Can Third Parties Read Our Emotions?

Jiayi Li, Yingfan Zhou, Pranav Narayanan Venkit, Halima Binte Islam, Sneha Arya
Shomir Wilson, Sarah Rajtmajer
Pennsylvania State University
{jpl6207, yxz5975, pranav.venkit, hpi5039, sfa5848, shomir, smr48}@psu.edu

Abstract

Natural Language Processing tasks that aim to infer an author's private states, e.g., emotions and opinions, from their written text, typically rely on datasets annotated by third-party annotators. However, the assumption that third-party annotators can accurately capture authors' private states remains largely unexamined. In this study, we present human subjects experiments on emotion recognition tasks that directly compare third-party annotations with first-party (author-provided) emotion labels. Our findings reveal significant limitations in third-party annotations—whether provided by human annotators or large language models (LLMs)—in faithfully representing authors' private states. However, LLMs outperform human annotators nearly across the board. We further explore methods to improve third-party annotation quality. We find that demographic similarity between first-party authors and third-party human annotators enhances annotation performance, while incorporating first-party demographic information into prompts leads to a marginal but statistically significant improvement in LLMs' performance. We introduce a framework for evaluating the limitations of third-party annotations and call for refined annotation practices to accurately represent and model authors' private states.

1 Introduction

Recognizing and interpreting subjective language used to express *private states*–users' internal experiences, e.g., opinions, emotions, evaluations, and speculations (Quirk et al., 1985)–has been a long-standing area of study in natural language processing (NLP) (Wiebe et al., 2004; Banfield, 1982). Examples of private state inference tasks include emotion classification (Mohammad et al., 2018; Demszky et al., 2020), emotion intensity detection (Mohammad and Bravo-Marquez, 2017), sarcasm detection (Bamman and Smith, 2015; Oprea and Magdy, 2019), sentiment analysis (Nemes and Kiss, 2021), political ideology detection (Iyyer et al., 2014), stance detection (AlDayel and Magdy, 2021), and many more. Approaches for these tasks have relied heavily on 'gold standard' datasets annotated by third-party annotators. The standard annotation process typically involves multiple human annotators, often recruited through crowdsourcing platforms or drawn from internal research teams (Rashtchian et al., 2010; Snow et al., 2008). More recently, large language models (LLMs) have been explored as scalable and cost-effective alternatives or complements to human annotators (Li et al., 2023). Studies have shown that LLMs can match or surpass human annotators on some annotation tasks (Hasanain et al., 2024; Gilardi et al., 2023).

While the adoption of third-party annotations is appropriate for many NLP tasks such as those with objective ground truth (e.g., named entity recognition (Nadeau and Sekine, 2009), part-of-speech tagging (Brill, 1994)) or for tasks that benefit from diverse third-party perception (e.g., toxicity detection (Pavlopoulos et al., 2020), hate speech detection (Davidson et al., 2017)), ***we suggest that third-party annotations (human or machine) have inherent limitations for tasks that seek to model an author's private state***. Fundamentally, the use of third-party annotations assumes that an author's private state can be identified from their writing by human or machine annotators. However, subjective language often lacks explicit linguistic cues (Wiebe et al., 2005; Balahur et al., 2012), leaving annotators to perform inference on textual cues, which may be implicit, ambiguous, or context-dependent. This challenge is compounded by individual differences in authors' expression of private states (DeAndrea et al., 2010; Bauer et al., 2003) as well as annotators' socio-demographic, cultural backgrounds, and personal beliefs which shape how they interpret authors' text (Shen and Rose, 2021; Ding et al., 2022; Oprea and Magdy, 2019).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21478–21499

Misalignment between an author's private state and its interpretation by third-party annotators is not merely a labeling error—it can propagate through learned models and compromise the reliability of downstream applications. However, little research has systematically examined this gap.

To this end, we conduct a series of studies to investigate the alignment between third-party annotations and first-party labels, i.e., authors' self-reported private states, in the context of emotion recognition tasks. Emotion recognition has a wide range of applications, especially in high-stakes contexts such as moderating online content, detecting deception, and powering therapeutic chatbots (Shaw and Lyons, 2017; Chiril et al., 2022; Zygadło et al., 2021). In these contexts, misinterpretations of users' emotions not only render the technology deficient but also make it socially pernicious, highlighting the critical need for ethical considerations in this domain (Mohammad, 2022). Moreover, emotion recognition is closely related to other tasks that aim to infer private states, such as mental health detection and sentiment analysis (Zhang et al., 2023; Venkit et al., 2023) so that we anticipate lessons learned in this domain will inform the next steps in others.

Our work is guided by the research questions:

RQ1: How do third-party human annotators and LLMs align with first-party labels, i.e., authors' self-reports, on emotion recognition tasks?

RQ2: Does demographic similarity between human annotators and authors improve alignment between third-party annotations and first-party labels?

RQ3: Does including authors' demographic information within prompts improve LLM annotations?

To explore these questions, we recruit social media users to share their own social media posts and label them with their own emotion labels (first-party labels). We then collect annotations of these same posts from third-party human annotators and LLMs and compare their performance. We investigate the impact of demographic similarity between third-party human annotators and first-party authors; similarly, we explore whether including first-party demographic information within prompts improves LLM annotations.

Our results reveal notable misalignment between first-party labels and third-party annotations, as both human annotators and LLMs achieve only low to fair Cohen's kappa values and F1 scores across different emotions. LLMs generally perform better than human annotators **(RQ1)**. In-group human annotators' performance is significantly better than out-group annotators, suggesting that demographic similarity between third-party human annotators and first-party individuals improves annotation performance (**RQ2**). We further observe that adding first-party information in prompts marginally improves LLMs' annotations (**RQ3**).

2 Related Work

2.1 Private State Annotations for NLP

Emotions, sentiment, opinions, evaluations, and speculations are fundamental aspects of an individual's internal world, collectively referred to as private states (Quirk et al., 1985). Many NLP tasks aim to identify an author's private states from their texts. Labeled datasets are needed to train and test models to perform these tasks.

Human-annotated datasets. For emotion recognition, Mohammad et al. (2018) introduce a set of subtasks, such as emotion classification and emotion intensity classification, that aim to infer the affective state of a person from their tweets and provide a human-annotated dataset for each of the subtasks. Aman and Szpakowicz (2007) introduce a corpus that consists of blog posts with human annotations of emotion at the sentence level. Demszky et al. (2020) build a dataset of Reddit posts labeled by third-party human annotators based on a fine-grained emotion taxonomy. These datasets have been widely used for model training and evaluation.

LLM-annotated datasets. Given LLMs' demonstrated zero-shot and few-shot capabilities for various NLP tasks (Brown et al., 2020), researchers have increasingly explored their use to augment or replace human annotators (Gilardi et al., 2023; Kim et al., 2024). Studies have highlighted the potential of LLMs as annotators even for subjective tasks that model authors' private states, such as emotion recognition (Niu et al., 2024) and stance detection (Li and Conrad, 2024).

Limitations of third-party annotations. Limited research has systematically investigated the limitations of third-party annotations for private states. Oprea and Magdy (2019) examined the differences between intended sarcasm (as labeled by the author) and perceived sarcasm (as labeled by third-party annotators). While sarcasm detection can benefit from explicit linguistic cues, such as irony and exaggeration, understanding private states often relies on implicit signals and is subject to individual interpretation. Another relevant study Joseph et al. (2021) explored the discrepancy between the publicly expressed stance, as inferred by third-party from social media posts, and the stance measured by public opinion polls.

2.2 Emotion Recognition

Emotion recognition is a long-studied task in NLP that aims to identify the emotions expressed by an author based on their written content (Alswaidan and Menai, 2020). While emotion recognition can be framed in multiple ways, our study specifically focuses on the task of emotion classification. Several third-party-annotated datasets have been developed for emotion classification across different domains and applications (Liu et al., 2019; Li et al., 2017; Buechel and Hahn, 2017). Among these, datasets such as those developed by Mohammad et al. (2018) and Demszky et al. (2020) are specifically based on social media posts.

Traditional approaches to emotion classification often rely on coarse-grained taxonomies such as Ekman's six basic emotions (Ekman, 1992) or Plutchik's wheel of emotions (Plutchik, 2001). More recent work in NLP has increasingly embraced fine-grained emotion taxonomies to better capture the nuances of emotional expression (Liew et al., 2016; Abdul-Mageed and Ungar, 2017). For example, Demszky et al. (2020) developed a taxonomy comprising 27 emotions plus a neutral category, designed to minimize overlap while preserving sufficient granularity to capture the diversity of emotions expressed in text.

Prior psychological studies have shown that emotion expression and interpretation can be influenced by cultural and social factors (Lim, 2016; Gendron et al., 2014; Mill et al., 2009; Barrett, 2004). Individuals from the same demographic background—such as age, gender, or cultural identity—tend to exhibit greater alignment in emotion expression and interpretation (Elfenbein and Ambady, 2002a,b; Sauter et al., 2010). Building on these findings, our study investigates whether annotators who share demographic traits with the author perform better on emotion recognition tasks than those who do not.

2.3 Incorporating Social Factors in LLMs

The importance of modeling and incorporating social factors in NLP tasks has been increasingly emphasized. Hovy and Yang (2021) have argued that understanding factors such as the speaker's and receiver's background information will ultimately be necessary if NLP models are to ever achieve human-level performance. Several studies have examined LLMs' ability to capture and align with the communication styles, perspectives, and preferences of specific demographic groups or even individuals (Hwang et al., 2023; Mukherjee et al., 2024). Beck et al. (2024) examined socio-demographic prompting, where socio-demographic information is incorporated into prompts to guide an LLM into adopting the perspective of a certain group or user, and found potential in this technique. In contrast, some studies indicate that demographic-infused prompts may not only fail to mitigate gender and racial biases in LLMs' annotations (Sun et al., 2023) but also reveal gendered stereotypes in emotion attribution (Plaza-del Arco et al., 2024).

3 Methodology

To investigate the alignment between first-party labels and third-party annotations in emotion recognition, we adopt the emotion recognition task from (Demszky et al., 2020) and conduct human subject experiments. Data collection occurred in two stages: (1) collection of first-party social media posts with self-reported emotion labels; and (2) collection of third-party annotations for these posts.

We analyze annotation performance using Cohen's kappa, F1 score, recall, and precision. We apply statistical analyses, including mixed linear models and the Wilcoxon signed-rank test, to further assess the impact of demographic similarity on third-party annotation performance. Additionally, we analyze patterns of misalignment to understand the role of linguistic cues in emotion recognition. All human subjects research was approved by the university's Institutional Review Board.

3.1 Emotion Taxonomy

We adopt the fine-grained emotion taxonomy introduced by Demszky et al. (2020), which includes 27 distinct emotions plus "neutral". The taxonomy provides definitions for each emotion, along with an associated emoji where applicable. The full taxonomy is provided in Appendix A. To evaluate the alignment between third-party annotations and first-party labels at a coarser level, we group the 28 emotion categories into seven groups: joy, love, anger, surprise, fear, sadness, and neutral. As there is no universally agreed-upon set of basic emotions, our grouping is based on the union of basic emotion categories proposed by Ekman (1992) and Shaver et al. (1987). The detailed grouping and analysis based on the aggregated emotion taxonomy are presented in Appendix H.

3.2 First-Party Data Collection

We recruited social media users through Connect[1], a crowdsourcing platform designed to facilitate research participation by connecting researchers with diverse and high-quality participant pools (Douglas et al., 2023; Eyal et al., 2021). We recruited participants in the United States from three age groups (18–27, 28–43, and 44–59), two gender categories (female and male), and three racial groups (Black, White, and Asian). We attempted to include additional groups, e.g., non-binary and Hispanic individuals, but they had insufficient representation on Connect. Our intersectional recruitment strategy aimed to achieve balanced representation across demographics. Specifically, we aimed to obtain approximately 10 responses for each combination of age, gender, and race. While we successfully reached most groups, some intersectional groups—such as Asian participants aged 44–59—proved challenging due to their underrepresentation on the platform. A full breakdown of the number of participants by age, gender, and race is offered in Table 5 of Appendix B.

Via online survey, participants were asked to provide basic demographic information and to name one or two social media platforms where they post most frequently. They were then asked to upload a minimum of 5 (maximum of 15) posts they had authored on the selected social media platform(s) in the past 12 months. Posts were submitted in the form of screenshots. For each uploaded post, participants were asked to review the content and select all emotions they believed were expressed in the post, using the list of 28 emotion categories, along with definitions. To ensure data quality and authenticity, posts were required to adhere to a number of criteria, related to language, multi-media, identifying information, and date of publication. These are detailed in Appendix C. Each participant was compensated $2.50 for completing the task, which took approximately 10–15 minutes.

Table 5 in Appendix B presents the number of participants who provided valid responses and the total number of valid posts for each intersectional group. In total, we collected 729 posts from 123 participants; 44% of these contained only text and the remaining posts contained both text and images. A quality check was manually performed on these posts and their associated labels; details of this process are provided in Appendix D.

[1] https://connect.cloudresearch.com

3.3 Third-Party Emotion Labels

Human annotations. We recruited human annotators through Connect. For each post, we assigned **six annotators**: **three in-group annotators**, who shared all three demographic traits (age group, gender, and racial group) with the author; and, **three out-group annotators**, who differed from the author on at least two of these traits.[2]The demographic distribution of in-group annotators mirrored that of first-party participants by design. We aimed to ensure a diverse and balanced sample of out-group annotators across different age groups, genders, and racial backgrounds. Figure 3 in Appendix B details the demographic distribution of out-group annotators.

Each annotation task consisted of 5–10 randomly assigned posts. Annotators were presented with the original screenshots of posts provided by authors and asked to identify emotions expressed by the author (first-party), using the 27 emotion categories plus "neutral" (if none). The list was accompanied by clear definitions of each category (see Appendix A). Annotators were instructed to follow three guidelines: (1) they could select multiple emotions, but only those they were reasonably confident were expressed; (2) if no emotions were expressed, they were to select "Neutral"; and (3) if they could not confidently assign any label from the list, they were to select "Unrecognizable". Annotators were compensated $2.00 for completing the task, which took approximately 8–10 minutes.

We disregarded annotations that were self-contradictory, such as those that selected both "neutral" and one or more emotions from the list of 27 categories. We interpreted such self-contradictory annotations as indicative of a lack of attention. Furthermore, annotations from annotators who frequently made such errors were removed. Summary statistics of human annotations is presented in Appendix B Table 6.

LLM annotations. To complement human annotations, we used LLMs to generate third-party

[2] As both first-party and third-party human annotators were recruited from the same platform, we ensured that first-party participants did not serve as in-group annotators for their own posts.

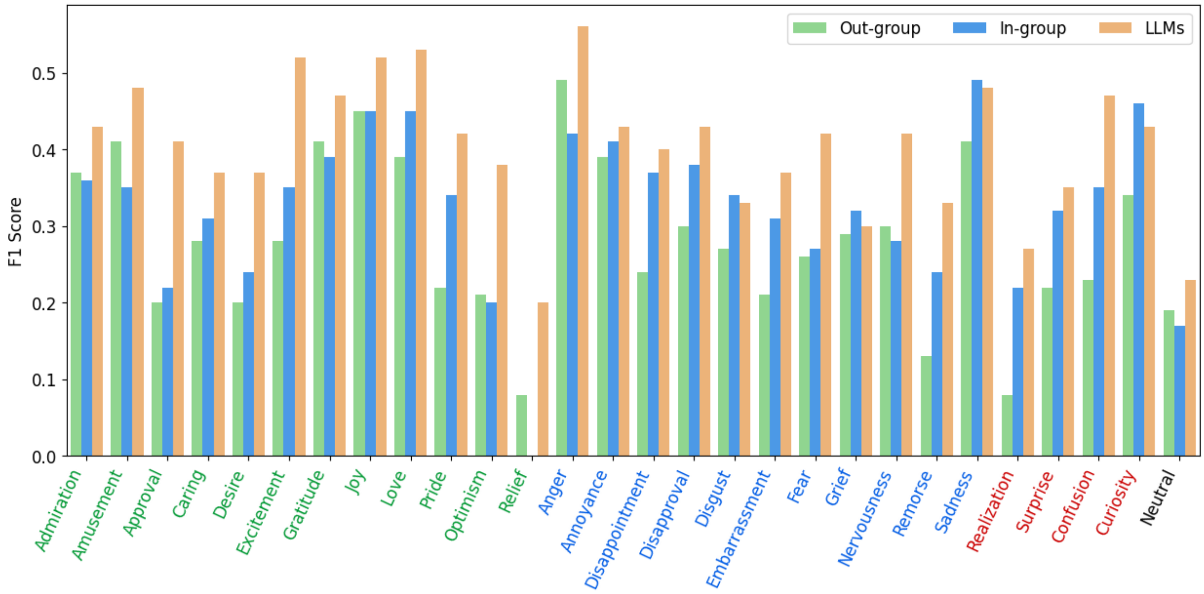

Figure 1: Alignment between third-party annotations with first-party labels, evaluated using F1 score.

emotion labels. Specifically, we used GPT-4 Turbo (OpenAI, 2023), GPT-4o (OpenAI, 2023), Gemini 1.5 Pro (Reid et al., 2024), Gemini 1.5 Flash (Reid et al., 2024), and Claude 3.5 Sonnet (Anthropic, 2024). These models were chosen based on their strong performance on other NLP tasks and their ability to process multimodal content.

Each post was independently annotated by each LLM. LLMs were providedscreenshots of posts submitted by authors, along with instructions defining the 27 emotion categories plus "neutral" and "unrecognizable". The prompt mirrored the instructions given to human annotators to maintain consistency (see Appendix I) .

4 Third-party Annotators' Performance

4.1 Third-party Annotators' Agreement

We estimated interrater agreement amongst human annotators and amongst LLMs, following the method used in (Demszky et al., 2020), i.e., calculating the Spearman correlation between each annotator's judgments and the mean judgments of other annotators, averaged across all posts labeled by the annotator. LLMs generally exhibit higher and more consistent interrater agreement compared to human annotators, but there is some variability in this with respect to positive vs. negative emotions. These results are further detailed in Appendix F.

4.2 First- and Third-Party Alignment (RQ1)

We analyze the alignment between first-party labels and third-party annotations. First-party labels are treated as the gold standard. As mentioned earlier, for each post, we collected 6 annotations from human annotators (3 in-group; 3 out-group) and one annotation from each of 5 LLMs. Each annotation provided by a third-party annotator for a post consists of a binary label for each of the 27 emotions plus "neutral". We employed two sets of metrics: (1) Cohen's kappa and (2) classification metrics: F1, recall, and precision.

To assess alignment, we compute Cohen's kappa for both human and LLM annotations relative to the first-party labels. We aggregate annotations for each post using majority voting—applied separately to human annotators and to LLMs. For human annotations, ties can occur after removal of low-quality annotations. Thus, Cohen's kappa was computed for each emotion based on posts where a majority-voted label for that emotion could be determined. Both human annotators and LLMs exhibit low to fair alignment with first-party labels; their Cohen's kappa scores range from 0 to 0.45. Details are provided in Appendix G.

We further evaluated first- and third-party alignment using F1 score, recall, and precision. The macro-average precision, recall, and F1-score are 0.38, 0.29, and 0.32 for in-group annotators; 0.36, 0.24, and 0.28 for out-group annotators; and 0.38, 0.50, and 0.40 for LLMs. Figure 1 illustrates the F1 scores achieved by human annotators (in-group, out-group) and LLMs across different emotions, where the emotions on the x-axis are color-coded to represent distinct semantic categories: positive, negative, and ambiguous emotions. F1 scores for LLMs range from 0.2 to 0.6, while the F1 scores fall between 0.1 and 0.5 for human annotators. Overall, LLMs outperform human annotators across most emotions. However, for grief, sadness, and curiosity, in-group human annotators demonstrate comparable or even superior performance. Realization, relief, and neutral exhibit consistently low F1 scores across all annotator types, suggesting that these emotions are particularly challenging to classify. We further examined the correlation between Cohen's kappa and F1 scores for both LLM and human annotations. In both cases, Cohen's kappa scores show a strong positive correlation with F1 scores (LLM: $r = 0.918$, $p < 0.001$; human: $r = 0.855$, $p < 0.001$). We further present confusion matrices showing the alignment between third-party annotations and first-party labels in Appendix G, along with detailed analyses.

To evaluate the alignment between third-party annotations and first-party labels at a coarser level, we replicated this analysis using the aggregated emotion taxonomy (see Appendix H.2). Third-party annotations from both human annotators and LLMs continued to exhibit significant misalignment with first-party labels even at this aggregated level. This suggests that the misalignment is not merely a result of difficulty distinguishing between fine-grained emotion categories, but instead reflects a more fundamental challenge: the inherent difficulty of any third-party attempt to access and accurately interpret another individual's internal emotional state.

4.3 In- vs. Out-Group Annotators (RQ2)

To address RQ2, we compare the performance of in- and out-group annotators at both the post and annotator levels.

Post-level comparison. Among 91% of the posts, both groups reached a majority decision for all 28 emotions. For each post, we computed F1, recall, and precision by comparing the majority-voted labels to the corresponding first-party labels for in-group and out-group. We also used Cohen's kappa to assess in-group and out-group annotation performance. For each emotion, we computed Cohen's kappa based on posts where the majority of annotators in both groups reached a decision on that emotion. On average, 99% of posts were included for this calculation.

We performed Wilcoxon signed-rank tests to assess the statistical significance of differences between in-group and out-group performance measured by F1, recall, precision, and Cohen's kappa. Table 1 presents the results of Wilcoxon signed-rank tests. Results indicate that the observed differences in performance of in-group and out-group annotations, measured based on F1, recall, precision, and Cohen's kappa, are statistically significant, allowing us to conclude in-group annotators outperform out-group annotators in recognizing the emotions expressed by first-party authors. Further details are provided in Figure 15 and Figure 16 in the Appendix.

Metric	In-group Median	Out-group Median	P-value
F1 Score	0.29	0.00	0.004*
Recall	0.25	0.00	0.001*
Precision	0.25	0.00	0.050
Cohen's Kappa	0.28	0.24	0.028*

Table 1: Comparison of performance between in- and out-group annotators at the post-level.

Metric	In-group Mean	Out-group Mean	P-value
Precision	0.32	0.30	0.052
F1 Score	0.30	0.28	0.023*
Recall	0.39	0.35	0.011*

Table 2: Comparison of performance between in- and out-group annotators at the annotator-level.

Annotator-level comparison. While post-level analysis provides insights into aggregated group performance, an annotator-level evaluation helps identify individual alignment with first-party labels. This approach accounts for individual annotator tendencies, which may not be apparent when only considering majority labels.

To compare in-group and out-group performance at the annotator-level, we preserved individual third-party annotations. For each annotation, we obtained F1, recall, and precision by comparing it to the corresponding first-party labels. We averaged the F1 score, precision, and recall across annotations provided within each annotation task for each annotator.

Figure 15 in the Appendix presents the distribution of average F1 scores, recall, and precision for individual annotators per task, grouped by in-group and out-group membership. Across all three metrics (precision, recall, and F1), the in-group annotators' distributions appear shifted slightly to the

Metric	In-group vs. LLMs			Out-group vs. LLMs		
	In-group Median	LLM Median	p-value	Out-group Median	LLM Median	p-value
F1 Score	0.29	0.40	8.34×10^{-12}***	0.00	0.40	4.62×10^{-25}***
Recall	0.25	0.33	4.95×10^{-31}***	0.00	0.33	2.76×10^{-39}***
Precision	0.25	0.67	0.71	0.00	0.67	3.00×10^{-3}**
Cohen's Kappa	0.28	0.32	2.75×10^{-5}***	0.23	0.32	7.45×10^{-9}***

Table 3: Comparison of in- and out-group annotators with LLMs.

right compared to the out-group annotators. The gap is most pronounced in recall and F1, whereas precision exhibits substantial overlap between the two groups. To formally test whether these differences are statistically significant, we employed linear mixed models to account for the nested structure of the data. Specifically, task ID and annotator ID were treated as random effects to account for variability across tasks and individual annotators.

The results of the mixed linear model, presented in Table 2, indicate that in-group annotators outperform out-group annotators in terms of recall and F1 score. However, the difference in precision between the two groups is not statistically significant. This observed advantage for in-group human annotators was also largely consistent when analyzing annotations grouped into the broader seven-category taxonomy (see H.3 for details). These findings suggest that shared identity traits (age, gender, and race) enhance annotators' ability to align with first-party expressed emotions, particularly in achieving higher recall and overall performance (as measured by F1).

4.4 Human Annotators vs. LLMs

We compare the performance of human annotators (both in-group and out-group) with that of LLMs by analyzing F1, recall, and precision metrics on a per-post basis. We compute performance metrics based on the majority-voted labels for each post for in-group, out-group, and LLMs. Comparison between LLMs and in-group is based on posts where the majority of in-group annotators reached a decision for all 28 emotions (94% of posts). Comparison between LLMs and out-group is based on posts where the majority of out-group annotators reached a decision for all 28 emotions (97% of posts).

For each emotion, we computed Cohen's kappa for in-group annotations using only posts where a majority of in-group annotators reached a decision (on average, 99% of posts). Similarly, Cohen's kappa for out-group annotations was calculated using posts where a majority of out-group annotators reached a decision (nearly all posts). Although LLMs provide decisions for all emotions on every post (since we aggregate responses from 5 LLMs), we restricted LLM evaluations to the same subsets of posts used for human annotators to allow direct comparison. We used the Wilcoxon signed-rank test to assess whether the differences in F1, recall, precision, and Cohen's kappa between LLMs and human annotator groups (in-group and out-group) were statistically significant. The results, in Table 3, suggest that LLMs perform significantly better in terms of F1, recall, and Cohen's kappa. However, their advantage does not extend to precision, as indicated by the lack of statistical significance in the in-group comparison.

4.5 Demographic Prompting (RQ3)

In this section, we examine the impact of incorporating first-party demographic information—specifically age, gender, and race—into the prompt. By explicitly including demographic details of first-party, we investigate whether the model's predictions align more closely with first-party emotion labels.

For each LLM, we generated an annotation for every post using demographic prompting. We applied majority voting across these LLM annotations to derive a majority-voted label for each post. We then compared these majority-voted labels with those obtained without demographic information in the prompt. We adopted the Wilcoxon signed-rank test to evaluate whether incorporating demographic information in the prompt improved alignment between the model's predictions and the first-party labels. The results indicate that demographic prompting results in a statistically significant difference in F1 score ($p = 0.0095$) and precision ($p < 0.001$), while no significant difference is observed in recall ($p = 0.9934$). However, the median scores for the demographically prompted condition (F1 score: 0.4, precision: 0.333, recall: 0.667) were

First-party social media post	First-party annotation	Third-party annotation
I'm always turning train into facts, one way or another	joy, optimism	pride, annoyance
Sometimes u gotta stay busy so you ain't got time to feel much. This is me lately. I feel emotionally numb and I just fight people in PC games when I'm not doing stuff around the house. I don't have many people to speak to other than my therapist as I finally transitioned my attention away from my hometown.	embarrassment, sadness	approval
As an ENFP who has been living with an INFJ for 4 years, the tiny argument likely had a lot more meaning for him than you realize. Give him a little space and try not to force talking about the issue until he initiates it next. Best of luck to you.	neutral	caring, optimism
2024 really has me out here googling shit like "how to attend house party" like I'm a 20 year old ex-homeschool kid (again) "how to introduce yourself to a stranger" "manage social anxiety birthday holiday barbecue no borax no glue" "how to not talk about Space Marines" "what do normal people talk about" "leave houseparty gracefully Irish goodbye French exit" "small talk 101 for weird nerds HELP" #thisisfine #I'm actually looking forward to tomorrow really I am enthusiastic #social anxiety #hermit recovery.	embarrassment, nervousness	amusement, confusion, nervousness, optimism

Table 4: Illustrative examples of misalignment between first-party Labels and third-party annotations.

similar to those obtained without demographic information in the prompt. Furthermore, the score distributions for both conditions, as presented in Figure 17 in the Appendix, also appear comparable. These observations suggest that while demographic information systematically impacts model predictions, the practical improvement in performance remains minimal.

5 Understanding Misalignment between First-Party Labels and Third-Party Annotations

We conducted a qualitative analysis to explore how the linguistic and contextual characteristics of posts may contribute to annotation discrepancies. We examined both cases where third-party annotations closely aligned with first-party labels and cases where they diverged significantly, aiming to identify patterns in post content that may explain these differences. Specifically, we defined high-alignment cases as posts where the majority-voted label received an F1-score greater than 0.6 and low-alignment cases as those with an F1-score below 0.2. These thresholds were determined based on the distribution of F1-scores across annotations. Content within these categories was manually reviewed by two authors.

For posts where third-party annotations (in-group, out-group, and LLMs) achieved near-perfect F1 scores, emotional expression was conveyed through explicit and unambiguous language. For instance, posts categorized by the first party as expressing "joy" explicitly included the word "happy," while expressions of "gratitude" featured phrases such as "thank you," and instances of "confusion" were marked by direct questions. In posts where third-party annotations achieved moderate to high F1 scores, a lack of clear linguistic cues led to subtle discrepancies rather than complete mismatches between the emotions identified by the author and those recognized by third-party annotators. E.g., third-party annotators selected additional emotions that the author did not specify, while in others, certain emotions identified by the author were not acknowledged by annotators.

Among posts exhibiting significant misalignment, while it is challenging to categorize all discrepancies due to their varied nature, we identified three notable patterns. First, posts lacking strong textual cues for emotion. Second, posts where essential context, crucial for interpretation, was unavailable to third-party annotators. For instance, the lack of conversational context, such as subreddit name and preceding conversational turns, in many Reddit posts may have posed a challenge for accurate emotion classification. In other instances, the necessary context can be personal to the first-party author, making the expressed emotion opaque to an external reader. Third, posts where authors self-reported "neutral" despite the presence of discernible emotional cues in the text. This particular pattern suggests that linguistic expressions of emotion do not necessarily reflect the author's internal emotional state. In addition, we observed that the majority of posts with first-party labels identified as spurious during manual quality check (illustrated in Appendix D) were found within this subset of misaligned posts, which is unsurprising, given that such labels would likely be misaligned with third-party interpretations. To illustrate annotation discrepancies described above, we present representative examples of posts with first-party and third-party annotations in Table 4.

6 Conclusion and Discussion

Our work challenges the assumption that third-party annotations (both human and LLM-based) can reliably infer authors' private states from their texts. By demonstrating a significant misalignment between third-party annotations and first-party (author-provided) labels, using both fine-grained and coarse-grained emotion taxonomies, we show the limitations of third-party interpretations across levels of granularity. We further explored methods to improve third-party annotation quality, leveraging first-party demographic information. We find that demographic similarity between first-party and third-party human annotators enhances annotation performance. Prompting first-party demographic traits marginally enhances LLMs' annotation performance.

A simple sentence like "I got a cup of coffee." can express different emotions depending on the speaker and the context, information that may or may not be transparent to a third party. Similarly, a statement like "I love morning classes." could be interpreted as expressing joy (genuine appreciation) or annoyance (sarcasm or frustration). These examples underscore the importance of clearly defining the modeling perspective: *Are we modeling the third-party's perception of the emotion expressed by the author, or the actual emotion expressed by the author in their written text?* Both first-party expressed and third-party perceived emotion can be valid modeling targets, depending on the application. For instance, empathetic support agents are designed to respond to a user's expressed emotion, while also ensuring their responses are perceived as supportive. However, it is crucial to recognize that perceived emotion cannot reliably serve as a substitute for an author's actual expressed emotion.

Third-party annotations struggle to differentiate between semantically similar emotions, such as joy and excitement, and to recognize the complexity of emotional expression—such as frustration underlying gratitude or anger coexisting with caring. This challenge highlights the inherent subjectivity in emotion recognition, where annotators bring their own interpretations, biases, and contextual assumptions to the task. Unlike first-party authors who experience the emotion firsthand, third-party annotators rely solely on textual cues, which may lack sufficient context for accurate inference. As a result, subtle emotional nuances, mixed emotions, or sarcasm often go unnoticed or misinterpreted. Moreover, third-party annotations may reflect individual, social, cultural influences on the perception of emotion rather than the actual emotional state intended by the author. For instance, an indirect expression of distress might not be recognized as sadness if it does not conform to expected linguistic patterns. This misalignment could undermine the reliability of models trained on third-party annotations and lead to unintended harm, particularly in high-stakes contexts, e.g., chatbots.

While our study focuses on emotion recognition, the challenges we identify have implications for other NLP tasks that attempt to infer authors' private states from text. Given these challenges, we argue that it is crucial to critically evaluate the extent to which third-party annotations can serve as a reliable ground truth for modeling authors' private states. Future research should explore methods, such as incorporating first-party feedback and explanation, to achieve a more truthful representation of the emotions actually expressed by the authors.

7 Limitations

The observed limitations of third-party annotations are based on first-party labels and third-party annotations collected from study participants. Even though we implemented strict quality control for first-party posts, we still observed first-party labels that may be spurious. These first-party data were retained in the analysis since whether first-party labels faithfully represent the authors' internal emotion states cannot be externally verified. Nevertheless, we anticipate that the inclusion of these data points has minimal impact on our overall findings. For third-party annotations, we removed annotations that did not follow annotation instructions, indicating a lack of attention. However, despite these quality control measures, some low-quality annotations may still remain, since inconsistencies in subjective judgment cannot be entirely eliminated.

In addition, we lack sufficient data within each demographic subgroup to conduct statistically robust analyses on how specific identity traits influence annotation performance. Furthermore, while this study examines differences between in-group and out-group annotators (based on age, gender, and race), our findings may not generalize to other demographic categorizations, broader populations, diverse cultural contexts, or different languages.

8 Ethical Consideration

The intention of our work is to highlight critical ethical concerns in inferring authors' private states (specifically emotions) via third-party annotations. Our research also brings to light additional ethical aspects that require careful consideration for future research and development in this field. The key issues include privacy risks when handling sensitive emotional data, potential bias from demographic mismatches between authors and annotators, and the unreliability of third-party labels (human or LLM) in capturing genuine emotions, risking harmful misjudgments in real-world applications. While our results show that LLMs outperform humans, in certain scenarios, they may perpetuate training data biases and foster overconfidence in flawed systems. Our study also underscores transparency gaps in methodology and LLM decision-making, labor concerns around displacing human annotators, and risks of mishandling demographic data. It calls for diverse annotation teams, participatory design, rigorous consent protocols, and bias audits to ensure ethical practices in modeling private states.

Acknowledgment

This work was supported by funding from the National Science Foundation award #2247723.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision support systems*, 53(4):742–753.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 574–577.

Ann Banfield. 1982. Unspeakable sentences : Narration and representation in the language of fiction.

Lisa Feldman Barrett. 2004. Feelings or words? understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology*, 87(2):266.

Patricia Bauer, Leif Stennes, and Jennifer Haight. 2003. Representation of the inner self in autobiography: Women's and men's use of internal states language in personal narratives. *Memory*, 11(1):27–42.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. *arXiv preprint cmp-lg/9406010*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sven Buechel and Udo Hahn. 2017. 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.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

David C DeAndrea, Allison S Shaw, and Timothy R Levine. 2010. Online language: The role of culture in self-expression and self-construal on facebook. *Journal of language and social psychology*, 29(4):425–442.

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

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.

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

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

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

Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, pages 1–20.

Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. Large language models for propaganda span annotation. In *EMNLP (Findings)*, pages 14522–14532. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1113–1122.

Kenneth Joseph, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (mis)alignment between stance expressed in social media data and public opinion surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 312–324, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176, St. Julians, Malta. Association for Computational Linguistics.

Mao Li and Frederick Conrad. 2024. Advancing annotation of stance in social media posts: A comparative analysis of large language models and crowd sourcing. *arXiv preprint arXiv:2406.07483*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Jasy Suet Yan Liew, Howard R Turtle, and Elizabeth D Liddy. 2016. Emotweet-28: A fine-grained emotion corpus for sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1149–1156.

Nangyeon Lim. 2016. Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integrative medicine research*, 5(2):105–109.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. Dens: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6293–6298.

Aire Mill, Jüri Allik, Anu Realo, and Raivo Valk. 2009. Age-related differences in emotion recognition ability: a cross-sectional study. *Emotion*, 9(5):619.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2009. A survey of named entity recognition and classification. In *Named Entities: Recognition, classification and use*, pages 3–28. John Benjamins publishing company.

László Nemes and Attila Kiss. 2021. Social media sentiment analysis based on covid-19. *Journal of Information and Telecommunication*, 5(1):1–15.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *CoRR*, abs/2408.17026.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Silviu Oprea and Walid Magdy. 2019. isarcasm: A dataset of intended sarcasm. *arXiv preprint arXiv:1911.03123*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *ACL*, pages 4296–4305. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, New York.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 139–147.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Disa A Sauter, Frank Eisner, Paul Ekman, and Sophie K Scott. 2010. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6):2408–2412.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Hannah Shaw and Minna Lyons. 2017. Lie detection accuracy—the role of age and the use of emotions as a reliable cue. *Journal of Police and Criminal Psychology*, 32:300–304.

Qinlan Shen and Carolyn Rose. 2021. What sounds "right" to me? experiential factors in the perception of political ideology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1762–1771, Online. Association for Computational Linguistics.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca Passonneau, and Shomir Wilson. 2023. The sentiment problem: A critical survey towards deconstructing sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763, Singapore. Association for Computational Linguistics.

Sophie F Waterloo, Susanne E Baumgartner, Jochen Peter, and Patti M Valkenburg. 2018. Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. *New media & society*, 20(5):1813–1831.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Artur Zygadło, Marek Kozłowski, and Artur Janicki. 2021. Text-based emotion recognition in english and polish for therapeutic chatbot. *Applied Sciences*, 11(21):10146.

Emotion Taxonomy with Definitions:

Admiration: 👏 Finding something impressive or worthy of respect.
Amusement: 😂Finding something funny or being entertained.
Approval: 👍Having or expressing a favorable opinion.
Caring: Displaying kindness and concern for others.
Desire: A strong feeling of wanting something or wishing for something to happen.
Excitement: 🤩Feeling of great enthusiasm and eagerness.
Gratitude: 🙏A feeling of thankfulness an d appreciation.
Joy: 😀A feeling of pleasure and happiness.
Love: ❤️A strong positive emotion of regard and affection.
Pride: Pleasure or satisfaction due to one's own achievements or the achievements of those with whom one is closely associated.
Optimism: 🤞 Hopefulness and confidence about the future or the success of something.
Relief: Reassurance and relaxation following release from anxiety or distress.
Anger: 😡 A strong feeling of displeasure or antagonism.
Annoyance: 😒Mild anger, irritation.
Disappointment: Sadness or displeasure caused by the nonfulfillment of one's hopes or expectations.
Disapproval: 👎 Having or expressing an unfavorable opinion.
Disgust: 🤮 Revulsion or strong disapproval aroused by something unpleasant or offensive.
Embarrassment: 😳 Self-consciousness, shame, or awkwardness.
Fear: 😨 Being afraid or worried.
Grief: Intense sorrow, especially caused by someone's death.
Nervousness: Apprehension, worry, anxiety.
Remorse: Regret or guilty feeling.
Sadness: 😢 Emotional pain, sorrow.
Realization: Becoming aware of something.
Surprise: 😮Feeling astonished, startled by something unexpected.
Confusion: 😕Lack of understanding, uncertainty.
Curiosity: A strong desire to know or learn something.
Neutral: No emotion.

Figure 2: Emotion taxonomy

A Emotion Taxonomy

We adopted the emotion taxonomy developed by Demszky et al. (2020), which is presented in Figure 2.

B Demographics and Annotation Statistics

In this section, we present the demographic breakdown of first-party participants, detailing the number of participants and the count of valid posts they contributed per group, in Table 5. Figure 3 illustrates the demographic composition of third-party annotators, showing the respective distributions for age, gender, and race. Additionally, Table 6 provides summary statistics of third-party annotations.

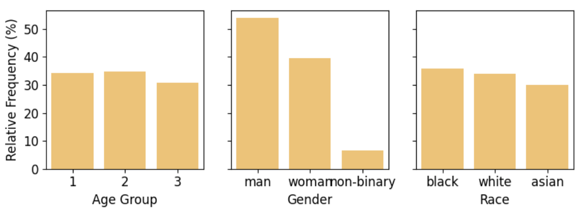

Figure 3: Distribution of third-party annotators by age group, gender, and race.

Age	Gender	Race	Participants	Posts
18-27	Man	Asian	8	33
18-27	Man	Black	7	55
18-27	Man	White	13	66
18-27	Woman	Asian	10	66
18-27	Woman	Black	8	48
18-27	Woman	White	8	57
28-43	Man	Asian	8	39
28-43	Man	Black	9	48
28-43	Man	White	8	59
28-43	Woman	Asian	9	44
28-43	Woman	Black	9	51
28-43	Woman	White	9	58
44-59	Man	White	8	51
44-59	Woman	White	9	54

Table 5: Demographic breakdown of first-party participants, showing the number of individuals and posts for each combination of age, gender, and race.

Number of annotators (unique)	399
Number of annotators	
- In-group role	236
- Out-group role	201
Average number of posts per annotator	
- In-group role	9.1
- Out-group role	10.7
Total annotations	
- By in-group annotators	2136
- By out-group annotators	2157

Table 6: Summary statistics of third-party annotations.

C Criteria and Verification for First-party Posts

To maintain data quality and authenticity, submitted posts were required to adhere to the following criteria: (1) originally created by the participant within the past 12 months (excluding shares, reposts, or non-original content); (2) in English and contain at least 5 words, excluding hashtags and URLs; (3) if multimedia is included, it should contain only images, with emotion conveyed through the images and text captured in the screenshot(s); (4) published at least 24 hours prior to submission; (5) fully captured in the screenshots. Additionally, participants were required to submit a screenshot of their social media account page, displaying their profile name, with the option to redact other information. The UI features of the account page helped us verify that

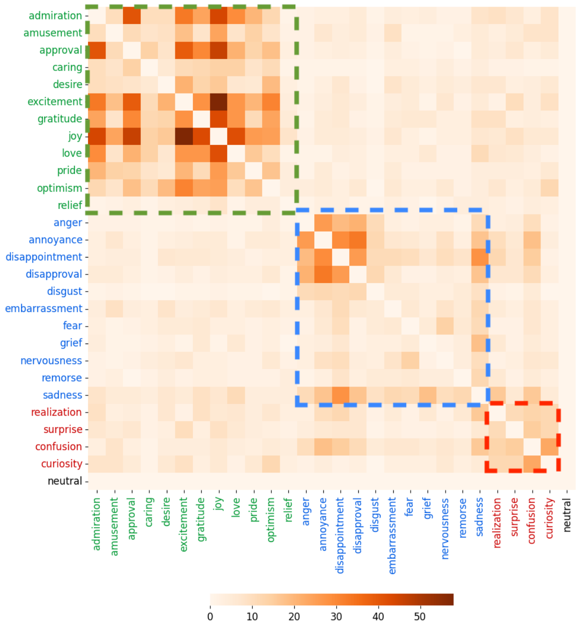

Figure 4: Co-occurrence of first-party emotion labels.

Emotion	First-party	In-group	Out-group	LLMs
Admiration	0.118	0.151	0.114	0.176
Amusement	0.102	0.103	0.078	0.162
Anger	0.070	0.060	0.059	0.097
Approval	0.126	0.106	0.089	0.195
Caring	0.064	0.047	0.047	0.187
Confusion	0.088	0.051	0.044	0.082
Curiosity	0.095	0.060	0.045	0.096
Desire	0.091	0.034	0.034	0.176
Disappointment	0.111	0.082	0.081	0.219
Disapproval	0.097	0.111	0.095	0.185
Disgust	0.041	0.040	0.030	0.051
Embarrassment	0.053	0.021	0.016	0.049
Excitement	0.150	0.106	0.088	0.226
Fear	0.044	0.032	0.029	0.041
Gratitude	0.126	0.130	0.108	0.089
Grief	0.034	0.019	0.022	0.029
Joy	0.176	0.187	0.165	0.332
Love	0.110	0.091	0.075	0.162
Nervousness	0.051	0.029	0.026	0.067
Neutral	0.063	0.038	0.043	0.010
Optimism	0.108	0.075	0.062	0.171
Pride	0.080	0.074	0.052	0.122
Realization	0.095	0.057	0.037	0.284
Relief	0.023	0.019	0.011	0.047
Remorse	0.034	0.014	0.008	0.025
Sadness	0.097	0.052	0.063	0.132
Surprise	0.063	0.043	0.030	0.086

Table 7: Distribution of emotion labels by annotator group.

participants had provided their own social media posts. We reviewed all submissions and removed posts that failed to meet one or more of the above criteria. We also excluded all submissions from participants whose submission authenticity could not be verified.

D Quality Check on First-party Labels

Two authors separately reviewed each post and its corresponding first-party label to identify spurious labels—cases where the first-party label could not be reasonably justified by the post content. They labeled only clear mismatches between first-party labels and the emotional expression in the posts, rather than cases of subtler misalignment where there might be room for interpretation. They agreed on 89% of the posts. For posts where disagreements occurred, a senior author conducted a final review to resolve discrepancies.

Among the 729 posts, 9.2% of first-party labels were flagged as spurious. This relatively low percentage suggests that the majority of first-party labels are reasonably aligned with the post content, indicating their overall high quality. We retain posts with labels identified as spurious in our analysis because a label being flagged by us as such (i.e., not clearly justified by the post content from an external perspective) does not definitively mean it fails to reflect the author's true internal emotional state. Indeed, we believe these instances are particularly valuable as they underscore the complexity of emotional expression and the inherent challenges, central to our study, of inferring emotions solely from textual content.

E Statistical Overview of First-party and Third-party Labels

First-party labels. Among all posts, 6.32% are labeled as Neutral, indicating no emotion is expressed. Of the remaining posts, 37.9% have a single emotion label, 25.3% have two labels, 14.6% have three labels, and 22.3% have four or more labels. Figure 4 shows the co-occurrence patterns of emotion labels, where the labels on the x- and y-axes are color-coded to represent distinct semantic categories: positive, negative, and ambiguous emotions. Positive emotions, such as joy, excitement, and admiration, occur more frequently overall compared to negative emotions like anger or disapproval. The heatmap reveals that emotions with similar conceptual or semantic tones often co-occur within the same post. For instance, positive emotions like joy, excitement, and admiration frequently appear together. Similarly, subsets of negative emotions, such as disappointment and annoyance, also exhibit notable co-occurrence. This pattern aligns with the observed tendency of social media users to predominantly share positive

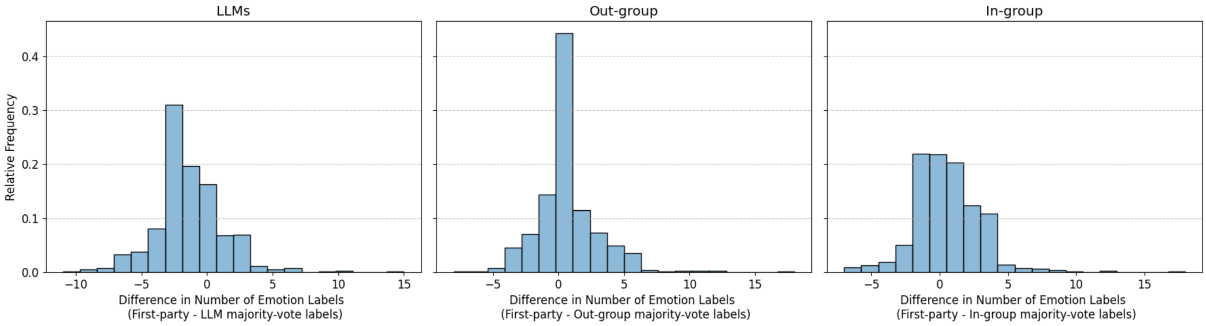

Figure 5: Distributions of the differences in the number of emotion labels assigned by first-party annotators and third-party annotators.

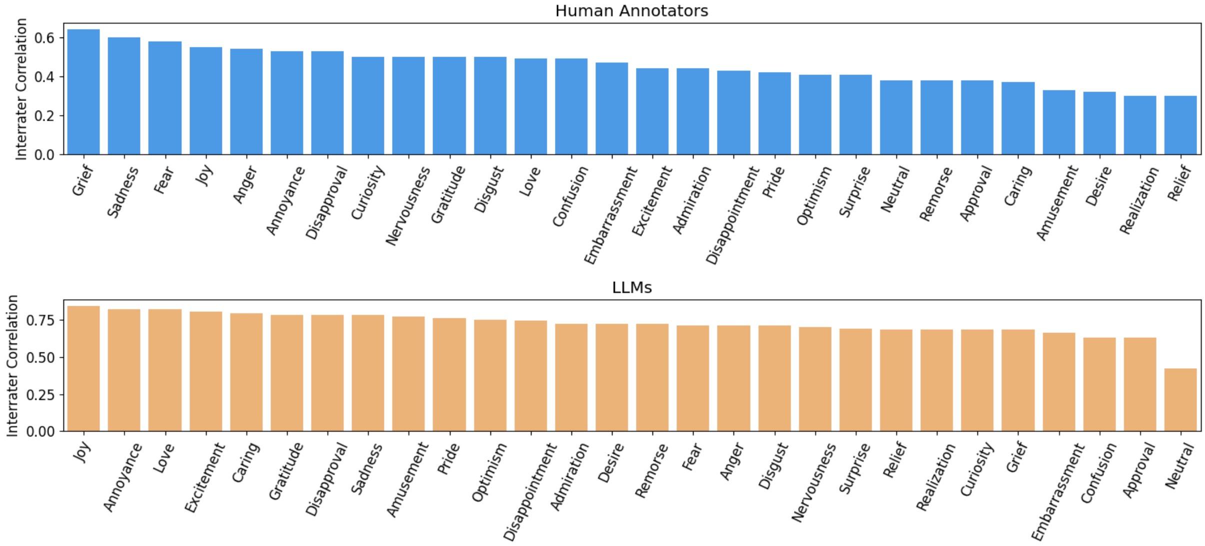

Figure 6: Interrater correlation among third-party annotators.

sentiments, while negative emotions appear less frequently (Waterloo et al., 2018). The co-occurrence of semantically related emotions suggests that users naturally cluster similar emotional tones in their posts, reflecting common patterns in emotional expression.

Third-party annotations. We applied majority voting separately to in-group annotations, out-group annotations, and LLM annotations. For in-group annotators, a majority decision was reached for all 28 emotions in 94% of posts, while for out-group annotators, a majority decision was reached for all 28 emotions in 97% of posts. In the case of LLMs, a majority label was assigned for all 28 emotions in every post, as we obtained annotations from 5 LLMs. Table 7 presents the label distribution for in-group, out-group, and LLM annotations.

For each annotator group, for posts not labeled as neutral, we calculated the difference in the number of emotion labels selected by first-party participants (authors) and by third-party annotators. The distribution of these differences is shown in Figure 5. The x-axis represents the difference in the number of emotion labels (calculated as first-party minus third-party), where negative values indicate over-labeling by third-party annotators, and positive values indicate under-labeling. LLMs exhibit a wider distribution with more over-labeling, while out-group annotators tend to align more closely with first-party labels, clustering around zero. In-group annotators show slightly more variation but tend to assign fewer labels than first-party participants.

F Agreement Among Third-party Annotators

Figure 6 presents the interrater agreement amongst human annotators and amongst LLMs, evaluated by interrater correlation. LLMs generally exhibit higher and more consistent interrater agreement across emotions compared to human annotators.

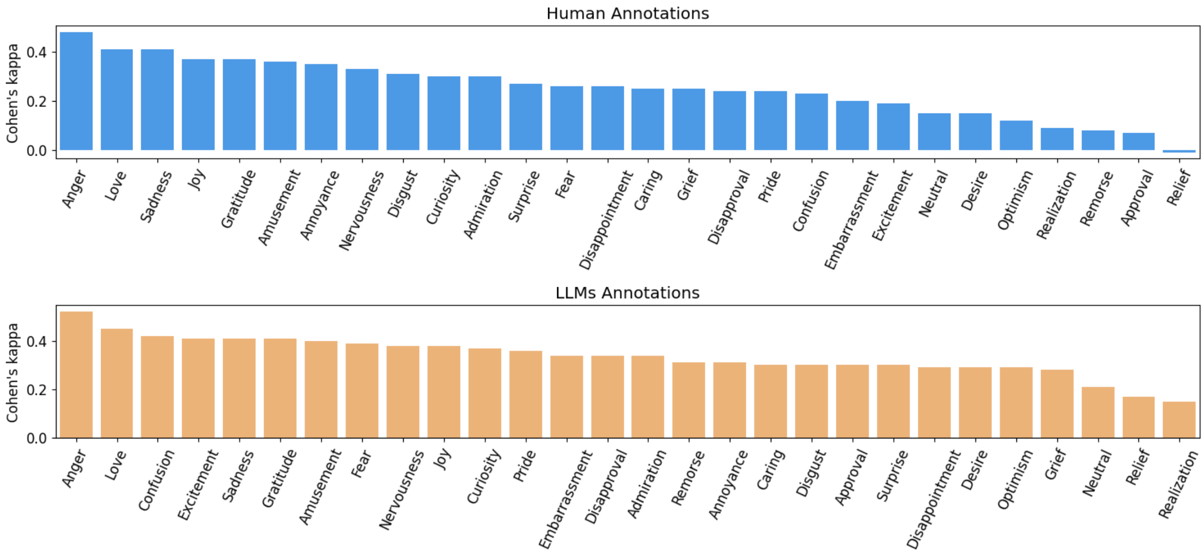

Figure 7: Alignment between third-party annotations with first-party labels, evaluated using Cohen's kappa.

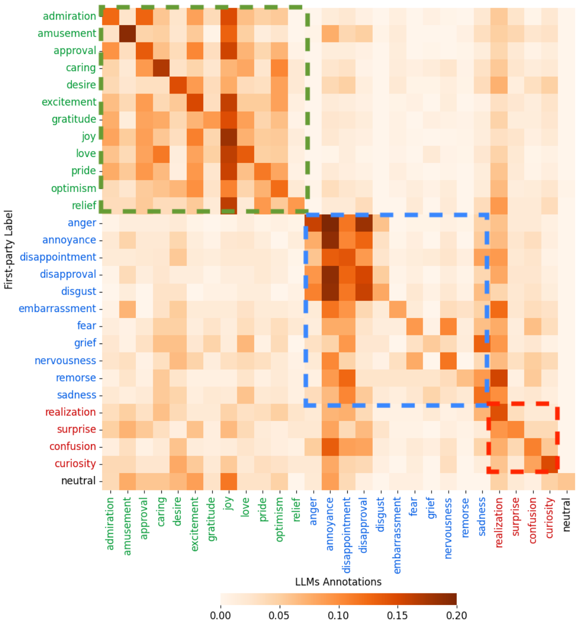

Figure 8: Confusion matrix showing the alignment between LLMs' annotations and first-party labels.

LLMs exhibit the highest agreement when labeling joy, love, and annoyance, while the lowest agreement is observed for neutral, confusion, and approval. In contrast, grief, sadness, and fear yield the highest interrater correlations among human annotators, whereas desire, realization, and relief yield the lowest.

Human annotators demonstrate higher agreement on negative emotions, while LLMs show higher agreement on positive emotions. Both human annotators and LLMs exhibit higher interrater agreement for emotions that are typically expressed with distinct textual cues and lower agreement for more nuanced and context-dependent emotions such as approval, relief, and realization. These emotions may be harder to infer externally due to their reliance on situational or implicit contextual cues.

We also computed Cohen's Kappa by randomly sampling two annotations from all human annotations for each post and calculating the agreement between these two sets of annotations. We followed an analogous procedure for LLM annotations. We computed the correlation between Cohen's kappa values and interrater correlation and found that Cohen's kappa and interrater correlation are highly correlated (LLMs: Pearson $r = 0.83$, $p < 0.001$; Human annotators: Pearson $r = 0.71$, $p < 0.001$). This strong correlation suggests that both metrics capture similar trends in annotator agreement.

G First- and Third-Party Alignment

Figure 7 presents the alignment between first-party labels and third-party annotations, evaluated by Cohen's Kappa. The levels of alignment between third-party annotations and first-party labels vary across different emotions. While LLMs demonstrate more consistent alignment across emotions, human annotators show greater variation, which potentially reflects a more subjective interpretation. Alignment is higher for emotions such as anger, love, and sadness for both human and LLM annotators, suggesting these emotions are more explicitly expressed and consistently recognized. Realization, relief, and approval yield the lowest alignment.

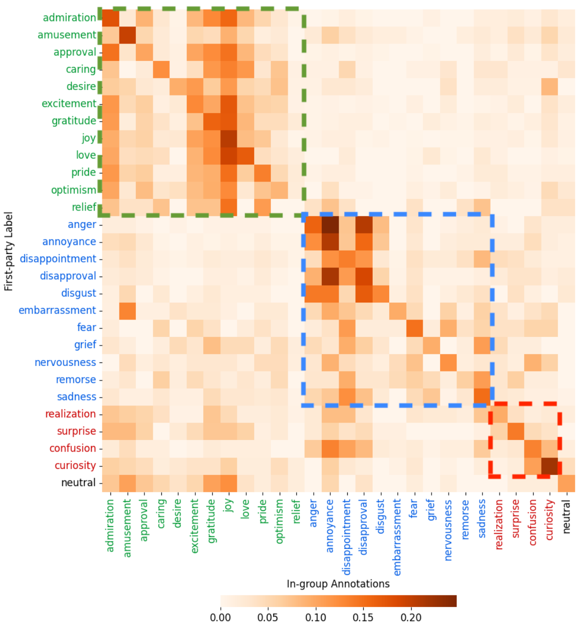

Figure 9: Confusion matrix showing the alignment between in-group annotations and first-party labels.

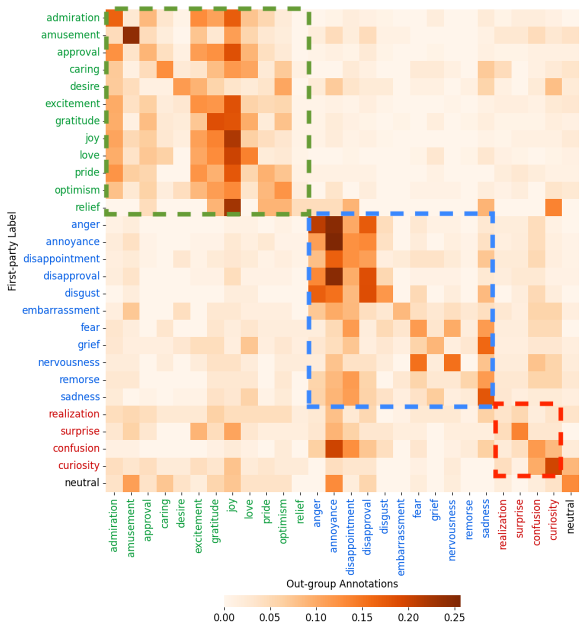

Figure 10: Confusion matrix showing the alignment between out-group annotations and first-party labels.

Since the Cohen's Kappa values are computed based on posts where a majority decision could be reached, we further present the percentage of posts where: (1) all annotators reached a decision; (2) a majority—but not all—of the annotators reached a decision; and (3) annotators failed to reach a decision due to ties. These percentages are reported in Table 8. For human annotations, on average, 27% of posts for a given emotion received a majority (but not unanimous) decision, 70% received a unanimous decision, and 3% resulted in ties. For LLM annotations, on average, 16% of posts yielded a majority decision and 84% a unanimous decision, with no ties occurring (as each post is annotated by 5 LLMs). The large proportion of posts where disagreement occurs, which aligns with the relatively low inter-rater agreement among third-party annotators, further highlights the highly subjective nature of emotion recognition.

We further examined the confusion matrices (Figures 8, 9, 10) to understand the patterns of alignment and misalignment between first-party labels and third-party annotations. As shown in the figures, when misalignments occur, third-party annotations often identify an emotion within the same broad semantic category as the first-party label. For example, positive emotions (e.g., joy, admiration, excitement) are frequently confused with one another, and similarly for negative emotions (e.g., anger, sadness, disappointment), as highlighted by the green and blue clusters. However, emotions that are more ambiguous or context-dependent, such as confusion, surprise, and realization, exhibit greater divergence between first-party and third-party labels, as seen in the red-outlined cluster. Additionally, LLMs show more consistent annotation patterns across emotions, whereas human annotators display greater variability, reflecting possible subjectivity in emotion interpretation.

H Alignment Between Third-party Annotations and First-party Labels on Coarse-Grained Emotions

H.1 Coarse-Grained Emotion Taxonomy

To evaluate the alignment between third-party annotations and first-party labels at a higher level, we group the 28 emotion categories into seven groups: joy, love, anger, surprise, fear, sadness, and neutral. Since there is no universally agreed-upon set of basic emotions, we construct our grouping based on the basic emotion categories proposed by Ekman (1992) (anger, disgust, fear, joy, neutral, sadness, surprise) and Shaver et al. (1987) (anger, love, fear, joy, sadness, and surprise). The mapping is illustrated in Table 9.

H.2 Third-party Annotators' Performance

We analyze the alignment between third-party annotations and first-party labels after mapping the original 28 emotion categories into 7 broader emotion

Emotion	Unanimous Agreement	Majority Agreement	Ties
Anger	0.76	0.22	0.02
Love	0.66	0.31	0.04
Sadness	0.79	0.19	0.02
Joy	0.53	0.41	0.06
Gratitude	0.56	0.37	0.06
Amusement	0.57	0.39	0.04
Annoyance	0.60	0.32	0.08
Nervousness	0.82	0.17	0.01
Disgust	0.80	0.19	0.01
Curiosity	0.77	0.20	0.02
Admiration	0.51	0.43	0.07
Surprise	0.73	0.24	0.03
Fear	0.83	0.16	0.01
Disappointment	0.62	0.33	0.05
Caring	0.68	0.29	0.02
Grief	0.87	0.12	0.01
Disapproval	0.66	0.26	0.08
Pride	0.65	0.30	0.05
Confusion	0.75	0.23	0.03
Embarrassment	0.86	0.13	0.01
Excitement	0.61	0.32	0.07
No emotion	0.71	0.28	0.02
Desire	0.71	0.26	0.02
Optimism	0.62	0.34	0.04
Realization	0.65	0.32	0.03
Remorse	0.87	0.12	0.01
Approval	0.54	0.40	0.06
Relief	0.78	0.21	0.01

Table 8: Alignment between third-party annotations with first-party labels, evaluated by Cohen's kappa.

groups. By aggregating annotations, we examine whether third-party annotators, while struggling to capture fine-grained emotions, align more closely with first-party labels when emotions are categorized at a higher level. For instance, annotations such as "Admiration" and "Approval" are grouped under "Joy". To assess alignment, similar to 4.2, we computed Cohen's kappa, F1, recall, and precision. Figure 11 presents Cohen's kappa scores for each emotion for human and LLM annotations. For human annotations, Cohen's kappa scores fall within the range of 0.15–0.6. For LLM annotations, Cohen's kappa scores fall within the range of 0.2–0.5.

The macro-average scores for LLM annotations are as follows: (F1: 0.54, Recall: 0.61, Precision: 0.57). For out-group human annotators, the scores are (F1: 0.45, Recall: 0.39, Precision: 0.55), while for in-group human annotators, they are (F1: 0.47, Recall: 0.41, Precision: 0.55). Figure 12 presents the F1 scores achieved by third-party human annotators (in-group, out-group) and LLMs across different emotions. These results suggest that third-party annotators struggle to identify the emotion expressed by the first party even at a higher level, indicating that this misalignment is not solely due to

Coarse Emotion	Fine-Grained Emotions
Love	Caring, Love, Desire, Gratitude
Fear	Fear, Nervousness
Joy	Joy, Amusement, Excitement, Optimism, Pride, Relief, Approval, Admiration
Sadness	Sadness, Grief, Disappointment, Remorse, Embarrassment
Surprise	Surprise, Confusion, Curiosity, Realization
Anger	Disapproval, Disgust, Anger, Annoyance
Neutral	Neutral

Table 9: Coarse Emotions and Corresponding Fine-Grained Emotions

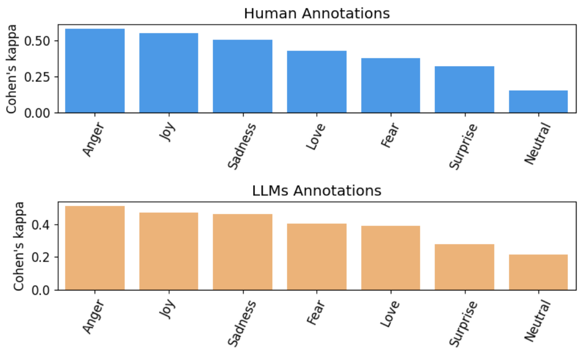

Figure 11: Alignment between third-party annotations with first-party labels, on coarse-grained emotions, evaluated using Cohen Kappa.

mistakenly interpreting similar emotions within the same broader category, but also stems from selecting emotions that are substantially different from those expressed by the first party. This highlights the inherent limitations in third-party annotations.

H.3 In- vs. Out-Group Annotators

Similar to 4.3, we explored the impact of demographic similarity between first-party and third-party annotators on the alignment between first-party labels and third-party annotations when emotions are grouped into higher-level categories. We follow the same procedure as in Section 4.3.

Post-level comparison. We applied the Wilcoxon signed-rank test to compare the F1, recall, and precision scores obtained by in-group majority-voted labels and out-group majority-voted labels for each post. The comparison is based on 93% of posts.

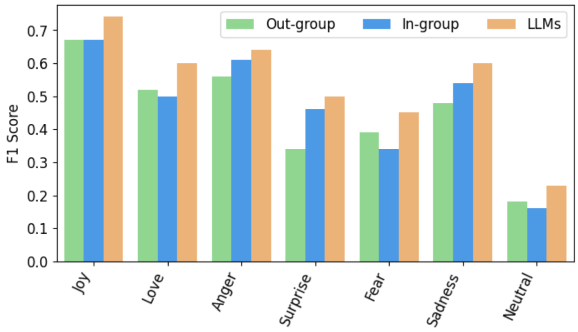

Figure 12: Alignment between third-party annotations with first-party labels, on coarse-grained emotions, evaluated using F1 score.

The Wilcoxon signed-rank test results indicate that in-group and out-group annotators perform similarly in terms of F1-score ($\text{Median}_{\text{in-group}} = 0.67$, $\text{Median}_{\text{out-group}} = 0.67$, $p = 0.05$), with the difference being marginally significant. Precision also shows no significant difference ($\text{Median}_{\text{in-group}} = 0.58$, $\text{Median}_{\text{out-group}} = 0.57$, $p = 0.602$). However, in-group annotators achieve significantly higher recall than out-group annotators ($\text{Median}_{\text{in-group}} = 1.00$, $\text{Median}_{\text{out-group}} = 0.50$, $p = 3 \times 10^{-3}$), suggesting that they are better at capturing first-party expressed emotions at a broader level.

Annotator-level comparison. We adopted the same method presented in 4.3 to test the impact of demographic similarity between third-party and first-party annotators at the annotator level. The results of the mixed linear models indicate that in-group annotators achieve significantly higher recall ($\text{Median}_{\text{in-group}} = 0.60$, $\text{Median}_{\text{out-group}} = 0.56$, $p = 0.006$) and marginally higher F1-score ($\text{Median}_{\text{in-group}} = 0.53$, $\text{Median}_{\text{out-group}} = 0.51$, $p = 0.04$) compared to out-group annotators. However, no significant difference is observed in precision ($\text{Median}_{\text{in-group}} = 0.54$, $\text{Median}_{\text{out-group}} = 0.52$, $p = 0.29$). These findings suggest that demographic similarity continues to play a role in improving annotation quality, even when evaluating alignment between third-party annotations and first-party labels at a coarse-grained level of emotion categorization.

I LLM Prompts

We introduce the LLM prompts used for annotations: Figure 13 presents the prompt incorporating first-party demographic information (age, gender, and race), while Figure 14 shows the baseline prompt used without these details. The specific prompt formulation for each of these two conditions was applied consistently across all LLMs.

```
messages=[{
                    "role": "system",
                    "content": You will be given a social media post written by an author. Your task is to  determine
                                  whether each of the following emotions is expressed by the author in the post. For each
                                  emotion in the emotion taxonomy below, indicate whether the emotion is expressed in
                                  the post by writing either 'yes' or 'no'. Respond exactly in the following format:
                                       1. Admiration: yes
                                       2. Amusement: no
                                        ...
                                       28. No emotion: no
                                  If you can't confidently assign any labels from the taxonomy, output
                                  'Unrecognizable.' Here is the [Emotion Taxonomy with Definitions]
                 },
                 {
                     "role": "user",
                     "content": Below is an {image} of the social media post written by an author who is a {race}
                                  {gender} from {age group}. The author’s text is highlighted with a red rectangle
                                  when other texts are present in the image or when the author’s text is less  noticeable.
                 }]
```

Figure 13: Prompt used to elicit LLM annotations.

```
messages=[{
                    "role": "system",
                    "content": You will be given a social media post written by an author. Your task is to  determine
                                  whether each of the following emotions is expressed by the author in the post. For each
                                  emotion in the emotion taxonomy below, indicate whether the emotion is expressed in
                                  the post by writing either 'yes' or 'no'. Respond exactly in the following format:
                                       1. Admiration: yes
                                       2. Amusement: no
                                        ...
                                       28. No emotion: no
                                  If you can't confidently assign any labels from the taxonomy, output
                                  'Unrecognizable.' Here is the [Emotion Taxonomy with Definitions]
                 },
                 {
                     "role": "user",
                     "content": Below is an {image} of the social media post. The author’s text is highlighted with a red
                                  rectangle when other texts are present in the image or when the author’s text is less
                                  noticeable.
                 }]
```

Figure 14: Prompt with first-party demographic information.

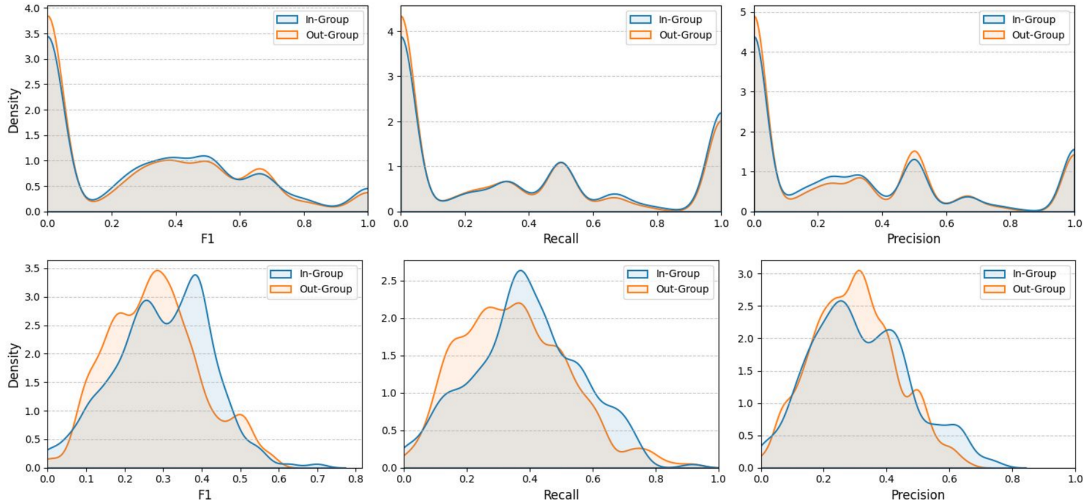

Figure 15: Density plots showing the distribution of F1 scores, recall for each post (first row). Density plots showing the distribution of averaged F1 scores, recall individual annotators per task (second row).

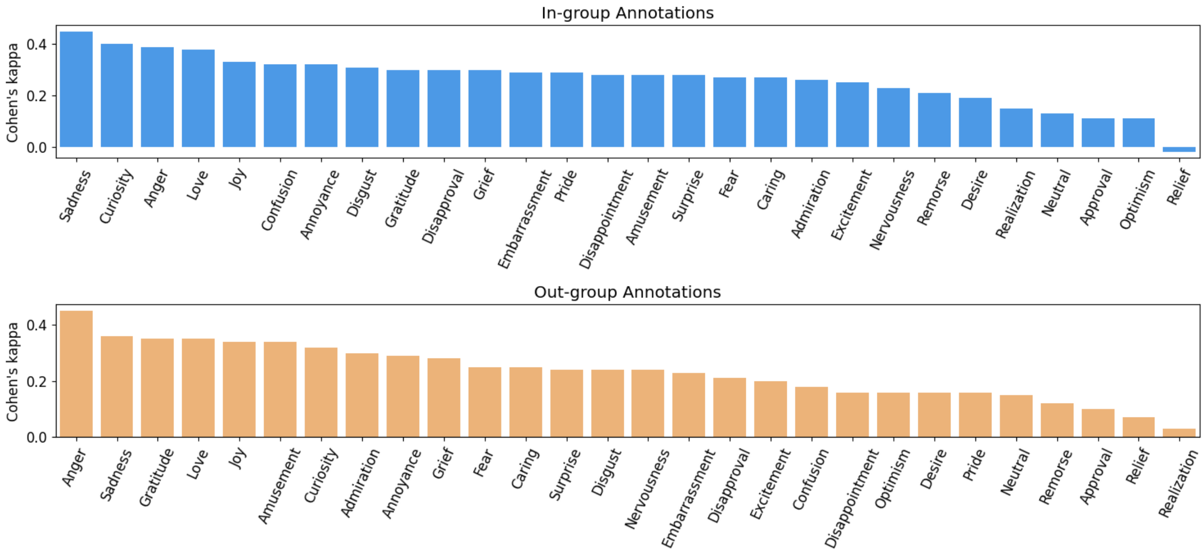

Figure 16: Performance comparison of in-group annotators and out-group annotators by Cohen's kappa

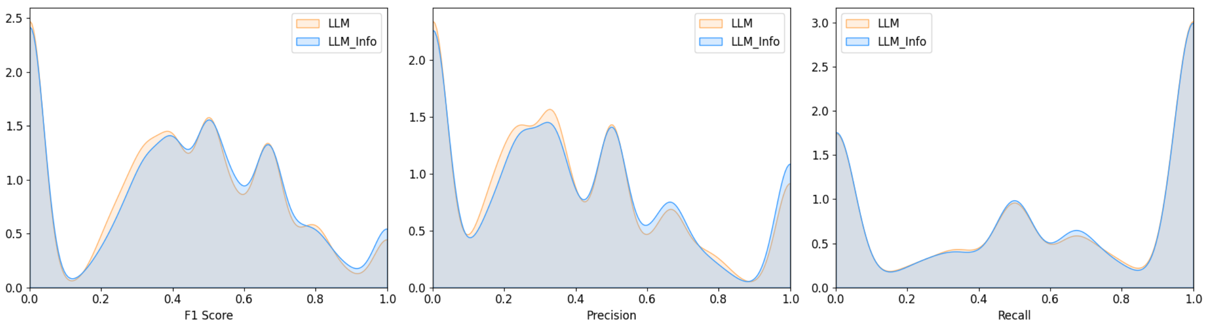

Figure 17: Density plots showing the distribution of F1 scores, recall, and precision for each.