritated, Exasperation	Irritation, Annoyed, Reluctance, Irritated
Curiosity	Suspicion, Surprise, Aggression, Distraction	Interest, Intrigue, Speculation, Curious	Interest, Slight Hopefulness, Curious, Fascination
Disapproval	Annoyance, Reluctance, Welcoming, Dismay	Judgment, Criticism, Condemnation, Dismay	Negative Opinion, Judgment, Criticism, Disagreement
Amusement	Laughter, Reluctant Admiration, Mock Frustration, Amused	Mild Amusement, Lack of Amusement, Lightheartedness, Mild Sarcasm	Mild Amusement, Lack of Amusement, Mock Frustration, Mock Seriousness
Love	Passion, Joy, Happiness, Pain	Affection, Passion, Adoration, Fondness	Affection, Joy, Good Wishes, Lust
Anger	Rage, Jealousy, Resentment, Contempt	Rage, Hostility, Hatred, Outrage	Rage, Hostility, Indignation, Hatred
Optimism	Cautious Optimism, Calmness, Hopefulness, Reassurance	Hopefulness, Hope, Optimistic, Hopeful	Hopefulness, Hope, Cautious Optimism, Hopeful
Joy	Hope, Grief, Calm, Celebration	Happiness, Celebration, Delight, Joyful	Happiness, Adoration, Celebration, Excitement
Sadness	Grief, Sad, Heartbreak, Disappointment	Sorrow, Heartbreak, Sad, Grief	Melancholy, Sad, Sorrow, Heartbreak
Confusion	Horror, Resentment, Jealousy, Impatience	Seeking Clarification, Seeking Help, Seeking Assistance, Seeking Advice	Confused, Awkwardness, Historical Trauma, Puzzlement
Disappointment	Regret, Frustration, Guilt, Relief	Dissatisfaction, Regret, Mild Disappointment, Disappointed	Dissatisfaction, Regret, Disappointed, Heartbreak
Realization	Disbelief, Distraction, Shock, Desperation	Surprise, Reflection, Denial, Acknowledgment	Denial, Shock, Focus, Shame
Surprise	Curiosity, Astonishment, Surprised, Realization	Realization, Astonishment, Mild Surprise, Shock	Mild Surprise, Shock, Astonishment, Realization
Caring	Denial, Kindness, Compassion, Protective	Empathetic, Considerate, Reassuring, Supportive	Considerate, Self-Assured, Affectionate, Accepting
Disgust	Irritation, Disgusted, Shame, Paranoia	Loathing, Disgusted, Horror, Hatred	Outrage, Disgusted, Condemnation, Dismay
Excitement	Delight, Anticipation, Panic, Astonishment	Joy, Delight, Enthusiasm, Triumph	Joy, Delight, Enthusiasm, Laughter
Desire	Lust, Jealousy, Longing, Affection	Lust, Longing, Craving, Eagerness	Longing, Lust, Passion, Misery
Fear	Panic, Distrust, Dread, Worry	Anxiety, Dread, Terror, Panic	Anxiety, Dread, Terror, Alarm
Remorse	Regret, Apology, Compassion, Guilt	Regret, Apology, Regretful, Apologetic	Regret, Apology, Regretful, Guilt
Embarrassment	Humiliation, Guilt, Concern, Frustration	Guilt, Humiliation, Self-Irritation, Shame	Guilt, Humiliation, Discomfort, Aversion
Nervousness	Tiredness, Helplessness, Awkwardness, Eagerness	Anxiety, Fear of Embarrassment, Numbness, Tiredness	Numbness, Desire for Emotional Relief, Fear of Embarrassment, Anxiety
Pride	Jealousy, Satisfaction, Empathy, Affection	Accomplishment, Confidence, Loyalty, Triumph	Reverence, Accomplishment, Belief, Authenticity
Relief	Relieved, Disappointment, Irritation, Grief	Gladness, Relieved, Glad, Satisfaction	Good, Relieved, Glad, Gladness
Grief	Heartbreak, Sadness, Desperation, Relief	Sorrow, Heartbreak, Sadness, Misery	Sorrow, Heartbreak, Sadness, Pain

Table 7.5: Top-4 most similar emotion terms retrieved for each target emotion using cosine similarity, using the BERT encoder, our contrastively distilled emotion model (768D), and its reduced 50D version.

Encoder is initialized with the BERT encoder and frozen during training, our encoder only differs from BERT by one linear layer. For each target emotion, we retrieve the top-4 most similar terms from the candidate pool using cosine similarity and the results are shown in Table 7.5.

Manual inspection suggests our model retrieves more emotionally aligned neighbors, compared to BERT. For instance, for "Admiration", our model returns "Reverence" and "Amazement", whereas BERT returns "Sympathy" and "Gratitude". We also observe that BERT tends to prioritize part-of-speech consistency, e.g., failing to retrieve "annoyed" for "annoyance" or "grateful" for "gratitude". In some cases, it even retrieves antonyms such as "Disapproval" for "Approval". These behaviors are likely due to semantic relatedness in general language space, but are unfavorable for emotion-specific use cases. Our contrastive training helps mitigate these effects. Additionally, we conduct the same experiment using our 50D model, aggressively compressing the emotion space. The retrieved neighbors remain largely consistent with those from the 768D model and, in our judgment, still align more closely with the intended target emotions than BERT.

These results demonstrate that our approach can learn a more efficient representation space that better preserves emotion saliency compared to language-focused embeddings.

7.4 Conclusion

We present a contrastive distillation framework that extracts emotional knowledge from LLMs' into a compact, BERT-sized model. Our method learns to map both text inputs and GPT-4 generated label descriptors into a joint representation space that enables zero-shot prediction across diverse label sets and task types. Experiments show strong performance across datasets and label spaces, outperforming comparable zero-shot baselines and approaching GPT-4 zero-shot performance while remaining far smaller in size. We discuss practical directions to further improve the model, as well as potential ethical considerations surrounding emotion annotation and model fairness.

7.4.1 Discussions

Zero-shot generalization is a highly desirable capability for ER systems, as it enables flexible adaptation to applications without the need for extra data collection or retraining. In this work, we design a compact model that distills emotional knowledge from GPT-4 and learns a flexible emotion space, achieving zero-shot generalization without the prohibitive scale of LLMs.

Our results are encouraging. First, we demonstrate that our contrastive learning framework enables the model to handle diverse ER setups, including both multi- and single-label classification, as well as regression. Second, through a nearest-neighbor retrieval analysis, we show that our model captures emotional saliency from general language representations, and this structure remains largely preserved even when the embedding dimensionality is reduced from 768 to 50. Together, these results suggest that our approach yields compact, generalizable emotion representations that can be readily applied to a variety of downstream tasks. In settings where further (few-shot) tuning is possible, these representations provide a strong and efficient starting point for adaptation.

Our results also invite reflection on what makes an effective representation space for emotions. While LLMs demonstrate strong emotion understanding, they operate in the full language space, which encodes a wide range of information beyond emotion. As a result, they can be larger and more resource-intensive than necessary for emotion tasks. In contrast, our approach distills a dedicated emotion space that focuses on emotionally salient features, showing that strong emotion understanding can be achieved with significantly less computation. We hope these findings encourage future research in this domain.

7.4.2 Limitation

For zero-shot inference on multi-label classification tasks, we calibrate prediction thresholds on a validation set for both our model and baseline models. We also ran experiments without calibration, where all models showed significant drops in F1 scores, but are still above random baseline and the relative ranking among the models remains the same. We think that threshold calibration is an essential step for multi-label classification under the current setups, where the model independently predicts each label without considering the full label set. However, human judgments are influenced by the full set of available alternatives, For example, in ISEAR, all positive samples are typically labeled as "joy", as it is the only available positive category. If given more fine-grained options, annotators may choose more accurate descriptions like "proud' or "excited" while dropping "joy". The calibration step serves as weak supervision to help models adjust to these differences. To remove this constraint, future work could explore methods that considers all label options jointly for each sample.

While our framework shows strong potential, there are several practical directions for further improving its generalizability and robustness. First, this work focuses on generalization to unseen labels, while new text domains can also pose challenges. Instead of using GoEmotions as the sole source of text for supervision, future work could incorporate more

diverse textual sources to improve generalization across varied contexts. Second, although our model is compatible with multiple emotion label spaces, the current loss design is best aligned with multi-label classification—where we also observe the strongest empirical performance. Future work could explore multi-stage training strategies to better prepare the model for specific downstream applications while preserving its zero-shot generalization ability.

CHAPTER 8

Conclusion

8.1 Summary of Contributions

This thesis explores three sources of inconsistencies commonly seen but not widely studied in affect modeling: annotation inconsistency, modality inconsistency, and interpersonal inconsistency. Chapter 1 first introduces the tasks, existing approaches, and challenges in quantifying and modeling effect. Chapter 2 and Chapter 3 explore the implications of modality inconsistency, proposing its use as a signal for mood disorder tracking and extending the idea to general-purpose speech representation learning. Chapter 4 and Chapter 5 focus on annotation inconsistency, considering the use of LLMs in annotation pipelines and examining how design choices in annotation experiments influence emotion annotations by both humans and LLMs. Chapter 6 focuses on interpersonal inconsistency, analyzing how annotator demographics and personality traits influence emotion perception and assessing the effectiveness of annotator-aware models in improving personalization and generalization in SER. Finally, Chapter 7 builds on findings from previous chapters and proposes a training framework that transfers emotion knowledge from LLMs into a compact model, enabling zero-shot emotion recognition across varies label spaces and producing efficient, emotion-salient representations.

In Chapters 2 and 3, we explored modality inconsistency, a common phenomenon where semantic and acoustic modalities in speech convey misaligned information. Chapter 2 built on clinical insights into Affect Blunting, focusing on emotional mismatch as an indicator of symptoms of bipolar disorder. In Chapter 3, we extended this finding to general speech representation learning. We proposed a multitask finetuning strategy to factorize acoustic and semantic embeddings, demonstrating its advantages in preserving distinct information and improving model robustness on multiple downstream tasks, including speech emotion recognition and emotional speech ASR.

In Chapters 4 and 5, we focused on annotation inconsistency, defined as disagreements

and low-quality labels in emotion annotation experiments, influenced by design choices in annotation experiments. We first conducted comprehensive studies that evaluated human and GPT-4 annotations across different emotion label spaces. Results showed a significant amount of noise in human emotion labels and strong potential of LLMs like GPT-4 to aid annotation processes, especially on large label spaces. On the other hand, the results also showed perspective differences between GPT-4 and humans in their labels, highlighting the continued importance of human perspectives in annotations. Following these insights, we proposed two methods to incorporate GPT-4 into human emotion annotation pipelines: label-level pre-filtering, which keeps the descriptive power of large emotion spaces while reducing the burden of human annotators, and sample-level post-filtering, which improves downstream training efficiency and performance.

In Chapter 6, we presented our work around interpersonal inconsistency, investigating how annotators' characteristics, such as demographics and personality traits, influence their emotion labeling. We introduced a set of new SER annotations with a large number of annotators per sample (compared to common datasets) and rich annotator information. Statistical analyses revealed significant effects of demographic and personality factors on dimensional emotion labels. Further, incorporating annotator information into SER models through lightweight adapters led to improved personalized SER performance.

In Chapter 7, we introduced a modeling approach which learns from LLMs' strong zero-shot emotion recognition capability through their generated, rich emotion labels (beyond fixed label sets). Our distilled model learned an efficient representation space with a focus on emotional saliency and can perform zero-shot inference on various unseen categorical and dimensional label spaces.

Together, this thesis calls attention to various inconsistencies in affect modeling, both by understanding their underlying causes from application perspectives and by developing strategies to mitigate or leverage them from signal processing and modeling perspectives. We explore several approaches that leverage the strong language capability of LLMs both to enhance human emotion annotations and to support the training of more robust emotion recognition models, highlighting their promise for future research in affective computing.

8.2 Limitations and Future Prospects

This thesis underscores the importance of understanding and addressing inconsistencies in affect modeling. We presented our exploration, yet several limitations and open questions remain.

Our work around modality inconsistency motivates further exploration of domain-specific

insights, such as leveraging modality-specific signals to understand emotion regulation in healthy individuals, extending beyond the domain of mood disorders. Further, the proposed multitask factorization method is a relatively coarse approach that does not achieve full separation of semantic and acoustic information. Future research could explore improved supervision tasks, loss functions (e.g., Mutual Information Loss [241]), or architectural innovations (e.g., domain-adversarial training [242]) to impose stricter disentanglement and better generalization across diverse tasks.

Our findings on annotation inconsistency highlight the sensitivity of annotation outcomes to the choice of emotion label spaces. However, designing an emotion annotation study involves many more considerations, such as sample selection, annotation interface design, pre-annotation training phases, quality control, etc. While previous work has explored some factors for human annotations, it is worth reconsidering them in the context of incorporating LLMs. Additionally, we demonstrated the potential of LLMs as powerful tools to assist human annotation and presented preliminary approaches. These findings can serve as a foundation for future refinements and advancements: designing emotion annotation studies with LLMs integrated into the pipeline may be of interest to both the emotion research and Human-Computer Interaction (HCI) communities.

For interpersonal inconsistency, Chapter 6 presents a preliminary study. Our dataset consists of a relative small group of annotators ($n = 200$) and number of unique samples ($n = 500$). As such, our experiments serve as a proof-of-concept, highlight the importance and potential benefits of annotator-aware SER, rather than to draw strong conclusions about specific demographic groups. Additionally, both our statistical analyses and modeling approaches focused on the effects of individual annotator factors in isolation, without examining potential interactions between them. We also expect that perception-level factors may interact with characteristics of the speaker, for example, people may respond differently to speakers who share their gender or racial background [214]. These limitations point to future work with larger-scale studies and more nuanced models that jointly consider multiple sources of variation.

8.3 Ethical Considerations

This thesis explores various aspects of affect modeling. While these explorations aim to improve the reliability and fairness of emotion recognition systems, they also raise important ethical considerations. We highlight two aspects below.

8.3.1 Privacy Protection and Data Use

Our experiments and potential applications of our proposed models involve the collection and use of human-subject data, for example, speech recordings, written content, and demographic information. These datasets may contain sensitive details, such as biometric identifiers (e.g., voice) and personally identifiable information (e.g., names or locations). To ensure ethical handling of such data, we adhere to institutional ethical guidelines. All experiments involved human subjects presented in this paper have been approved by University of Michigan's IRB. We also follow guidelines proposed by researchers in the Affective Computing community, including proper consent, secure storage, responsible use, etc. [243, 244].

8.3.2 Bias and Fairness

In Chapter 6, we demonstrated that emotion annotations can be influenced by demographic and personality traits. While our exploration aimed to incorporate annotator-specific information to improve accuracy and fairness, there also exist risks of reinforcing harmful biases. Overgeneralizing the difference in emotion perception across different demographic groups risks misrepresenting individuals and communities, potentially leading to inconsistent findings, ignorance of confounding factors, and even discriminatory outcomes [245, 246, 183]. Similarly, models trained on biased datasets may learn and amplify these biases, making it essential to implement strategies that mitigate such risks [91].

The use of LLMs in annotation collection and generation can also introduce potential bias. In Chapter 7, we distill emotion supervision from GPT-4. This approach has potential benefits: LLMs like GPT-4 are trained on large-scale, diverse data and are explicitly designed with fairness considerations in mind [247], which may reduce some forms of annotator-specific bias. However, any representational or linguistic biases in GPT-4 will propagate into the distilled model. In our annotations, we also observe that GPT-4 tends to generate emotion terms that are more complex or infrequent, sometimes diverging from how many people naturally express emotions in everyday settings. While such richness can enhance expressiveness, it may also reflect linguistic or cultural preferences that do not generalize across populations. These patterns underscore the need for careful reflection on the sources and implications of emotion supervision, especially when deploying models in sensitive applications.

BIBLIOGRAPHY

[1] Gustavo Assunção, Bruno Patrão, Miguel Castelo-Branco, and Paulo Menezes. An overview of emotion in artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 3(6):867–886, 2022.

[2] Chiara Zucco, Barbara Calabrese, and Mario Cannataro. Sentiment analysis and affective computing for depression monitoring. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1988–1995. IEEE, 2017.

[3] Emily Mower Provost, Sarah H Sperry, James Tavernor, Steve Anderau, Anastasia Yocum, and Melvin G McInnis. Emotion recognition in the real-world: Passively collecting and estimating emotions from natural speech data of individuals with bipolar disorder. *IEEE Transactions on Affective Computing*, 2024.

[4] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[5] Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. Affective computing in education: A systematic review and future research. *Computers & education*, 142: 103649, 2019.

[6] Yuanyuan Liu, Ke Wang, Lin Wei, Jingying Chen, Yibing Zhan, Dapeng Tao, and Zhe Chen. Affective computing for healthcare: Recent trends, applications, challenges, and beyond. *arXiv preprint arXiv:2402.13589*, 2024.

[7] Florian Eyben, Martin Wöllmer, Tony Poitschke, Björn Schuller, Christoph Blaschke, Berthold Färber, and Nhu Nguyen-Thien. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, 2010(1):263593, 2010.

[8] Leimin Tian, Sharon Oviatt, Michal Muszynski, Brent Chamberlain, Jennifer Healey, and Akane Sano. *Applied Affective Computing*. Morgan & Claypool, 2022.

[9] Dacher Keltner and Daniel T Cordaro. Understanding multimodal emotional expressions: Recent advances in basic emotion theory. 2017.

[10] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2462–2465. IEEE, 2010.

[11] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. The state of speech in hci: Trends, themes and challenges. *Interacting with computers*, 31 (4):349–371, 2019.

[12] Sameera A Abdul-Kader and John C Woods. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, 6(7), 2015.

[13] Susan R Fussell. The verbal communication of emotion: Introduction and overview. In *The verbal communication of emotions*, pages 9–24. Psychology Press, 2002.

[14] JO-ANNE BACHOROWSKI and Michael J Owren. Sounds of emotion: production and perception of affect-related vocal acoustics. *Annals of the New York Academy of Sciences*, 1000(1):244–265, 2003.

[15] Sona Patel, Klaus R Scherer, Eva Björkner, and Johan Sundberg. Mapping emotions into acoustic space: The role of voice production. *Biological psychology*, 87(1):93–98, 2011.

[16] Ze-Jing Chuang and Chung-Hsien Wu. Multi-modal emotion recognition from speech and text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, pages 45–62, 2004.

[17] Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. Multimodal emotion recognition with high-level speech and text features. In *2021 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 350–357. IEEE, 2021.

[18] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15:99–117, 2012.

[19] K Sreenivasa Rao and Shashidhar G Koolagudi. *Emotion recognition using speech features*. Springer Science & Business Media, 2012.

[20] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[21] Manas Jain, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, Rajesh Kumar Muthu, et al. Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*, 2020.

[22] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *2005 international conference on machine learning and cybernetics*, volume 8, pages 4898–4901. IEEE, 2005.

[23] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.

[24] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech 2015*, 2015.

[25] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Interspeech*, 2021.

[26] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023.

[27] Janneke Wilting, EJ Krahmer, and MGJ Swerts. Real vs. acted emotional speech. In *Proceedings of the international conference on spoken language processing (interspeech 2006)*. ISCA, 2006.

[28] Pashiera Barkhuysen, EJ Krahmer, and MGJ Swerts. Cross-modal perception of emotional speech. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS'07)*, pages 2133–2136. ICPhS, 2007.

[29] James J Gross and Oliver P John. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348, 2003.

[30] Itziar Fernández, Pilar Carrera, Flor Sánchez, Darío Paez, and Luis Candia. Differences between cultures in emotional verbal and non-verbal reactions. *Psicothema*, 12(Su1): 83–92, 2000.

[31] Peter Kuppens, Iven Van Mechelen, John B Nezlek, Dorien Dossche, and Tinneke Timmermans. Individual differences in core affect variability and their relationship to personality and psychological adjustment. *Emotion*, 7(2):262, 2007.

[32] Claudia Marinetti, Penny Moore, Pablo Lucas, and Brian Parkinson. Emotions in social interactions: Unfolding emotional experience. *Emotion-oriented systems: The humaine handbook*, pages 31–46, 2011.

[33] UnYoung Chavez-Baldini, Marieke Wichers, Ulrich Reininghaus, Johanna TW Wigman, Genetic Risk, and Outcome of Psychosis Investigators. Expressive suppression in psychosis: The association with social context. *Plos one*, 15(3):e0230102, 2020.

[34] Judith A Hall, Jason D Carter, and Terrence G Horgan. Gender differences in nonverbal communication of emotion. *Gender and emotion: Social psychological perspectives*, pages 97–117, 2000.

[35] Jacob W Gurera and Derek M Isaacowitz. Emotion regulation and emotion perception in aging: A perspective on age-related differences and similarities. *Progress in brain research*, 247:329–351, 2019.

[36] A George Gitter, Harvey Black, and David Mostofsky. Race and sex in the perception of emotion. *Journal of Social Issues*, 28(4):63–78, 1972.

[37] Carlos Busso, Murtaza Bulut, Shrikanth 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*, 7(17): 110–127, 2013.

[38] Lowri Williams, Michael Arribas-Ayllon, Andreas Artemiou, and Irena Spasić. Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification*, 36:619–648, 2019.

[39] Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower 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)*, pages 7415–7419. IEEE, 2019.

[40] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

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

[42] Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion review*, 3(4):364–370, 2011.

[43] Jessica L Tracy and Daniel Randles. Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt. *Emotion review*, 3(4):397–405, 2011.

[44] Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *EACL 2017*, page 578, 2017.

[45] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi 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)*, pages 986–995, 2017.

[46] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

[47] Harald G Wallbott and Klaus R Scherer. How universal and specific is emotional experience? evidence from 27 countries on five continents. *Social Science Information*, 25(4), 1986.

[48] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.

[49] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

[50] Roddy Cowie and Randolph R Cornelius. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1-2):5–32, 2003.

[51] Lisa Feldman Barrett and James A Russell. The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science*, 8(1):10–14, 1999.

[52] J.A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110:145, 2003.

[53] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[54] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[55] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[56] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell. On the importance of both dimensional and discrete models of emotion. *Behavioral sciences*, 7(4): 66, 2017.

[57] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[58] American Psychological Association. Mood, n.d. URL https://dictionary.apa.org/mood. Accessed: 2024-12-30.

[59] Zhuochao Peng, Pieter MA Desmet, and Haian Xue. Mood in experience design: A scoping review. *She Ji: The Journal of Design, Economics, and Innovation*, 9(3): 330–378, 2023.

[60] RC Kessler, P Berglund, O Demler, et al. Mood disorders: Bipolar and major depressive disorders. *Jama*, 289(23):3095–105, 2003.

[61] Muk Yan Wong. The mood-emotion loop. *Philosophical Studies*, 173:3061–3080, 2016.

[62] W Michael Vanderlind, Yael Millgram, Arielle R Baskin-Sommers, Margaret S Clark, and Jutta Joormann. Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation. *Clinical psychology review*, 76:101826, 2020.

[63] B. H Morris, L. M. Bylsma, and J. Rottenberg. Does emotion predict the course of major depressive disorder? a review of prospective studies. *British Journal of Clinical Psychology*, 48(3):255–273, 2009.

[64] Sheri L Johnson, June Gruber, and Lori R Eisner. Emotion and bipolar disorder. 2007.

[65] Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, and Emily Mower Provost. The priori emotion dataset: Linking mood to emotion detected in-the-wild. In *Interspeech*, 2018.

[66] Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. Multimodal markers of irony and sarcasm. 2003.

[67] Maedeh Kazemitabar, Susanne P Lajoie, and Tenzin Doleck. Analysis of emotion regulation using posture, voice, and attention: a qualitative case study. *Computers and Education Open*, 2:100030, 2021.

[68] Maria Spinelli, Mirco Fasolo, Gabrielle Coppola, and Tiziana Aureli. It is a matter of how you say it: Verbal content and prosody matching as an index of emotion regulation strategies during the adult attachment interview. *International Journal of Psychology*, 54(1):102–107, 2019.

[69] Michael Cronquist Christensen, Hongye Ren, and Andrea Fagiolini. Emotional blunting in patients with depression. part ii: relationship with functioning, well-being, and quality of life. *Annals of General Psychiatry*, 21(1):20, 2022.

[70] H. Berenbaum and T.F. Oltmanns. Emotional experience and expression in schizophrenia and depression. *Journal of abnormal psychology*, 101(1):37, 1992.

[71] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, 2019.

[72] Emily Öhman. Challenges in annotation: Annotator experiences from a crowdsourced emotion annotation task. In *DHN*, pages 293–301, 2020.

[73] Zoraida Callejas and Ramon Lopez-Cozar. Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50(5):416–433, 2008.

[74] Jeffrey A Brooks, Holly Shablack, Maria Gendron, Ajay B Satpute, Michael H Parrish, and Kristen A Lindquist. The role of language in the experience and perception of emotion: A neuroimaging meta-analysis. *Social cognitive and affective neuroscience*, 12(2):169–183, 2017.

[75] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology*, 70:153–163, 2017.

[76] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[77] Stephan Hamann and Turhan Canli. Individual differences in emotion processing. *Current opinion in neurobiology*, 14(2):233–238, 2004.

[78] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE, 2018.

[79] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[80] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv [cs.LG]*, July 2018.

[81] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv [cs.CL]*, April 2019.

[82] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33: 12449–12460, 2020.

[83] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[84] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-Tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-Yi Lee. SUPERB: Speech processing universal PERformance benchmark. *arXiv [cs.CL]*, May 2021.

[85] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. *arXiv [eess.AS]*, May 2023.

[86] Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP*, pages 1–5. IEEE, 2023.

[87] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL*, pages 3881–3906, 2024.

[88] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Bias in emotion recognition with ChatGPT. 2023.

[89] Sally Olderbak and Oliver Wilhelm. Emotion perception and empathy: An individual differences test of relations. *Emotion*, 17(7):1092, 2017.

[90] Rod A Martin, Glen E Berry, Tobi Dobranski, Marilyn Horne, and Philip G Dodgson. Emotion perception threshold: Individual differences in emotional sensitivity. *Journal of Research in Personality*, 30(2):290–305, 1996.

[91] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 973–987, 2022.

[92] World Health Organization et al. World mental health report: transforming mental health for all: executive summary. In *World mental health report: transforming mental health for all: executive summary*. 2022.

[93] M.R. Morales and R. Levitan. Speech vs. text: A comparative analysis of features for depression detection systems. In *IEEE spoken language technology workshop (SLT)*, pages 136–143. IEEE, 2016.

[94] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67, 2021.

[95] Y.R. Tausczik and J.W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54, 2010.

[96] O. Flanagan, A. Chan, P. Roop, F. Sundram, et al. Using acoustic speech patterns from smartphones to investigate mood disorders: Scoping review. *JMIR mHealth and uHealth*, 9(9):e24352, 2021.

[97] S. Shiffman, A.A. Stone, and M.R. Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[98] J. Gideon, H.T. Schatten, M.G. McInnis, and E. Mower Provost. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In *Interspeech*, 2019.

[99] N.C. Andreasen. The scale for the assessment of negative symptoms (sans): conceptual and theoretical foundations. *The British journal of psychiatry*, 155(S7):49–52, 1989.

[100] K. M'Bailara, T. Atzeni, F. Colom, J. Swendsen, S. Gard, A. Desage, and C. Henry. Emotional hyperreactivity as a core dimension of manic and mixed states. *Psychiatry research*, 197(3):227–230, 2012.

[101] Z.N. Karam, E. Mower Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M.G. Mcinnis. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *IEEE international conference on acoustics, speech and signal processing*, pages 4858–4862, 2014.

[102] Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

[103] R.C. Young, J.T. Biggs, V.E. Ziegler, and D.A. Meyer. A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry*, 133, 1978.

[104] S.L. McElroy, B.E. Martens, R.S Creech, J.A. Welge, L. Jefferson, A.I. Guerdjikova, and P.E. Keck Jr. Randomized, double-blind, placebo-controlled study of divalproex extended release loading monotherapy in ambulatory bipolar spectrum disorder patients with moderate-to-severe hypomania or mild mania. *The Journal of clinical psychiatry*, 71(5):16734, 2010.

[105] M. Zimmerman, J.S. Martinez, D. Young, I. Chelminski, and K. Dalrymple. Severity classification on the hamilton depression rating scale. *Journal of affective disorders*, 150(2):384–388, 2013.

[106] K. Matton, M.G. McInnis, and E. Mower Provost. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. In *Interspeech*, 2019.

[107] S Jagabathula, L Subramanian, and A Venkataraman. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 18(1):3233–3299, 2017.

[108] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[109] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[110] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B.W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *arXiv:2203.07378*, 2022.

[111] M. Perez, M. Jaiswal, M. Niu, C. Gorrostieta, M. Roddy, K. Taylor, R. Lotfian, J. Kane, and E. Mower Provost. Mind the gap: On the value of silence representations to lexical-based speech emotion recognition. *Interspeech*, pages 156–160, 2022.

[112] John Gideon, Emily Mower Provost, and Melvin McInnis. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2359–2363. IEEE, 2016.

[113] Abdelrahman Mohamed, Hung-Yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N Sainath, and Shinji Watanabe. Self-supervised speech representation learning: A review. *IEEE J. Sel. Top. Signal Process.*, 16(6):1179–1210, October 2022.

[114] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic evaluation of audio representations. *arXiv [cs.SD]*, March 2022.

[115] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[116] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921, December 2021.

[117] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using Wav2vec 2.0 embeddings. *arXiv [cs.SD]*, April 2021.

[118] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[119] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333. IEEE, 2018.

[120] Zhaoxi Mu, Xinyu Yang, Sining Sun, and Qing Yang. Self-supervised disentangled representation learning for robust target speech extraction. *AAAI*, 38(17):18815–18823, March 2024.

[121] Yuying Xie, Thomas Arildsen, and Zheng-Hua Tan. Disentangled speech representation learning based on factorized hierarchical variational autoencoder with self-supervised objective. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2021.

[122] SiCheng Yang, Methawee Tantrawenith, Haolin Zhuang, Zhiyong Wu, Aolan Sun, Jianzong Wang, Ning Cheng, Huaizhen Tang, Xintao Zhao, Jie Wang, et al. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. *arXiv preprint arXiv:2208.08757*, 2022.

[123] Jiahao Xun, Shengyu Zhang, Yanting Yang, Jieming Zhu, Liqun Deng, Zhou Zhao, Zhenhua Dong, Ruiqi Li, Lichao Zhang, and Fei Wu. Discover: Disentangled music representation learning for cover song identification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 453–463, 2023.

[124] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[125] Jo-Anne Bachorowski and Michael J Owren. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological science*, 6(4):219–224, 1995.

[126] Sandeep Kumar Pandey, Hanumant Singh Shekhawat, and SR Mahadeva Prasanna. Deep learning techniques for speech emotion recognition: A review. In *2019 29th international conference RADIOELEKTRONIKA (RADIOELEKTRONIKA)*, pages 1–6. IEEE, 2019.

[127] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[128] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, June 2006.

[129] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[130] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27:600–604, 2020.

[131] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv [eess.AS]*, October 2022.

[132] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, 2015.

[133] Alec Burmania, Mohammed Abdelwahab, and Carlos 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)*, pages 5190–5194. IEEE, 2016.

[134] Giannis Haralabopoulos, Myron Tsikandilakis, Mercedes Torres Torres, and Derek McAuley. Objective assessment of subjective tasks in crowdsourcing applications. In *LREC 2020 Workshop on" Citizen Linguistics in Language Resource Development"*, 2020.

[135] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[136] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv:2212.10403*, 2022.

[137] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada, July 2023. Association for Computational Linguistics.

[138] Tom Brown et al. Language models are Few-Shot learners. May 2020.

[139] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are Zero-Shot learners. September 2021.

[140] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is ChatGPT equipped with emotional dialogue capabilities? April 2023.

[141] Jen-Tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Emotionally numb or empathetic? evaluating how LLMs feel using EmotionBench. August 2023.

[142] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, January 2023.

[143] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia M C Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. 2024.

[144] Shutong Feng, Guangzhi Sun, Nurul Lubis, Chao Zhang, and Milica Gašić. Affect recognition in conversations using large language models. September 2023.

[145] Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. Can large language models aid in annotating speech emotional data? uncovering new frontiers. July 2023.

[146] Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Bjorn W Schuller. Refashioning emotion recognition modelling: The advent of generalised large models. August 2023.

[147] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv:1912.00741*, 2019.

[148] Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of EACL*, pages 578–585, Valencia, Spain, April 2017.

[149] Tiantian Feng and Shrikanth Narayanan. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. September 2023.

[150] Sven Buechel and Udo Hahn. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th linguistic annotation workshop*, pages 1–12, 2017.

[151] Manikanta Loya, Divya Anand Sinha, and Richard Futrell. Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variation and hyperparameters. *arXiv:2312.17476*, 2023.

[152] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):1–35, January 2023.

[153] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U. S. A.*, 120(6):e2218523120, February 2023.

[154] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[155] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong 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)*, pages 11173–11195, July 2023.

[156] Santiago González-Carvajal and Eduardo C Garrido-Merchán. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

[157] Henry Oluchukwu Ikediego, Mustafa Ilkan, A Mohammed Abubakar, and Festus Victor Bekun. Crowd-sourcing (who, why and what). *International Journal of Crowd Science*, 2(1):27–41, 2018.

[158] Angeliki Metallinou and Shrikanth 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.

[159] Ian D Wood, John P McCrae, Vladimir Andryushechkin, and Paul Buitelaar. A comparison of emotion annotation approaches for text. *Information*, 9(5):117, 2018.

[160] Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. From text to emotion: Unveiling the emotion annotation capabilities of llms. In *Proc. Interspeech*, pages 2650–2654, 2024.

[161] Lea Canales, University of Alicante, Alicante, Spain, Walter Daelemans, Ester Boldrini, and Patricio Martínez-Barco. Towards the improvement of automatic emotion preannotation with polarity and subjective information. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*. Incoma Ltd. Shoumen, Bulgaria, November 2017.

[162] Lea Canales, Walter Daelemans, Ester Boldrini, and Patricio Martinez-Barco. EmoLabel: Semi-automatic methodology for emotion annotation of social media text. *IEEE Trans. Affect. Comput.*, 13(2):579–591, April 2022.

[163] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, 2022.

[164] Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, 25(2):333–358, 2015.

[165] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. 2015.

[166] Isil Yakut Kilic and Shimei 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*, pages 4532–4539, 2022.

[167] Rajwa Alharthi and Abdulmotaleb El Saddik. A multi-layered psychological-based reference model for citizen need assessment using ai-powered models. *SN Computer Science*, 1(5):291, 2020.

[168] Héctor Ricardo Murrieta Bello, Lukas Heilmann, and Esben Ronan. Detecting fake news spreaders with behavioural, lexical and psycholinguistic features. In *CLEF (Working Notes)*, 2020.

[169] Priya Ranganathan, CS Pramesh, and Rakesh Aggarwal. Common pitfalls in statistical analysis: logistic regression. *Perspectives in clinical research*, 8(3):148–151, 2017.

[170] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*, 2023.

[171] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David 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*, pages 11370–11403, 2023.

[172] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

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

[174] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11 (2):37–50, 1912.

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

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

[177] Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[178] Sven Buechel and Udo Hahn. Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016*, pages 1114–1122. IOS Press, 2016.

[179] Sonja Biersack and Vera 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.

[180] Carlos Busso and Shrikanth S Narayanan. The expression and perception of emotions: comparing assessments of self versus others. In *Interspeech*, pages 257–260, 2008.

[181] Doo Yon Kim and Christian Wallraven. Label quality in affectnet: results of crowd-based re-annotation. In *Asian Conference on Pattern Recognition*, pages 518–531. Springer, 2021.

[182] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6): e2218523120, 2023.

[183] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[184] Tara M Chaplin and Amelia Aldao. Gender differences in emotion expression in children: a meta-analytic review. *Psychological bulletin*, 139(4):735, 2013.

[185] Seungyoun Kim, Jennifer L Geren, and Bob G Knight. Age differences in the complexity of emotion perception. *Experimental Aging Research*, 41(5):556–571, 2015.

[186] David Matsumoto. Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an american sample. *Motivation and emotion*, 17(2):107–123, 1993.

[187] MGJ Swerts and EJ Krahmer. Gender-related differences in the production and perception of emotion. In *Proceedings of the international conference on spoken language processing (Interspeech 2008)*, pages 334–337. ISCA, 2008.

[188] Yi Lin, Hongwei Ding, and Yang Zhang. Unisensory and multisensory stroop effects modulate gender differences in verbal and nonverbal emotion perception. *Journal of Speech, Language, and Hearing Research*, 64(11):4439–4457, 2021.

[189] Michaela Riediger, Manuel C Voelkle, Natalie C Ebner, and Ulman Lindenberger. Beyond "happy, angry, or sad?": Age-of-poser and age-of-rater effects on multidimensional emotion perception. *Cognition & emotion*, 25(6):968–982, 2011.

[190] Cassie Mogilner, Sepandar D Kamvar, and Jennifer Aaker. The shifting meaning of happiness. *Social Psychological and Personality Science*, 2(4):395–402, 2011.

[191] David Matsumoto, Jeff LeRoux, Carinda Wilson-Cohn, Jake Raroque, Kristie Kooken, Paul Ekman, Nathan Yrizarry, Sherry Loewinger, Hideko Uchida, Albert Yee, et al. A new test to measure emotion recognition ability: Matsumoto and ekman's japanese and caucasian brief affect recognition test (jacbart). *Journal of Nonverbal behavior*, 24:179–209, 2000.

[192] Irene Trilla, Anne Weigand, and Isabel Dziobek. Affective states influence emotion perception: evidence for emotional egocentricity. *Psychological research*, 85(3):1005–1015, 2021.

[193] Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6(1):26, 2023.

[194] Woan-Shiuan Chien and Chi-Chun Lee. Achieving fair speech emotion recognition via perceptual fairness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[195] Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. Impact of annotator demographics on sentiment dataset labeling. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2):1–22, November 2022.

[196] Biqiao Zhang, Yuqing Kong, Georg Essl, and Emily Mower Provost. f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5725–5732, 2019.

[197] Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, and Yushi Aono. Speech emotion recognition based on multi-label emotion existence model. In *INTERSPEECH*, pages 2818–2822, 2019.

[198] Navin Raj Prabhu, Nale Lehmann-Willenbrock, and Timo Gerkmann. End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning. *IEEE Transactions on Affective Computing*, 15 (2):579–592, 2023.

[199] Wen Wu, Chao Zhang, Xixin Wu, and Philip C Woodland. Estimating the uncertainty in emotion class labels with utterance-specific dirichlet priors. *IEEE Transactions on Affective Computing*, 14(4):2810–2822, 2022.

[200] Jun Deng, Wenjing Han, and Björn Schuller. Confidence measures for speech emotion recognition: A start. In *Speech Communication; 10. ITG Symposium*, pages 1–4. VDE, 2012.

[201] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897, 2017.

[202] James Tavernor, Yara El-Tawil, and Emily Mower Provost. The whole is bigger than the sum of its parts: Modeling individual annotators to capture emotional variability. In *Proc. Interspeech 2024*, pages 3195–3199, 2024.

[203] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 international joint conference on neural networks (IJCNN)*, pages 566–570. IEEE, 2016.

[204] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[205] Minh Tran, Yufeng Yin, and Mohammad Soleymani. Personalized adaptation with pre-trained speech encoders for continuous emotion recognition. *Proc. INTERSPEECH 2023*, 2023.

[206] Tsungcheng Yao, Ernest Foo, and Sebastian Binnewies. Personalised abusive language detection using llms and retrieval-augmented generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 92–98, 2024.

[207] OP John. The big five inventory-versions 4a and 54. *University of California, Berkeley, Institute of Personality and Social Research*, 1991.

[208] Jiaxin Pei and David Jurgens. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, 2023.

[209] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *J. Res. Pers.*, 41(1):203–212, 2007.

[210] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

[211] Ingo Siegert, Ronald Böck, and Andreas Wendemuth. Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *Journal on Multimodal User Interfaces*, 8:17–28, 2014.

[212] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

[213] Srinivas Parthasarathy and Carlos Busso. Jointly predicting arousal, valence and dominance with multi-task learning. In *Interspeech*, volume 2017, pages 1103–1107, 2017.

[214] Hillary Anger Elfenbein and Nalini Ambady. Is there an in-group advantage in emotion recognition? 2002.

[215] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601, 2018.

[216] Mirosław Płaza, Robert Kazała, Zbigniew Koruba, Marcin Kozłowski, Małgorzata Lucińska, Kamil Sitek, and Jarosław Spyrka. Emotion recognition method for call/contact centre systems. *Applied Sciences*, 12(21):10951, 2022.

[217] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005.

[218] Yüksel Yurtay, Hüseyin Demirci, Hüseyin Tiryaki, and Tekin Altun. Emotion recognition on call center voice data. *Applied Sciences*, 14(20):9458, 2024.

[219] Abdur Rasool, Saba Aslam, Naeem Hussain, Sharjeel Imtiaz, and Waqar Riaz. nbert: Harnessing nlp for emotion recognition in psychotherapy to transform mental health care. *Information*, 16(4):301, 2025.

[220] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, 2019.

[221] Manh Tu Vu, Marie Beurton-Aimar, and Serge Marchand. Multitask multi-database emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3637–3644, 2021.

[222] Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496, 2024.

[223] Katie Hoemann, Evan Warfel, Caitlin Mills, Laura Allen, Peter Kuppens, and Jolie B Wormwood. Using freely generated labels instead of rating scales to assess emotion in everyday life. *Assessment*, page 10731911241283623, 2024.

[224] Eimear Stanley, Eric DeMattos, Anita Klementiev, Piotr Ozimek, Georgia Clarke, Michael Berger, and Dimitri Palaz. Emotion label encoding using word embeddings for speech emotion recognition. In *Proceedings of the INTERSPEECH*, pages 2418–2422, 2023.

[225] Ankita Bhaumik and Tomek Strzalkowski. Towards a generative approach for emotion detection and reasoning. *arXiv preprint arXiv:2408.04906*, 2024.

[226] Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8):1–35, 2025.

[227] Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. Challenges and opportunities of text-based emotion detection: a survey. *IEEE access*, 12:18416–18450, 2024.

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

[229] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[230] Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. Fofo: A benchmark to evaluate llms' format-following capability. In *62nd Annual Meeting of the Association for Computational Linguistics*, pages 680–699, 2024.

[231] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[232] Justin Olah, Sabyasachee Baruah, Digbalay Bose, and Shrikanth Narayanan. Cross domain emotion recognition using few shot knowledge transfer. *arXiv [cs.CL]*, October 2021.

[233] Flor Miriam Plaza Del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, 2022.

[234] Patrick Bareiß, Roman Klinger, and Jeremy Barnes. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1318–1326, 2024.

[235] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[236] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[237] Senait Gebremichael Tesfagergish, Jurgita Kapočiūtė-Dzikienė, and Robertas Damaševičius. Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*, 12(17):8662, 2022.

[238] Sven Buechel, Luise Modersohn, and Udo Hahn. Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249, 2021.

[239] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3914–3923. Association for Computational Linguistics, 2019.

[240] Georgios Chochlakis, Gireesh Mahajan, Sabyasachee Baruah, Keith Burghardt, Kristina Lerman, and Shrikanth Narayanan. Using emotion embeddings to transfer knowledge between emotions, languages, and annotation formats. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[241] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[242] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[243] Desmond C Ong. An ethical framework for guiding the development of affectively-aware artificial intelligence. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[244] Kat Roemmich and Nazanin Andalibi. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.

[245] Agneta H Fischer, Mariska E Kret, and Joost Broekens. Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis. *PloS one*, 13(1):e0190712, 2018.

[246] Ted Ruffman, Julie D Henry, Vicki Livingstone, and Louise H Phillips. A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews*, 32(4):863–881, 2008.

[247] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*, 2024.